On the Choice of Sampling Rates in Optimal Linear Systems

Åström, Karl Johan

1963

IBM RESEARCH

ON THE CHOICE OF SAMPLING RATES IN OPTIMAL LINEAR SYSTEMS

Karl J. Åström

RJ 243

April 25, 1963

SCIENCES

MATERIAL TECHNOLOGIES

MATH AND PROGRAMMING

ENGINEERING STUDIES

MACHINES AND MACHINE ORGANIZATION

PRINTED IN U. S. A.

ON THE CHOICE OF SAMPLING RATES IN OPTIMAL LINEAR SYSTEMS

by

Karl J. Åström

International Business Machines Corporation
San Jose Research Laboratory
San Jose, California

ABSTRACT: In a sample data system we are essentially restricted to control signals which are constant over the sampling interval. The purpose of this report is to analyze the effect of this restriction. The situation where the admissible control signals are piecewise continuous is used as a reference case and the problem is approached from a variational formulation of the control problem where the object of the control is to minimize a loss function. The minimal values of the loss function over the classes of piecewise constant and piecewise continuous control signals are compared. It is assumed that the system is linear and the criterion quadratic. Asymptotic formulas for small sampling intervals and Fortran programs for the evaluation of the values of the loss function in the discrete and continuous cases are given.

Research Report
RJ-243
April 25, 1963

I. INTRODUCTION:

When controlling a physical system, we are faced with the following situation: The system can be influenced by applying control signals to it, and its status can be observed by measuring certain output signals. The basic problem of control is to generate a control signal on the basis of the observed outputs in such a way that the objective of the system is achieved. To implement the solution of the problem, one frequently uses schemes where the data processing and the function generation are performed at discrete instants of time, so called sampled data systems. There are several reasons for this: The complexity and accuracy of the computations might require a digital computer. The output signals are available only at discrete times. For example, this is the case in the control of many chemical processes where analytical instruments are used for composition measurements, or in fire control systems where target data are available only once per revolution of the radar antenna.

The choice of sampling rate is a fundamental problem in the design of sampled data systems. There are many considerations which affect this choice. In some applications, the maximum sampling rate is given by the available equipment; for example, in radar applications where the rate is given by the angular velocity of the antenna, or in the control of a chemical process where the sampling rate is given by the procedure used for the composition analysis of the product. It is, however, important to determine what sampling rates are intrinsically required for the solution of a particular problem. This information can be used for example to determine the specifications of the computer required for the implementation of the system or to find the incentive in an improved measurement technique.

There exist at present no rational procedure for the choice of sampling rates in discrete systems. Shannon's sampling theorem has been suggested for this purpose.[10]

In essence, Shannon's theorem states that signals whose power spectra are zero outside the frequency interval $(-f, f)$ can be represented by the values of the time function at time intervals with the spacing $\frac{1}{2f}$. The filter required to reconstruct the continuous signal from such a representation is, however, not physically realizable. The approximation of the ideal filter by physically realizable filters leads to considerable delay of the reconstructed signal. There is no difficulty in an open loop system. In a closed loop system, however, this delay may give rise to stability problems. When compromising between a slower sampling rate with a sophisticated data reconstruction and a higher sampling rate with a simpler data reconstruction, one, therefore, frequently chooses the latter alternative. For example, in carrier frequency servos, it is not unusual to have a bandwidth of the servo which is a tenth of the carrier frequency.

In recent years, control problems have been successfully treated as variational problems, the object of the control being to minimize a scalar loss function. The basic idea of this paper is to exploit this approach in order to see if it can give some insight into the choice of sampling rates in a discrete system. By using a sampled data system instead of a continuous system, we restrict the data processing and this should, therefore, affect the loss function. To be able to persue the subject analytically, it is assumed that the system can be described by linear differential equations and that the objective is to minimize the integral of a quadratic form in the state-and control variables. It has been shown[3, 5] that the data processing problem in this case can be separated into two problems; first, the minimum mean square estimation of the state from the observed outputs, and, second, the deterministic control problem, i.e., knowing the state to determine the optimal control. Furthermore, it has been demonstrated[7] that these problems are dual.

The influence of the sampling rate in the deterministic optimal control problem is analyzed in Section II. In the discrete case, the control signal has to be a prescribed function of time over the sampling interval: Constant, linear, quadratic, exponential, etc. All cases can be reduced to the situation where the control signal is constant over the sampling interval by expanding the state of the system and we will, therefore, only analyze this case. (The reduction of the other cases is demonstrated by examples in Section III.) The main problem is thus to compare the minimal values of the loss function over the classes of piecewise continuous and piecewise constant control signals, respectively. Let V and $\tilde{V}$ denote the loss function in the two cases. It is shown that

$$1 + \beta_1 \leq \frac{\tilde{V}}{V} \leq 1 + \beta_2 \quad , \quad \beta_1 > 0 \quad , \quad \beta_2 > 0$$

The quantities β_1 and β_2, which can be interpreted respectively as the minimum and maximum relative increase in the loss function due to sampling, are related analytically to the parameters of the criterion and system equations. It is shown that they converge to zero as the sampling intervals tend to zero. In the case of stationary systems with constant sampling intervals of length h, we obtain an asymptotic estimate

$$\beta_i = \alpha_i h^2 + O(h^3)$$

which is very useful for the determination of sampling rates in practical problems.

In Section III, we demonstrate the application of the results of Section II to some examples. It is found that the values of the sampling rates obtained by the methods of Section II in some cases are widely different from those obtained from considerations based on the sampling theorem.

In Section IV, we analyze the influence of the sampling rate on the solution of the discrete estimation problem. The main problem here is to analyze the effect of having to operate on the observations at discrete instants of time as compared to a continuous data processing. As can be expected from the principle of duality, the problem is essentially the same as the problem discussed in Section II and the results previously obtained are extensively used.

A straightforward combination of the results of Sections II and IV will give the influence of the sampling rate in the linear stochastic optimal control problem.

The paper makes extensive use of results developed recently in the linear theory of optimal control. As there is not yet available a unified, comprehensive treatment of this, we have included in the Appendices A, B, C, and D statements of the main results used in this paper.

In Appendix A, we give the main results of linear continuous optimal control. The treatment is based on Reference 6 to which we refer for proofs and further details.

In Appendix B, we give the transformation to the discrete problem and the main results of the theory of time discrete linear optimal control. This problem was first solved in Reference 8; proofs and control theoretic interpretations are also found in References 1 and 2.

In Appendices C and D, we have stated the main results on the continuous and discrete filtering. Most of the results stated in these sections are based on Reference 7.

The results obtained in Section II are in a form which is well suited for numerical computations on a digital computer. For the determination of suitable sampling rates for a stationary system, we have developed a set of computer programs. In Appendix E, we give a program for the computation of β_1 and β_2 , and in Appendix F, we give a program for the evaluation of the asymptotic bounds α_1 and α_2.

The main conclusion of the results of this paper is that the formulation of a control problem as a variational problem will give a rational way to determine the sampling rate in a discrete system by analyzing the increase in the loss function due to the sampling.

II. THE INFLUENCE OF THE SAMPLING RATE IN THE DISCRETE CONTROL PROBLEM:

Consider a linear system described by

$$\frac{dx}{dt} = F(t)\,x + G(t)\,u \tag{2.1}$$

where $x(t)$ and $u(t)$ for fixed t are real n and r vectors, called the state vector and the control vector, respectively. The elements of the matrices $F(t)$ and $G(t)$ are assumed to be continuous and bounded.

Given the initial condition

$$x(t_o) = x_o \tag{2.2}$$

and given $t_1 > t_o$. Now consider the following functional of the solution $x_u(t_o, x_o)$

$$V(x_o, t_o, t_1, u) = x^T(t_1)Q_o\,x(t_1) + \int_{t_o}^{t_1} \begin{pmatrix} x(s) \\ u(s) \end{pmatrix}^T Q(s) \begin{pmatrix} x(s) \\ u(s) \end{pmatrix} ds \tag{2.3}$$

where $Q(t)$ is a positive semi-definite symmetrical matrix, with bounded elements

$$Q(t) = \begin{pmatrix} Q_{11}(t) & Q_{12}(t) \\ Q_{21}(t) & Q_{22}(t) \end{pmatrix} \tag{2.4}$$

$$Q_{21}(t) = Q_{12}^T(t) \tag{2.5}$$

Let $u(t)$ belong to the class of admissible controls U.

PROBLEM: Given the system (2.1) and the criterion (2.3), find the difference between the minimal values of the functional (2.3) with respect to the class of admissible controls U when (a) $U = U_c$ is the class of piecewise continuous functions and (b) $U = U_d$ is the class of piecewise constant functions.

The (a) portion of the problem is referred to as the continuous problem, and the (b) portion as the discrete problem. The control schemes defined by the solutions are called continuous control and discrete or sampled data control, respectively.

The object of this paper is to provide a solution to the problem stated above. We start with some preliminaries.

Let $V^o(x_o, t_o; t_1)$ and $\tilde{V}^o(x_o, t_o; t_1)$ be the minimal values of the functional (2.3) in the following two cases:

$$V^o(x_o, t_o; t_1) = \min_{u \in U_c} V(x_o, t_o; t_1, u) \tag{2.6}$$

$$\tilde{V}^o(x_o, t_o; t_1) = \min_{u \in U_d} V(x_o, t_o; t_1, u) \tag{2.7}$$

It is well known that under certain regularity conditions[6] the functions $V^o(x, t; t_1)$ and $\tilde{V}^o(x, t; t_1)$ are quadratics in the initial state, i.e.,

$$V^o(x, t; t_1) = x^T S(t; t_1) x \tag{2.8}$$

$$\tilde{V}^o(x, t; t_1) = x^T \tilde{S}(t; t_1) x \tag{2.9}$$

where the matrices $S(t;t_1)$ and $\tilde{S}(t;t_1)$ are positive definite and bounded. Analytical expressions for them in terms of $F(t)$, $G(t)$, and $Q(t)$ are also available. Summaries of these well-known results, which are frequently used in this paper, are given in Appendices A and B. As $U_d \subset U_c$, the difference

$$W(x, t;t_1) = \tilde{V}^o(x, t;t_1) - V^o(x, t;t_1) = x^T \, T(t;t_1)\, x \tag{2.10}$$

is positive semi-definite. The quantity W can obviously be interpreted as the additional increase of the function (2.3) due to the constraint that the control signal has to be constant over the sampling interval or "the additional loss due to sampling".

We have

$$\lambda_{min}(TS^{-1}) \leq \frac{\tilde{V}^o - V^o}{V^o} \leq \lambda_{max}(TS^{-1}) \leq \|TS^{-1}\| \tag{2.11}$$

where $\lambda_{min}(A)$ and $\lambda_{max}(A)$ denote the magnitudes of the maximum and minimum eigenvalues of A, and we have temporarily dropped the arguments t_o and t_1 in T, S, $\tilde{V}^o$, and V^o. As the quantity W depends on the initial state x, we will frequently characterize W by its maximum $\lambda_{max}(TS^{-1})$ which can be interpreted physically as the maximum relative increase in the loss function due to sampling. We will now consider the solution of the problem.

We have

<u>THEOREM 2.1:</u>

$$\tilde{V}(x, t, t_1) \rightarrow V(x, t;t_1) \quad \text{as} \quad \max_i (\tau_i) \rightarrow 0$$

PROOF:

The discrete version of the canonical equations (B. 17) can be regarded as a difference approximation to the Euler equations (A. 10). It is easily verified that the continuity of $F(t)$, $G(t)$, and $Q(t)$ implies that (B. 17) is a consistent difference approximation of (A. 10). As (B. 17) is linear, it now follows from a well-known result in numerical analysis (Henrici,[4] Theorem 3.2, p. 124), that

$$\tilde{\sum}(t_k;t_1) \rightarrow \sum(t_k;t_1) \quad \text{as} \quad \max_i (T_i) \rightarrow 0$$

But $\tilde{V}$ and V are expressed in terms of $\tilde{\sum}(t;t_1)$ and $\sum(t;t_1)$ by (A.15) and (B. 22), which proves the theorem.

The theorem implies that the value of the functional (2. 3) in the discrete case can be made arbitrarily close to its value in the continuous case by choosing the sampling interval small enough.

We will now give an estimate of the difference between the discrete and the continuous case. To simplify the algebraical work, we will assume that the dynamical system (2. 1) is stationary, that $Q(t)$ is a constant matrix, and that the sampling period is constant h.

Before giving the estimate, we will make an observation which simplifies the formal work. The inclusion of $Q_{12}(t)$ is immaterial in the continuous problem statement. If Q_{12} is non-zero, we can make the transformation

$$F^* = F - GQ_{22}^{-1} Q_{21} \tag{2. 12}$$

$$Q_{11}^* = Q_{11} - Q_{12}Q_{22}^{-1} Q_{21} \tag{2. 13}$$

and we have a problem without Q_{12}. Therefore, throughout this section we will assume that $Q_{12} = 0$.

THEOREM 2.2:

Let the matrices F, G, and Q be stationary, assume that Q_{22} is positive definite and that the sampling intervals are constant $\tau_i = h$, then

$$S(t;t_1) = [\Sigma_{21}(t;t_1) + \Sigma_{22}(t;t_1)Q_o]\ [\Sigma_{11}(t;t_1) + \Sigma_{12}(t;t_1)Q_o]^{-1} \tag{2.14}$$

$$\tilde{S}(t;t_1) = [\tilde{\Sigma}_{21}(t;t_1) + \tilde{\Sigma}_{22}(t;t_1)Q_o]\ [\tilde{\Sigma}_{11}(t;t_1) + \tilde{\Sigma}_{12}(t;t_1)Q_o]^{-1} \tag{2.15}$$

where

$$\frac{d}{dt}\sum(t;t_1) = A \cdot \sum(t;t_1) \quad , \quad \sum(t_1;t_1) = I \tag{2.16}$$

$$\tilde{\sum}(t;t_1) = \sum(t;t_1) + \frac{h^2}{12} E(t;t_1) + O(h^3) \tag{2.17}$$

$$\frac{d}{dt} E(t;t_1) = A \cdot E(t;t_1) + B\sum(t,t_1) \quad , \quad E(t_1;t_1) = 0 \tag{2.18}$$

$$A = \begin{pmatrix} F & -GQ_{22}^{-1}G^T \\ -Q_{11} & -F^T \end{pmatrix} \tag{2.19}$$

$$B = \begin{pmatrix} FGQ_{22}^{-1}G^TQ_{11} & FGQ_{22}^{-1}G^TF^T \\ -Q_{11}GQ_{22}^{-1}G^TQ_{11} & -Q_{11}GQ_{22}^{-1}G^TF^T \end{pmatrix} \tag{2.20}$$

PROOF:

The transition matrix $\tilde{\sum}(t+h;t)$ is given by

$$\tilde{\Sigma}_{11}(t_k+h;t_k) = A_1^T + A_2 A_1^{-1} A_3$$

$$\tilde{\Sigma}_{12}(t_k+h;t_k) = A_2 A_1^{-1}$$

$$\tilde{\Sigma}_{21}(t_k+h;t_k) = -A_1^{-1} A_3$$

$$\tilde{\Sigma}_{22}(t_k+h;t_k) = A_1^{-1}$$

where

$$A_1 = [\Phi - \Gamma \tilde{Q}_{22}^{-1} \tilde{Q}_{21}]^T$$

$$A_2 = \Gamma \tilde{Q}_{22}^{-1} \Gamma^T$$

$$A_3 = \tilde{Q}_{11} - \tilde{Q}_{12} \tilde{Q}_{22}^{-1} \tilde{Q}_{21}$$

and

$$\Phi(t) = \exp(Ft)$$

$$\Gamma(t) = \left[\int_o^h \Phi(t)\, dt \right] G$$

$$\tilde{Q}_{11} = \int_o^h \Phi^T(t)\, Q_{11}\, \Phi(t)\, dt$$

$$\tilde{Q}_{12} = \int_0^h \Phi^T(t)\, Q_{11}\, \Gamma(t)\, dt$$

$$\tilde{Q}_{22} = \int_0^h [\Gamma^T(t)\, Q_{11}\, \Gamma(t) + Q_{22}]\, dt$$

The equation (B.17) which gives the solution of the discrete problem can be regarded as a difference approximation of (A. 10), yielding the solution of the continuous problem. We have

$$\sum(t+h;t) = \sum_{n=0}^{\infty} \frac{1}{n!} A^n h^n \tag{2.21}$$

A tedious but trivial series expansion of $\tilde{\sum}(t+h,t)$ gives

$$\tilde{\sum}(t+h;t) = I + hA + \frac{1}{2!} h^2 A^2 + \frac{1}{3!} h^3 (A^3 + B) + O(h^4) \tag{2.22}$$

This implies that (B. 17), regarded as a difference approximation of (A. 10), has a local truncation error

$$\frac{1}{12} h^3 B + O(h^4)$$

The asymptotic formula (2.17) for $\tilde{\sum}(t;t_1)$ now follows from a well-known result in numerical analysis. (See Henrici,[4] Theorem 3.4, p. 135.)

It should be noted that:

1. The assumption of a constant sampling interval is not essential. Theorem 2.2 still holds if $h = \max_i |\tau_i|$.

2. Results similar to Theorem 2.2 can also be obtained in the case of time varying coefficients using the same technique. In the general case, the difference W will, however, be of the order of h. The asymptotic formula corresponding to (2.17) has, however, a complicated analytical form.

Let $E(t;t_1)$ be the solution of (2.18) and let the matrix $C(t;t_1)$ be defined by

$$C = \{E_{21} + E_{22} Q_o - S[E_{11} + E_{12} Q_o]\} [\overset{\Sigma}{\cancel{E}}_{11} + \overset{\Sigma}{\cancel{E}}_{12} Q_o]^{-1} \qquad (2.23)$$

where the arguments t and t_1 are temporarily dropped in the matrices C, S, E_{ij}, and Σ_{ij}. It now follows from Theorem 2.2:

<u>Corollary 2.3:</u>

$$\tilde{S}(t;t_1) = S(t;t_1) + \frac{h^2}{12} C(t;t_1) + O(h^3) \qquad (2.24)$$

<u>Corollary 2.4:</u>

$$\max_x \frac{\tilde{V}(x,t;t_1)}{V(x,t;t_1)} = 1 + \frac{h^2}{12} \lambda_{\max} (S^{-1}C) + O(h^3) \qquad (2.25)$$

$$\min_x \frac{\tilde{V}(x,t;t_1)}{V(x,t;t_1)} = 1 + \frac{h^2}{12} \lambda_{\min} (S^{-1}C) + O(h^3) \qquad (2.26)$$

The asymptotic formulas of Theorem 2.2 and its corollaries essentially provides the solution of the stated problem when F, G, and Q are constant, and they can be used to estimate the sampling rates required for a particular application. The results can be utilized in different ways depending on the way the computations are arranged.

The asymptotic formulas can be evaluated. If only the discrete problem is solved, we can repeat the solution for different values of h. Knowing that the asymptotic behavior is $O(h^2)$, we can then estimate the difference by a Richardson extrapolation. Some examples of these applications are given in the following section.

We notice from the proof of Theorem 2.2 that the asymptotic formulas hold whenever

$$\sum (t+h;t)$$

can be approximated with sufficient accuracy by a few terms of its series expansion, which apparently is true if

$$h \|A\| < 1$$

This observation is useful in order to find the magnitudes of h for which the asymptotic formulas are valid. We will also use $\|A\|$ to normalize h.

III. EXAMPLES:

We will now consider some examples which illustrate the application of the results of the previous section.

Example 1.

Consider the first order system

$$\dot{x} = u \tag{3.1}$$

with the criterion

$$V(x_o, u) = \int_o^T [q_1 x^2(t) + q_2 u^2(t)]\, dt + q_o x^2(T) \tag{3.2}$$

when (a) u belongs to the class of continuous functions $u \in U_c$ and (b) u belongs to the class of piecewise constant functions $u \in U_d$. To fix the ideas, it is assumed in (b) that $u(t)$ is constant on the semi-open intervals $[kh, (k+1)h)$, $k = 1, \ldots, n-1$, $nh = T$.

We have

$$A = \begin{pmatrix} 0 & -q_2^{-1} \\ -q_1 & 0 \end{pmatrix} \quad , \quad B = \begin{pmatrix} 0 & 0 \\ -q_1^2 q_2^{-1} & 0 \end{pmatrix}$$

The solution of the canonical equation (2.16) becomes

$$\sum(t, T) = \begin{pmatrix} \cosh a & \frac{\sinh a}{\sqrt{q_1 q_2}} \\ \sqrt{q_1 q_2} \sinh a & \cosh a \end{pmatrix} \tag{3.3}$$

where

$$a = \sqrt{\frac{q_1}{q_2}} (T-t) \tag{3.4}$$

The minimal value of the functional (3.2) in case (a) is

$$\min_{u \in U_c} V(x_o, u) = x_o^2 \, S(0,T) \quad ,$$

where

$$S(t;T) = \sqrt{q_1 q_2} \cdot \frac{q_o + \sqrt{q_1 q_2} \quad \tanh \alpha}{\sqrt{q_1 q_2} \; + \; q_o \tanh \alpha}$$

Now consider the discrete case. We get

$$\Phi = 1$$

$$\Gamma = h$$

$$Q_{11} = q_1 h$$

$$Q_{12} = \frac{1}{2} q_1 h^2$$

$$Q_{22} = q_2 h + \frac{1}{3} q_1 h^3$$

and the minimal value of (3.2) is

$$\min_{u \in U_d} V(x_o, u) = x_o^2 \, \tilde{S}(0)$$

where

$$\tilde{S}(t-h) = \tilde{S}(t) + q_1 h - \frac{[h\tilde{S}(t) + \frac{1}{2} q_1 h^2]^2}{[h^2 \tilde{S}(t) + q_2 h + \frac{1}{3} q_1 h^3]}$$

$$\tilde{S}(T) = q_o$$

Using the asymptotic formulas (2.15) and (2.17) for $\tilde{S}(t)$ we get

$$C = \frac{q_1}{2q_2}\sqrt{q_1 q_2} \cdot \frac{q_1 q_2 (\sinh \alpha \cosh \alpha - \alpha) + 2q_0\sqrt{q_1 q_2}\,\sinh^2 \alpha + q_0^2 (\sinh \alpha \cosh \alpha - \alpha)}{q_1 q_2 \cdot \cosh^2 \alpha + 2q_0 \sqrt{q_1 q_2}\,\sinh \alpha \cosh \alpha + q_0^2 \sinh^2 \alpha}$$

If the interval $(0, T)$, over which the optimization is performed, is increased to infinity, we get

$$C = \frac{q_1}{2q_2}\sqrt{q_1 q_2}$$

and the asymptotic formula gives

$$\tilde{S}(t, \infty) = \sqrt{q_1 q_2}\ (1 + \frac{h^2}{24}\ \frac{q_1}{q_2}) + O(h^3)$$

The exact value of $\tilde{S}(t, \infty)$ is

$$\tilde{S}(t, \infty) = \sqrt{q_1 q_2}\quad \sqrt{1 + \frac{1}{12}\ \frac{q_1}{q_2}\ h^2}$$

To judge the influence of the sampling rate on the criterion, we have, in the table below, given the relative increase of the loss function for different values of the sampling rate. To demonstrate the asymptotic formula, we have also in the second column given the corresponding values calculated from the asymptotic formula.

$h\sqrt{\frac{q_1}{q_2}}$	$\frac{S-S}{S}$	$h^2 \frac{C}{12S}$
0.1	.00041658	.00041667
0.2	.0016653	.00016667
0.5	.010363	.010416
1.0	.04043	.04125
2.0	.1547	.1667
5.0	.756	1.042
10.0	2.05	4.17

The asymptotic formula will give results correct within one percent if

$$h\sqrt{\frac{q_1}{q_2}} < 0.7$$

Notice that $\sqrt{\frac{q_1}{q_2}}$ is the magnitude of the largest eigenvalue of the A-matrix.

Example 2.

Consider the system

$$\dot{x} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} x + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u$$

with the criterion

$$\int_0^T [q_1 x_1^2(t) + q_2 x_2^2(t) + r u^2(t)] \, dt \quad , \quad r > 0, q_1 > 0, q_2 > 0$$

when: (a) u belongs to the class of continuous functions U_c, and (b) u belongs to the class of piecewise constant functions U_d. To fix the ideas, it is assumed in (b) that $T = nh$ and that u is constant on the semi-open intervals $[kh, (k+1)h)$, $k = 1, \ldots, n-1$. As $r > 0$, both problems are regular and there exists a unique solution.

We have

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & -r^{-1} \\ -q_1 & 0 & 0 & 0 \\ 0 & -q_2 & -1 & 0 \end{pmatrix}$$

$$B = \begin{pmatrix} 0 & r^{-1}q_2 & r^{-1} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -r^{-1}q_2^2 & -r^{-1}q_2 & 0 \end{pmatrix}$$

The eigenvalues of A are given by

$$r\lambda^4 - q_2\lambda^2 + q_1 = 0$$

Letting $T \to \infty$, we get

$$S(t, \infty) = \begin{pmatrix} \sqrt{q_1 q_2} + 2q_1\sqrt{r q_1} & \sqrt{r q_1} \\ \sqrt{r q_1} & \sqrt{r q_2} + 2r\sqrt{r q_1} \end{pmatrix}$$

20

The integrals (B.8) through (B.11), giving the transformation to the discrete problem, can be evaluated in closed form as follows:

$$\Phi = \begin{pmatrix} 1 & h \\ 0 & 1 \end{pmatrix}$$

$$\Gamma = \begin{pmatrix} \frac{1}{2} h^2 \\ h \end{pmatrix}$$

$$\tilde{Q}_{11} = \begin{pmatrix} q_1 h & \frac{1}{2} q_1 h^2 \\ \frac{1}{2} q_1 h^2 & q_2 h + \frac{1}{3} q_1 h^3 \end{pmatrix}$$

$$\tilde{Q}_{12} = \begin{pmatrix} \frac{1}{2} q_1 h^2 \\ \frac{1}{2} q_1 h^3 + q_2 h \end{pmatrix}$$

$$\tilde{Q}_{22} = \frac{1}{20} q_1 h^5 + \frac{1}{3} q_2 h^3 + r h$$

To find the influence of the sampling rate, we have calculated $\lambda_{max}(TS^{-1})$ and $\lambda_{min}(TS^{-1})$ for the case $q_1 = q_2 = r = 1$, $T = 10$. The results are summarized in Figure 1. Notice that for this choise of parameters, the eigenvalues of the A-matrix will all have the magnitude 1. The figure shows, for example, that a sampling rate of 0.3 will give an increase of the loss function of 1% at most, while a sampling interval of 1.0 may give a 10% increase of the loss function.

Example 3.

Consider the problem of Example 2, but assume in the discrete problem that the minimum is taken with respect to piecewise linear functions.

We handle this situation by transforming it to the case of piecewise constant controls in the following way. Introduce

$$\begin{aligned} u &= x_3 + u_1 \\ \dot{x}_3 &= u_2 \end{aligned}$$

By assuming that u_1 and u_2 are piecewise constant, we will thus achieve that u is piecewise linear. We adjoin the variable x_3 to the state variables x_1 and x_2 and obtain

$$\begin{aligned} \dot{x}_1 &= x_2 \\ \dot{x}_2 &= x_3 + u_1 \\ \dot{x}_3 &= u_2 \end{aligned}$$

and the criterion becomes

$$\int_0^T [q_1 x_1^2(t) + q_2 x_2^2(t) + r x_3^2(t) + 2 r x_3(t) u_1 + r u_1^2] \, dt$$

The problem is thus in standard form with

$$F = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \qquad G = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$Q_{11} = \begin{pmatrix} q_1 & 0 & 0 \\ 0 & q_2 & 0 \\ 0 & 0 & r \end{pmatrix} \qquad Q_{12} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ r & 0 \end{pmatrix} \qquad Q_{22} = \begin{pmatrix} r & 0 \\ 0 & 0 \end{pmatrix}$$

In this case, the transformation to the discrete problem can also be carried out analytically, and we get

$$\Phi = \begin{pmatrix} 1 & h & \frac{1}{2} h^2 \\ 0 & 1 & h \\ 0 & 0 & 1 \end{pmatrix} \qquad \Gamma = \begin{pmatrix} \frac{1}{2} h^2 & \frac{1}{6} h^3 \\ h & \frac{1}{2} h^2 \\ 0 & h \end{pmatrix}$$

$$\tilde{Q}_{11} = \begin{pmatrix} q_1 h & \frac{1}{2} q_1 h^2 & \frac{1}{6} q_1 h^3 \\ \frac{1}{2} q_1 h^2 & \frac{1}{3} q_1 h^3 + q_2 h & \frac{1}{8} q_1 h^4 + \frac{1}{2} q_2 h^2 \\ \frac{1}{6} q_1 h^3 & \frac{1}{8} q_1 h^4 + \frac{1}{2} q_2 h^2 & \frac{1}{20} q_1 h^5 + \frac{1}{3} q_1 h^3 + r h \end{pmatrix}$$

$$\tilde{Q}_{12} = \begin{pmatrix} \frac{1}{6} q_1 h^3 & \frac{1}{24} q_1 h^4 \\ \frac{1}{8} q_1 h^4 + \frac{1}{2} q_2 h^2 & \frac{1}{30} q_1 h^5 + \frac{1}{6} q_2 h^3 \\ \frac{1}{20} q_1 h^5 + \frac{1}{3} q_2 h^3 + r h & \frac{1}{72} q_1 h^6 + \frac{1}{8} q_1 h^4 + \frac{1}{2} r h^2 \end{pmatrix}$$

$$\tilde{Q}_{22} = \begin{pmatrix} \frac{1}{40} q_1 h^5 + \frac{1}{3} q_2 h^3 + r h & \frac{1}{72} q_1 h^6 + \frac{1}{8} q_2 h^4 + \frac{1}{2} r h^2 \\ \frac{1}{72} q_1 h^6 + \frac{1}{8} q_2 h^4 + \frac{1}{2} r h^2 & \frac{1}{252} q_1 h^7 + \frac{1}{20} q_2 h^5 + \frac{1}{3} r h^3 \end{pmatrix}$$

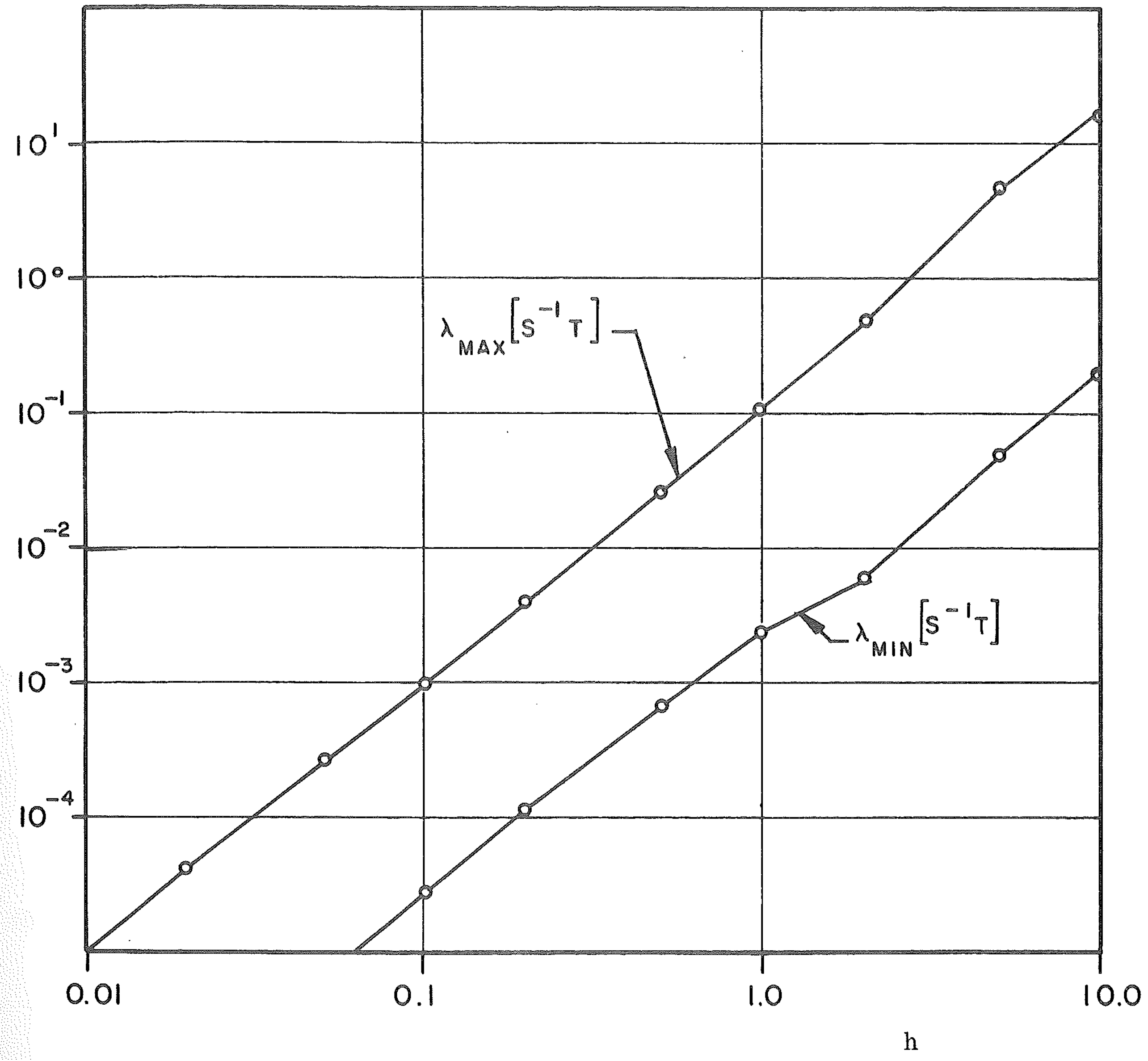

10^{1}
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
$\lambda_{MAX}[S^{-1}T]$
$\lambda_{MIN}[S^{-1}T]$
0.01
0.1
1.0
10.0
h

FIGURE 1.

As Q_{22} is singular, the continuous problem is not regular, and Theorem 2.2 does not apply. It is easily seen, however, that there are positive h such that Q_{22} is positive definite, which means that the discrete problem is regular for the values of h.

For $q_1 = q_2 = r = 1$, we find that Q_{22} is positive definite, at least for $0 < h \leq 5$, and the discrete problem is thus regular for h in this range. In Figure 2, we have plotted $\lambda_{\max}(TS^{-1})$. Also, for the purpose of comparison, we have shown in the same figure $\lambda_{\max}(TS^{-1})$ for the case of piecewise constant functions. Compare Example 2.

To have at the most a 1% increase of the loss function, we find that the sampling interval has to be smaller than $h = 1.6$, compared to $h = 0.3$ in the case of piecewise constant functions. Also notice that this problem is degenerate in the sense that the second-order term in h vanishes and that we have

$$\tilde{S}(t) = S(t) + h^3 D + O(h^4)$$

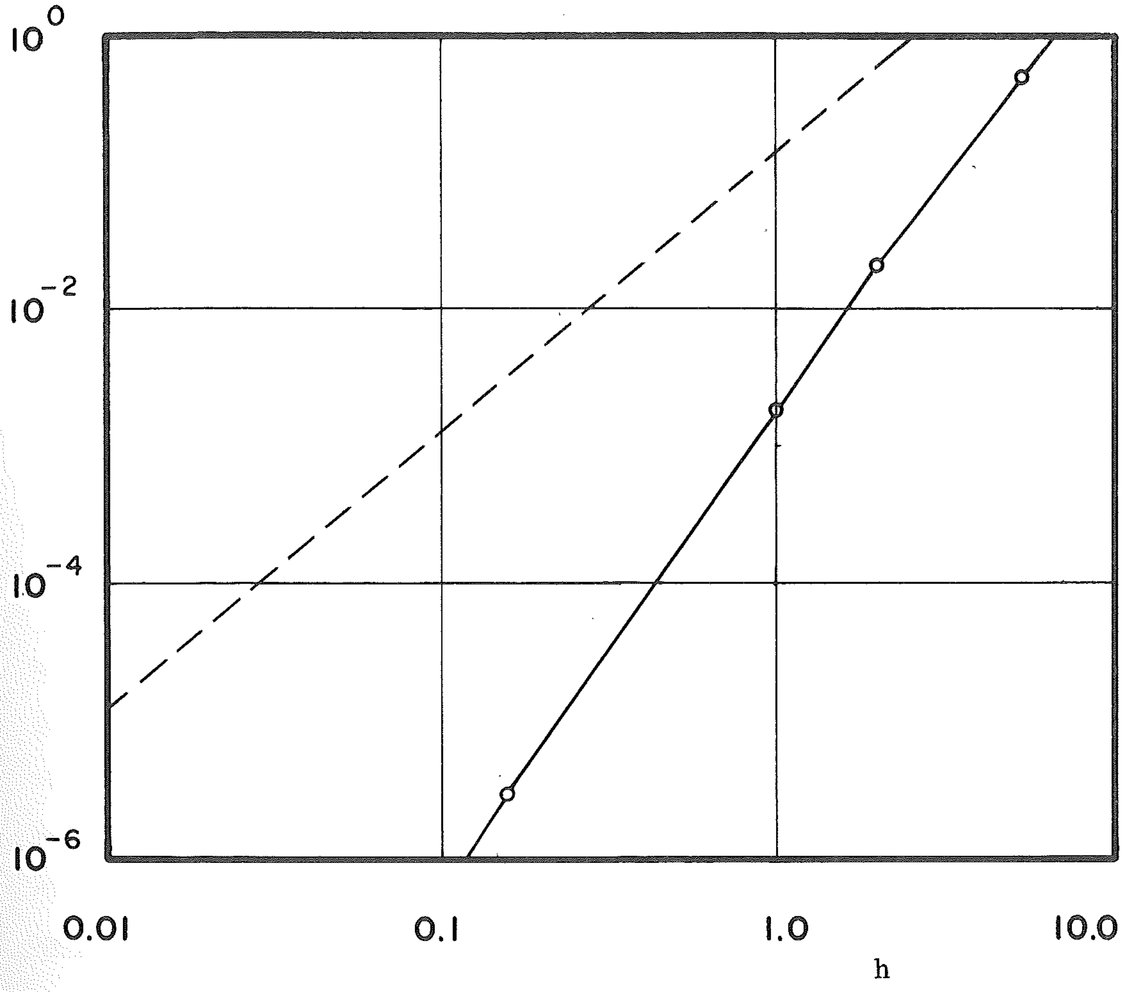


FIGURE 2.

IV. THE INFLUENCE OF THE SAMPLING RATE IN THE DISCRETE ESTIMATION PROBLEM:

We will now consider the influence of the sampling rate in the implementation of the solution of the minimum mean square estimation problem.

Consider a system described by

$$\frac{dx}{dt} = F(t)\,x + v_1 \tag{4.1}$$

$$y = H(t)\,x + v_2 \tag{4.2}$$

where $x(t)$ and $y(t)$ are n and p vectors referred to as the state vector and the output signal, respectively. The variables $v_1(t)$ and $v_2(t)$ are vector valued second-order random functions with zero average and the covariance functions

$$\operatorname{cov} v_1(t)\, v_1^T(t+\tau) = R_{11}(t)\,\delta(\tau) \tag{4.3}$$

$$\operatorname{cov} v_1(t)\, v_2^T(t+\tau) = R_{12}(t)\,\delta(\tau) \tag{4.4}$$

$$\operatorname{cov} v_2(t)\, v_2(t+\tau) = R_{22}(t)\,\delta(\tau) \tag{4.5}$$

where $\delta(\tau)$ is the Dirac measure and the initial state is a random variable with

$$E[x(0)] = m$$

$$\operatorname{cov} x(0)\, x^T(0) = R_o \tag{4.6}$$

The matrices $F(t)$ and $H(t)$ are assumed to be continuous and bounded. Consider the following:

PROBLEM:

Given the observations of the output $y(t)$ over the interval $(0, t)$ find the minimum mean square estimate, $\hat{x}(t)$, of the state vector $x(t)$ when (a) the estimate is formed by operating continuously on the observations, and (b) the estimate is formed by operating on the observations only at discrete instants of time. Find the difference between the estimation errors of Case (a) and Case (b).

Case (a) is called the continuous problem. Case (b), called the discrete problem, refers to the situation when the solution of the estimation problem is implemented by a data processing equipment operating sequentially in time. To fix the ideas, we will in Case (b) assume that the data processing equipment produces an estimate $\hat{x}(t_k)$ of the state variable $\hat{x}(t)$ at discrete instants of time

$$t_n = t_o + \sum_{i=0}^{n-1} \tau_i$$

where τ_i are fixed numbers referred to as the sampling intervals. The data to be processed by the computer is fed from the measuring equipment via sample and hold circuits. Let the components of the p-vector $z(t)$ be the numbers obtained in the computer registers after these operations. We will assume that $z(t)$ is related to the output signals $y(t)$ by the linear transformation

$$z(t_k) = \int_{t_{k-1}}^{t_k} k(t_k;s)\, y(s)\, ds \qquad (4.7)$$

Different kernels $k(t, s)$ correspond to different methods of sampling. For example:

1. "Inpulse Sampling"

We have

$$k(t_k;t) = \delta(t_k-t)$$

and we get

$$z(t_k) = y(t_k) \tag{4.8}$$

2. "Average Sampling"

We have

$$k(t_k;t) = 1$$

and we get

$$z(t_k) = \int_{t_{k-1}}^{t_k} y(s)\, ds \tag{4.9}$$

For a more detailed discussion of the sampling process, we refer to Ragazzini and Franklin.[10] In this section, we will arbitrarily assume that $z(t_k)$ is formed by (4.9), i.e., average sampling. This assumption is not essential. One reason for this choice is that it leads to the formal dual of the optimal control problem discussed in the previous section, where the corresponding assumption was that the control signal is constant over the sampling intervals.

The solutions of the discrete and the continuous estimation problems are well known; the minimum mean square estimation errors are given by

$$P(t) = E\{[x(t) - \hat{x}(t)]\ [x(t) - \hat{x}(t)]^T\} \tag{4.10}$$

and

$$\tilde{P}(t_k) = E\{[x(t_k) - \hat{x}(t_k)]\ [x(t_k) - \hat{x}(t_k)]^T\} \tag{4.11}$$

respectively.

Analytic expressions for the matrices $P(t)$ and $\tilde{P}(t)$ in terms of F, G, and B are available. See Kalman.[7] As we are going to make extensive use of them, they are included in Appendices C and D. For the purpose of comparing P and $\tilde{P}$, we introduce the norm

$$\| P \|^2 = \max_{\| x \| = 1} x^T P x \tag{4.12}$$

It immediately follows from the problem statement that

$$\| \tilde{P}(t_k) \|^2 \geq \| P(t_k) \|^2 \tag{4.13}$$

The estimation problem of this section is the formal dual of the control problem of Section II. Using the dual transformation,

$$\begin{aligned} t^* &= -t \\ F^*(t^*) &= F^T(t) \\ H^*(t^*) &= G^T(t) \\ R^*(t^*) &= Q(t) \\ P(t^*) &= S(t) \end{aligned} \tag{4.14}$$

the results of Section II can immediately be used and we obtain

<u>THEOREM 4.1:</u>

$$\tilde{P}(t_k;t_o) \rightarrow P(t_k;t_o) \quad \text{as} \quad \max_i(\tau_i) \rightarrow 0 \tag{4.15}$$

In the case of stationary systems, we have the following asymptotic estimate of the difference between $\tilde{P}$ and P, THEOREM 4.2:

Let the matrices F, G, and R be stationary, let $R_{12}=0$, let R_{22} be positive definite, and let all sampling intervals have the same length, $\tau_i = h$, then

$$P(t;t_o) = [\Lambda_{21}(t;t_o) + \Lambda_{22}(t;t_o)R_o][\Lambda_{11}(t;t_o) + \Lambda_{12}(t;t_o)R_o]^{-1} \tag{4.16}$$

$$\tilde{P}(t;t_o) = [\tilde{\Lambda}_{21}(t;t_o) + \tilde{\Lambda}_{22}(t;t_o)R_o][\tilde{\Lambda}_{11}(t;t_o) + \tilde{\Lambda}_{12}(t;t_o)R_o]^{-1} \tag{4.17}$$

where

$$\frac{d}{dt}\Lambda(t;t_o) = A^*\Lambda(t;t_o) \quad , \quad \Lambda(t_o;t_o) = I \tag{4.18}$$

$$\tilde{\Lambda}(t;t_o) = \Lambda(t;t_o) + \frac{h^2}{12}E^*(t;t_o) + O(h^3) \tag{4.19}$$

$$\frac{d}{dt}E^*(t;t_o) = A^*E(t;t_o) + B^*\Lambda(t;t_o) \quad , \quad E(t_o;t_o) = 0 \tag{4.20}$$

$$A^* = \begin{pmatrix} -F^T & H^T R_{22}^{-1} H \\ R_{11} & F \end{pmatrix} \tag{4.21}$$

$$B^* = \begin{pmatrix} -F^T H^T R_{22}^{-1} H R_{11} & -F^T H^T R_{22}^{-1} H F \\ R_{11} H^T R_{22}^{-1} H R_{11} & R_{11} H^T R_{22}^{-1} H F \end{pmatrix} \tag{4.22}$$

Again we notice that the assumptions of constant sampling intervals and $R_{12} = 0$ are not essential.

Further, introduce the matrix $C^*(t;t_o)$ defined by

$$C^* = \{E^*_{21} + E^*_{22}R_o - P[E^*_{11} + E_{12}R_o]\}[\Lambda_{11} + \Lambda_{12}R_o]^{-1} \qquad (4.23)$$

where the arguments $t;t_o$ are temporarily dropped in C, D, E_{ij} and Λ_{ij}. It now follows from Theorem 4.2,

<u>COROLLARY 4.3:</u>

$$\tilde{P}(t;t_o) = P(t;t_o) + \frac{h^2}{12} C^*(t;t_o) + O(h^3) \qquad (4.24)$$

<u>COROLLARY 4.4:</u>

$$\max_x \frac{x^T\tilde{P}x}{x^TPx} = 1 + \frac{h^2}{12} \lambda_{max}(P^{-1}C^*) + O(h^3) \qquad (4.25)$$

$$\min_x \frac{x^T\tilde{P}x}{x^TPx} = 1 + \frac{h^2}{12} \lambda_{min}(P^{-1} C^*) + O(h^3) \qquad (4.26)$$

APPENDIX A

The Continuous Control Problem:

The continuous problem, i.e., the minimization of the functional (2.3) with respect to the class of piecewise continuous functions, is a classical variational problem. For a detailed discussion, including proofs and control theoretic interpretations, we refer to Kalman.[6] In this appendix we will summarize the main results used in this paper.

The Hamiltonian of the variational problem is

$$2H(x,p,t,u) = x^T Q_{11} x + 2x^T Q_{12} u + u^T Q_{22} u + 2p^T (Fx+Gu) \quad . \tag{A.1}$$

The Hamiltonian is minimized for $u = u^o$ where

$$Q_{21} x + Q_{22} u^o + G^T p = 0 \quad . \tag{A.2}$$

If Q_{22} is positive definite, the control u^o which minimizes the Hamiltonian is uniquely given by

$$u^o(x,p,t) = -Q_{22}^{-1} [Q_{21} x + G^T p] \tag{A.3}$$

The condition that Q_{22} is positive definite is the regularity condition for the variational problem. Notice that in (A.3) the optimal control signal is defined in terms of the values of the state vector x and the canonical coordinate p, which means that (A.3) in fact defines a control law or a feedback solution.

The minimal value of the Hamiltonian is given by

$$2H^{o}(x,p,t) = 2 \min_{u} H(x,p,t,u) =$$

$$= x^{T}[Q_{11}-Q_{12}Q_{22}^{-1}Q_{21}]x + 2p^{T}[F-GQ_{22}^{-1}Q_{21}]x - p^{T}GQ_{22}^{-1}G^{T}p \quad . \tag{A.4}$$

The Hamilton-Jacobi equation is

$$V_{t}^{o} + H^{o}(x, V_{x}, t) = 0 \quad , \tag{A.5}$$

where $V^{o}(x,t;t_1)$ is the minimal value of the loss function, i.e.,

$$V^{o}(x,t;t_1) = \min_{u \in U} V(x,t,t_1,u) \quad . \tag{A.6}$$

The Hamilton-Jacobi equation has the solution

$$V^{o}(x,t;t_1) = x_{o}^{T} S(t;t_1) x_{o} \quad , \tag{A.7}$$

where $S(t;t_1)$ is a symmetric matrix which satisfies the Riccati equation

$$\frac{dS}{dt} + [F-GQ_{22}^{-1}Q_{21}]^{T}S + S[F-GQ_{22}^{-1}Q_{21}] - SGQ_{22}^{-1}G^{T}S$$
$$+ Q_{11} - \overset{Q_{12}}{\wedge} Q_{\underset{2}{\cancel{1}}2}^{-1}Q_{21} = 0 \quad , \tag{A.8}$$

with the boundary condition

$$S(t_1 ; t_1) = Q_{o} \quad . \tag{A.9}$$

Conditions for the existence of a solution of (A.8) for $t_{o} \leq t \leq t_1$ are found in reference 6.

The canonical equations, or Euler's equations, whose solutions are the characteristics of the Hamilton-Jacobi equation, are

$$\dot{x} = [F-GQ_{22}{}^{-1}Q_{21}]x - GQ_{22}{}^{-1}G^T p \quad ,$$

$$\dot{p} = [Q_{11}-Q_{12}Q_{22}{}^{-1}Q_{21}]x - [F-GQ_{22}{}^{-1}Q_{21}]^T p \quad , \qquad (A.10)$$

with the boundary conditions

$$x(t_o) = x_o \quad , \qquad (A.11)$$

$$p(t_1) = Q_o x(t_1) \quad . \qquad (A.12)$$

Let

$$\sum(t;t_1) = \begin{pmatrix} \Sigma_{11}(t;t_1) & \Sigma_{12}(t;t_1) \\ \Sigma_{21}(t;t_1) & \Sigma_{22}(t;t_1) \end{pmatrix} \qquad (A.13)$$

be the fundamental solution to (A.10). Using the boundary condition on p(t), we get

$$p(t) = S(t;t_1)x(t) \quad , \qquad (A.14)$$

where

$$S(t;t_1) = [\Sigma_{21}(t;t\ ) + \Sigma_{22}(t;t_1)Q_o]\ [\Sigma_{11}(t;t_1) + \Sigma_{12}(t;t_1)Q_o]^{-1} \ . \qquad (A.15)$$

The matrix $S(t;t_1)$ satisfies the Riccati equation (A.8) with the boundary condition (A.9).

Using (A.14), the control law (A. 3) becomes

$$u^{o}(x, p, t) = -L(t)\, x(t) \quad , \tag{A. 16}$$

where

$$L(t) = Q_{22}^{-1}(t)\, [Q_{21}(t) + G^{T}(t)\, S(t;t_{1})] \quad , \tag{A. 17}$$

The equation of motion of the optimal system is

$$\dot{x}(t) = [F(t) - G(t)\, L(t)]\, x(t) \quad . \tag{A. 18}$$

APPENDIX B

The Discrete Control Problem:

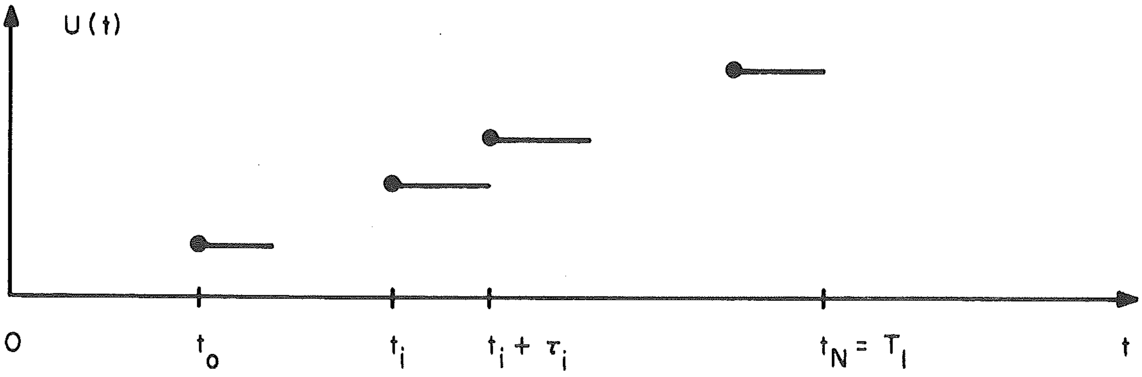
Now consider the discrete problem, i.e., the minimization of the functional (2.3) with respect to control functions, u(t), which are piecewise constant. This problem was first solved by Kalman and Koepcke[8] using Bellman's technique of dynamic programming[1] and has since been subject to extensive treatment. In the following, we will state the main results.

It is assumed that the time $t_1 = T_1$ is fixed and that the interval $[t_o, T_1]$ is divided into subintervals $(t_i, t_i + \tau_i)$ by the division points

$$t_k = t_o + \sum_{i=0}^{k-1} \tau_i \quad , \quad t_N = T_1 \tag{B.1}$$

where $\{\tau_i\}$ is a given sequence of numbers called the sampling intervals.

The control functions u(t) are assumed to be constant over the semi-open intervals $[t_i, t_i + \tau_i)$. See Figure B.1.



Utilizing the fact that u(t) is constant over the sampling intervals, the equation (2.1) describing the motion of the dynamical system can be integrated, and we get

$$x(t) = \Phi(t;t_k)\, x(t_k) + \Gamma(t;t_k)\, u(t_k) \quad , \quad t_k \leq t \leq t_{k+1} \tag{B.2}$$

where

$$\frac{d}{dt}\, \Phi(t;t_k) = F(t)\, \Phi(t;t_k) \quad , \quad t_k \leq t \leq t_{k+1} \tag{B.3}$$

$$\Phi(t_k;t_k) = I \tag{B.4}$$

and

$$\Gamma(t;t_k) = \int_{t_k}^{t} \Phi(t;s)\, G(s)\, ds \quad , \quad t_k \leq t \leq t_{k+1} \tag{B.5}$$

The functional (2.3) can now be expressed as

$$\tilde{V}(t_o, x_o, T_1, u) = x^T(t_N)\, Q_o\, x(t_N) + \sum_{k=0}^{N-1} \begin{pmatrix} x(t_k) \\ u(t_k) \end{pmatrix}^T \tilde{Q}(t_k) \begin{pmatrix} x(t_k) \\ u(t_k) \end{pmatrix} \tag{B.6}$$

where

$$\tilde{Q}(t_k) = \begin{pmatrix} \tilde{Q}_{11}(t_k) & \tilde{Q}_{12}(t_k) \\ \tilde{Q}_{21}(t_k) & \tilde{Q}_{22}(t_k) \end{pmatrix} \tag{B.7}$$

and

$$\tilde{Q}_{11}(t_k) = \int_{t_k}^{t_{k+1}} \Phi^T(s;t_k)\, Q_{11}(s)\, \Phi(s;t_k)\, ds \tag{B.8}$$

$$\tilde{Q}_{12}(t_k) = \int_{t_k}^{t_{k+1}} [\Phi^T(s;t_k)\, Q_{11}(s)\, \Gamma(s;t_k) + \Phi^T(s;t_k)\, Q_{12}(s)]\, ds \tag{B.9}$$

$$\tilde{Q}_{21}(t_k) = Q_{12}^T(t_k) \tag{B.10}$$

$$\tilde{Q}_{22}(t_k) = \int_{t_k}^{t_{k+1}} [\Gamma^T(s;t_k)\, Q_{11}(s)\, \Gamma(s;t_k) + \Gamma^T(s;t_k)\, Q_{12}(s) + Q_{21}(s)\Gamma(s;t_k) + Q_{22}(s)]\, ds \tag{B.11}$$

The discrete problem can now be stated as

> <u>PROBLEM:</u> Given the discrete time dynamical system described by the difference equation (B.2), find the minimal value of the functional (B.6) and the sequence of controls for which the minimal is attained.

The minimal value of the functional (B.6) is

$$\tilde{V}(x_o, t_o; T_1) = x_o^T\, \tilde{S}(t_o, T_1)\, x_o \tag{B.12}$$

where the symmetric matrix $S(t_o, T_1)$ is given by the recursive equation

$$\tilde{S}(t_k;T_1) = \Phi^T \tilde{S}(t_{k+1};T_1)\Phi - \tilde{L}^T[\Gamma^T \tilde{S}(t_{k+1};T_1)\Gamma + \tilde{Q}_{22}]\tilde{L} + \tilde{Q}_{11}$$

$$= \Psi^T \tilde{S}(t_{k+1};T_1)\Psi + \tilde{Q}_{11} - \tilde{L}^T \tilde{Q}_{21} - \tilde{Q}_{12}\tilde{L} + \tilde{L}^T \tilde{Q}_{22}\tilde{L}$$

$$= \Psi^T \tilde{S}(t_{k+1};T_1)\Phi + \tilde{Q}_{11} - \tilde{L}^T \tilde{Q}_{21} \qquad \text{(B. 13)}$$

and

$$\tilde{L} = [\Gamma^T \tilde{S}(t_{k+1};T_1)\Gamma + \tilde{Q}_{\underset{22}{\cancel{11}}}]^{-1} \; [\Gamma^T \tilde{S}(t_{k+1};T_1)\Phi + \tilde{Q}_{21}]$$

$$= \tilde{Q}_{22}^{-1}[\Gamma^T \tilde{S}(t_{k+1};T_1)\Psi + \tilde{Q}_{21}] \qquad \text{(B. 14)}$$

$$\Psi = \Phi - \Gamma\tilde{L} \qquad \text{(B. 15)}$$

The initial condition is

$$\tilde{S}(T_1, T_1) = Q_o \qquad \text{(B. 16)}$$

The canonical equations are

$$\left\{\begin{aligned} x(t_{k+1}) &= [\Phi - \Gamma\, \tilde{Q}_{22}^{-1}\, \tilde{Q}_{21}]\, x(t_k) - \Gamma\, \tilde{Q}_{22}^{-1}\, \Gamma^{T}\, p(t_k) \\ p(t_{k-1}) &= [\tilde{Q}_{11} - \tilde{Q}_{12}\, \tilde{Q}_{22}^{-1}\, \tilde{Q}_{21}]\, x(t_k) + [\Phi - \Gamma\, \tilde{Q}_{22}^{-1}\, \tilde{Q}_{21}]^{T}\, p(t_k) \end{aligned}\right. \tag{B.17}$$

with the boundary conditions

$$x(t_o) = x(t_o) \quad , \tag{B.18}$$

$$p(t_{N-1}) = Q_o\, x(t_N) \quad . \tag{B.19}$$

In the formulas above, the arguments of the functions are as follows:

$$\begin{aligned} \Phi &= \Phi(t_{k+1};t_k) \\ \Psi &= \Psi(t_{k+1};t_k) \\ \Gamma &= \Gamma\,(t_{k+1};t_k) \\ \tilde{L} &= \tilde{L}(t_{k\cancel{+1}}) \\ \tilde{Q}_{11} &= \tilde{Q}_{11}\,(t_k) \\ \tilde{Q}_{12} &= \tilde{Q}_{12}\,(t_k) \\ \tilde{Q}_{22} &= \tilde{Q}_{22}\,(t_k) \end{aligned}$$

Let

$$\tilde{\sum}\,(t_k;t_N) = \begin{pmatrix} \tilde{\Sigma}_{11}\,(t_k;t_N) & \tilde{\Sigma}_{12}\,(t_k;t_N) \\ \tilde{\Sigma}_{21}\,(t_k;t_N) & \tilde{\Sigma}_{22}\,(t_k;t_N) \end{pmatrix} \tag{B.20}$$

be the fundamental solution of (B.17) with

$$\tilde{\Sigma}(t_N;t_N) = I$$

then,

$$p(t_{k-1}) = \tilde{S}(t_k;t_N)\, x(t_k) \qquad (B.21)$$

where

$$\tilde{S}(t_k;t_N) = [\tilde{\Sigma}_{21}(t_k;t_N) + \tilde{\Sigma}_{22}(t_k;t_N)Q_o]\,[\tilde{\Sigma}_{11}(t_k;t_N) + \tilde{\Sigma}_{12}(t_k;t_N)Q_o]^{-1} \qquad (B.22)$$

If $\tilde{Q}_{22}$ and $\tilde{S}(t_k;t_N)$ are positive definite, then $\tilde{S}(t_k;t_N)$ satisfies (B.13) with the initial condition (B. 16) which explains the notation (B.21).

The minimum of the loss function is assumed for the control law

$$u(t_k) = -\tilde{L}(t_k)\, x(t_k) \qquad (B.23)$$

and the equation of motion of the optimal system is

$$\hat{x}(t_{k+1}) = \Psi(t_{k+1};t_k)\, \hat{x}(t_k) \qquad (B.24)$$

The regularity condition which assures a unique control law is that the matrix

$$\Gamma^T \tilde{S}(t_{k+1};T_1)\Gamma + \tilde{Q}_{22} \qquad (B.25)$$

is positive definite for all k. Notice, however, that the minimum might exist even if this is not the case. One control law yielding the minimal value to the loss function might be obtained by substituting the inverse in (B.14) by a generalized inverse. See Penrose.[9]

Also notice that the regularity of the continuous problem implies the regularity of the discrete problem, but that the converse statement is not true. By the limitation of the class of admissible controls, the conditions for the regularity of the variational problem have been considerably weakened. If $Q_{22} = 0$, the continuous problem is irregular, but the discrete problem might well be regular.

APPENDIX C

The Continuous Estimation Problem:

Consider the system described by (4.1) and (4.2). The minimum mean square estimate $\hat{x}(t)$ of the state variable $x(t)$ is given by

$$\frac{d}{dt}\,\hat{x}(t) = F(t)\,\hat{x}(t) + K(t)\,[y(t) - \hat{y}(t)] \qquad (C.1)$$

$$\hat{x}(0) = m \qquad (C.2)$$

$$\hat{y}(t) = H(t)\,\hat{x}(t) \qquad (C.3)$$

$$K(t) = [P(t;t_o)\,H(t) + R_{12}(t)]\,R_{22}^{-1}(t) \qquad (C.4)$$

where $P(t;t_o)$ is the minimum mean square estimation error, i.e.,

$$P(t;t_o) = E[x(t) - \hat{x}(t)]\,[x(t) - \hat{x}(t)]^T \qquad (C.5)$$

The symmetric matrix $P(t;t_o)$ satisfies the Riccati equation

$$\frac{d}{dt}\,P = [\,F \overset{-}{(+)} R_{12}\,R_{22}^{-1}\,H^{\cancel{T}}]\,P + P[\,F \overset{-}{(+)} R_{12}\,R_{22}^{-1}\,H^{\cancel{T}}]^T -$$

$$- PH^T R_{22}^{-1} HP + R_{11} - R_{12}\,R_{22}^{-1}\,R_{21} \qquad (C.6)$$

with the initial condition

$$P(t_o, t_o) = R_o \qquad (C.7)$$

In (C. 7), the argument t is temporarily dropped in all the functions. Notice the formal similarity with (A. 8), from which we can conclude that the solution of (C. 6) can be represented as

$$P(t;t_o) = [\Lambda_{2\underset{1}{\cancel{2}}}(t;t_o) + \Lambda_{2\underset{2}{\cancel{1}}}(t;t_o)\, R_o]\, [\Lambda_{11}(t;t_o) + \Lambda_{12}(t;t_o)\, R_o]^{-1} \tag{C. 8}$$

where $\Lambda(t;t_o)$ is the fundamental solution of

$$\frac{d}{dt}\Lambda(t;t_o) = \begin{pmatrix} -[F \overset{-}{\cancel{+}} R_{12}\, R_{22}^{-1}\, H^{\cancel{T}}]^T & H^T R_{22}^{-1} H \\ [R_{11} - R_{12}\, R_{22}^{-1}\, R_{21}] & [F \overset{-}{\cancel{+}} R_{12}\, R_{22}^{-1}\, H^{\cancel{T}}] \end{pmatrix} \Lambda(t;t_o) \tag{C. 9}$$

APPENDIX D

The Discrete Estimation Problem:

Consider the discrete estimation problem, i.e., to produce an estimate of the state of (4.1) when we are only operating on the data at discrete intervals of time. To fix the idea, we assume that it is possible to operate on the data at the instants

$$t_k = t_o + \sum_{i=0}^{k-1} \tau_i \quad , \tag{D. 1}$$

where $\{\tau_i\}$ is a given sequence of numbers called the sampling intervals. We also assume "average sampling", i.e., the number produced in the data processing equipment as the result of the action of the sample and hold circuits is related to the output of the plant by

$$z(t_k) = \int_{t_{k-1}}^{t_k} y(t)\,dt \quad . \tag{D. 2}$$

Integrating the equations of motion (4.1) and (4.2), we get

$$x(t_{k+1}) = \Phi(t_{k+1};t_k)\,x(t_k) + e_1(t_k) \tag{D. 3}$$

$$z(t_k)_{+1} = \Theta(t_{k+1};t_k)\,x(t_k) + e_2(t_k) \tag{D. 4}$$

where

$$\frac{d}{dt}\,\Phi(t;t_k) = F(t)\,\Phi(t;t_k) \qquad t_k \leq t \leq t_{k+1} \tag{D. 5}$$

$$\Phi(t_k;t_k) = I \tag{D. 6}$$

$$\Theta(t_{k+1};t) = \int_t^{t_{k+1}} H(s)\,\Phi(s;t)\,ds \qquad t_k \leq t \leq t_{k+1} \tag{D. 7}$$

The variables $e_1(t_k)$ and $e_2(t_k)$ are second order vector valued random functions with zero means and the covariance functions

$$E \; e_1(t_k)\, e_1^T(t_s) = \tilde{R}_{11}(t_k)\, \delta_{ks} \tag{D.8}$$

$$E \; e_1(t_k)\, e_2^T(t_s) = \tilde{R}_{12}(t_k)\, \delta_{ks} \tag{D.9}$$

$$E \; e_2(t_k)\, e_2^T(t_s) = \tilde{R}_{22}(t_k)\, \delta_{ks} \tag{D.10}$$

where δ_{ks} is the Kronecker delta and

$$\tilde{R}_{11}(t_k) = \int_{t_k}^{t_{k+1}} \Phi(t_{k+1};s)\, R_{11}(s)\, \Phi^T(t_{k+1};s)\, ds \tag{D.11}$$

$$\tilde{R}_{12}(t_k) = \int_{t_k}^{t_{k+1}} [\Phi(t_{k+1};s)R_{11}(s)\,\Theta^T(t_{k+1};s) + \Phi(t_{k+1};s)R_{12}(s)]\, ds \tag{D.12}$$

$$\tilde{R}_{22}(t_k) = \int_{t_k}^{t_{k+1}} [\Theta(t_{k+1};s)R_{11}(s)\,\Theta^T(t_{k+1};s) + \Theta(t_{k+1};s)R_{12}(s) + + R_{21}(s)\,\Theta^T(t_{k+1};s) + R_{22}(s)]\, ds \tag{D.13}$$

$$\tilde{R}_{21}(t_k) = \tilde{R}_{12}^T(t_k) \tag{D.14}$$

To arrive at these results, we have to interchange the operations of calculating mathematical expectations and integration with respect to time.

After these preliminaries, the discrete problem can now be stated as follows:

PROBLEM:

Consider the discrete dynamical system (D.3) with the output signal (D.4). Given a sequence of observed outputs $y(t_o), y(t_1), \ldots, y(t_k)$, find the minimum mean square estimate of the state vector at time t_k.

The solution of this problem is given by Kalman.[7] The estimate is obtained from

$$\begin{aligned}\hat{x}(t_{k+1}) &= \Phi(t_{k+1};t_k)\,\hat{x}(t_k) + \tilde{K}(t_{k+1})\,[y(t_{k+1}) - \hat{y}(t_{k+1})] \\ &= \Psi(t_{k+1};t_k)\,\hat{x}(t_k) + \tilde{K}(t_{k+1})\,y(t_{k+1})\end{aligned}$$

where

$$\hat{y}(t_{k+1)} = \Theta(t_{k+1};t_k)\;\Phi(t_{k+1};t_k)\,\hat{x}(t_k)$$

$$\Psi(t_{k+1};t_k) = \Phi(t_{k+1};t_k) - \tilde{K}(t_{k+1})\,\Theta(t_{k+1};t_k)$$

and

$$\begin{aligned}\tilde{K}(t_{k+1}) &= [\Phi\tilde{P}(t_k;t_o)\,\Theta^T + \tilde{R}_{12}]\;[\Theta\tilde{P}(t_k;t_o)\Theta^T + \tilde{R}_{22}]^{-1} \\ &= [\Psi\tilde{P}(t_k;t_o) + \tilde{R}_{12}]\;\tilde{R}_{22}^{-1}\end{aligned}$$

$$\begin{aligned}\tilde{P}(t_{k+1};t_o) &= \Phi\,\tilde{P}(t_k;t_o)\Phi^T - \tilde{K}(t_{k+1})\,[\Theta\tilde{P}(t_k;t_o)\Theta^T + \tilde{R}_{22}]\,\tilde{K}^T(t_{k+1}) + \tilde{R}_{11} \\ &= \Psi\,\tilde{P}(t_k;t_o)\,\Phi^T - \tilde{K}(t_{k+1})\tilde{R}_{21} + \tilde{R}_{11} \\ &= \Psi\,\tilde{P}(t_k;t_o)\,\Psi^T + \tilde{R}_{11} - 2\tilde{K}(t_{k+1})\tilde{R}_{21} + \tilde{K}(t_{k+1})\tilde{R}_{22}\,\tilde{K}^T(t_{k+1})\end{aligned}$$

To abbreviate the writing, the arguments have been omitted in the formulas above. The arguments are as follows:

$$\Phi = \Phi(t_{k+1};t_k)$$

$$\Psi = \Psi(t_{k+1};t_k)$$

$$\Theta = \Theta(t_{k+1};t_k)$$

$$\tilde{R}_{11} = \tilde{R}_{11}(t_k)$$

$$\tilde{R}_{12} = \tilde{R}_{12}(t_k)$$

$$\tilde{R}_{22} = \tilde{R}_{22}(t_k)$$

As before, we notice that $P(t_{k+1};t_o)$ is the covariance of the estimation error, i.e.,

$$\mathrm{cov}[x(t_k) - \hat{x}(t_k)] = \tilde{P}(t_k;t_o)$$

This fact can be used, for example, to calculate confidence intervals of the estimate.

APPENDIX E

Program TSAMPN:

This program gives a numerical solution of the discrete and continuous linear optimal control problems for stationary systems. The program also compares the values of the loss function for the purpose of finding the influence of the sampling time.

The program has three sets of data cards:

1. The first card gives the formats FMT, FST, and FTT for the print-out of the results and the data cards to follow, respectively. Further, it gives the integers N, NU, and NT specifying the dimensions of the data. The number C 0, giving the accuracy of the iteration for LAMDAMAX, and T 0, the length of the interval over which the optimization is performed.

2. The second group of data cards gives the elements of the matrices F, G, Q, and Q_o as specified by the format FST.

F	N × N	matrix
G	N × NU	matrix
Q1	N × N	matrix
Q12	N × NU	matrix
Q2	NU × NU	matrix
Q0	N × N	matrix

3. The third group of data cards gives the elements of the matrix T(NT × 1), i.e., the lengths of the sampling intervals to be used in the discrete problem.

The program contains the subroutines DYN, EXPR, INVPD, LMAX, NORM, RICCE, and TRANS which are described in detail below. A FORTRAN listing of the programs is given at the end of this section.

Subroutine DYN:

Given the matrices Φ, Γ, $\tilde{Q}_{11}$, $\tilde{Q}_{12}$, $\tilde{Q}_{22}$, and $\tilde{S}(t_{k+1})$ this program computes the matrices $\tilde{L}$ and $\tilde{S}(t_k)$ from

$$\tilde{L} = [\Gamma^T\tilde{S}(t_{k+1})\Gamma + \tilde{Q}_{22}]^{-1} [\Gamma^T\tilde{S}(t_{k+1}) \Phi + \tilde{Q}_{12}^T] \tag{E.1}$$

$$\tilde{S}(t_k) = \Phi^T S(t_{k+1})\Phi - L[\tilde{Q}_{12}^T + \Gamma^T\tilde{S}(t_{k+1})\Phi] + \tilde{Q}_{11} \tag{E.2}$$

The matrix inversion is made by the subroutine INVPD based on the square root method. The ratio R between the determinant and the norm of the matrix to be inverted is calculated at each step. If this value is less than 10^{-6}, a print-out

THE MATRIX S2 IN SUBROUTINE DYN ILLCONDITIONED

DET/NORM = ...

is generated.

The following notations are used in the FORTRAN program.

F	=	Φ	N × N	matrix
G	=	Γ	N × NU	matrix
Q1	=	$\tilde{Q}_{11}$	N × N	matrix
Q12	=	$\tilde{Q}_{12}$	N × NU	matrix
Q2	=	$\tilde{Q}_{22}$	NU × NU	matrix
S1	=	$\tilde{S}(t_{k+1})$	N × N	matrix
S	=	$\tilde{S}(t_k)$	N × N	matrix
AL	=	$\tilde{L}$	NU × N	matrix

Subroutine EXPR:

This routine calculates the exponential function of a matrix by the series

$$\exp A = \sum_{n=0}^{\infty} \frac{1}{n!} A^n \qquad \text{(E.3)}$$

The series is truncated after N terms where N is the smallest number less than 36, such that

$$R \leq 10^{-8} \qquad \text{(E.4)}$$

where

$$R = \left\| \frac{1}{N} A^N \right\| \left\| \sum_{n=0}^{N} \frac{1}{n!} A^n \right\|^{-1} \qquad \text{(E.5)}$$

and $\|A\|$ is the matrix norm (E.7).

If the condition (E.4) is not satisfied for any number less than 36, a print-out

SUBROUTINE EXPR TERMINATED AFTER 35 TERMS R = ...

is generated.

If the (E.7) norm of A is greater than 9, the program computes the exponential function as

$$\exp A = [\exp \frac{1}{n} A]^n$$

where

$$n = \text{largest integer less than } \frac{\|A\|}{4.5}$$

and a print-out of the norm of A is obtained.

Subroutine INVPD:

This subroutine calculates the inverse of a positive definite symmetrical matrix by the square root method.

Subroutine LMAX:

The subroutine calculates the largest eigenvalue of a symmetrical matrix by the Reyleigh Ritz method as described in Faddeeva, p. 212. The iteration is terminated whenever the difference between two consequtive iterates is less than C0. The initial value of the eigenvector is taken as

$$x(0) = \begin{pmatrix} 1.1415963 \\ 1.0000000 \\ \cdot \\ \cdot \\ \cdot \\ 1.0000000 \end{pmatrix}$$

The routine iterates a maximum of 100 times;if the prescribed accuracy is not obtained after 100 iterations, a print-out

SUBROUTINE LMAX TERMINATED AFTER 100 STEPS C = ...

is generated.

Subroutine NORM:

This subroutine calculates the matrix norm

$$\|A\| = \min [\max_i \sum_j |a_{ij}| , \max_j \sum_i |a_{ij}|] \qquad (E.7)$$

Subroutine RICCE:

This program integrates the differential equation

$$\begin{cases} \dot{S} = F^T S + SF - SQ_2S + Q_1 \\ S(0) = Q_o \end{cases} \qquad (E.8)$$

The program utilizes the result that the solution to the Riccati equation can be written as

$$S(t) = [\Sigma_{21}(t) + \Sigma_{22}(t) Q_o] [\Sigma_{11}(t) + \Sigma_{12}(t) Q_o]^{-1} \qquad (E.9)$$

where

$$\begin{pmatrix} \Sigma_{11}(t) & \Sigma_{12}(t) \\ \Sigma_{21}(t) & \Sigma_{22}(t) \end{pmatrix}$$

is the fundamental solution to the linear differential equation

$$\dot{x} = \begin{pmatrix} F & -Q_2 \\ -Q_{11} & -F^T \end{pmatrix} x \qquad (E.10)$$

This linear equation is integrated by the exponential subroutine EXPR and the solution is then formed from (E.9).

For test purposes, we utilized the result that the solution of the Riccati equation also can be expressed as the Taylor series

$$S(t) = \sum_{n=0}^{\infty} A_n t^n \qquad (E.11)$$

where the coefficients A_n are given by

$$A_n = \begin{cases} Q_o & n = 0 \\ F^T Q_o + Q_o F - A_o Q_2 A_o + Q_1 & n = 1 \\ \frac{1}{n}[F^T A_{n-1} + A_{n-1} F - \sum_{k=0}^{n-1} A_k Q_2 A_{n-1-k}] & n > 1 \end{cases} \qquad (E.12)$$

This method was found to be slower than the technique based on the exponential routine.

Subroutine TRANS:

General Description:

This program performs the transformation from the "continuous problem" to the "discrete problem". Given the $n \times n$ matrices ($n \leq 20$) F, G, Q_{11}, Q_{12}, and Q_{22}, which define the "continuous problem", the program computes the $n \times n$ matrices Φ, Γ, Q_{11}, Q_{12}, and Q_{22}, which defines the "discrete problem", from the equations

$$\Phi(t) = \exp Ft \qquad (E.13)$$

$$\Gamma(t) = \left[\int_o^t \Phi(s)\, ds\right] G \qquad (E.14)$$

$$\tilde{Q}_{11}(t) = \int_o^t \Phi^T(s)\, Q_{11}\, \Phi(s)\, ds \qquad (E.15)$$

$$\tilde{Q}_{12}(t) = \int_o^t \Phi^T(s)\, [Q_{11}\, \Gamma(s) + Q_{12}]\, ds \qquad (E.16)$$

$$\tilde{Q}_{22}(t) = \int_o^t [\Gamma^T(s) Q_{11}\, \Gamma(s) + \Gamma^T(s)\, Q_{12} + Q_{21} \Gamma(s) + Q_{22}]\, ds \qquad (E.17)$$

The integrands, being integral functions, can be expanded in power series converging for all t. For all finite t, the series can also be integrated termwise and we can express the integrals as sums of the type

$$I = \sum_{n=0}^{\infty} A_n t^n \qquad (E.18)$$

where the coefficients A_n are given by recursive equations as described in detail below. In the numerical computations, the series are truncated after N terms, where N is the smallest number less than 36 such that

$$\|A_N t^N\| \leq 10^{-8} \|I_N\| \qquad (E.19)$$

where I_N is the N-th partial sum and $\|A\|$ is the matrix norm (E.7).

If there is no N less than 36 which satisfies the inequality (E.19), the computation is terminated and a print-out

COMPUTATION OF ... TERMINATED AFTER 35 TERMS

is generated.

The following notations are used in the FORTRAN listing:

C	=	n		
TS	=	t		
T	=	t^n		
F	=	F	N × N	matrix
G	=	G	N × NU	matrix
Q1	=	Q	N × N	matrix
Q12	=	Q	N × NU	matrix
Q2	=	Q	NU × NU	matrix
FD	=	Φ	N × N	matrix
GD	=	Γ	N × NU	matrix
QID	=	$\tilde{Q}_{11}$	N × N	matrix
Q12D	=	$\tilde{Q}_{12}$	N × NU	matrix
Q2D	=	$\tilde{Q}_{22}$	NU × NU	matrix

Computation of $\tilde{Q}_{22}$:

A trivial calculation shows that the integral (E.17) has the following series expansion:

$$\tilde{Q}_{22}(t) = Q_{22}\, t + \sum_{n=2}^{\infty} [G^T A_n G + G^T B_n + C_n G^T]\, t^n \qquad \text{(E. 20)}$$

where the coefficients A_n, B_n, and C_n are given by the recursive equations

$$A_n = \begin{cases} 0 & n = 2 \\ \frac{1}{n}[F^T A_{n-1} + F A_n] + \frac{1}{n!} Q_{11} F^{n-3} + \frac{1}{n!} (F^T)^{n-3} Q_{11} & n > 2 \end{cases} \qquad \text{(E. 21)}$$

$$B_n = \frac{1}{n!} (F^T)^{n-2} Q_{12} \qquad (E.22)$$

$$C_n = \frac{1}{n!} Q_{12}^T F^{n-2} = B_n^T \qquad (E.23)$$

In the FORTRAN listing, the following notations are used:

$S1 = A_n$

$S2 = A_{n-1}$ and $[G^T A_n G + G^T B_n + C_n G^T] t^n$

$S3 = \frac{1}{n!} F^{n-2}$

$FD = \frac{1}{(n-1)!} F^{n-3}$

Computation of $\tilde{Q}_{12}$:

The integral (E.16) has the series expansion

$$\tilde{Q}_{12}(t) = \sum_{n=1}^{\infty} [A_n G + B_n] t^n \qquad (E.24)$$

where the coefficients A_n and B_n are given by the recursive equations

$$A_n = \begin{cases} 0 & n = 1 \\ \frac{1}{n} [F^T A_{n-1} + B_{n-1}] + \frac{1}{n!} (F^T)^{n-2} Q_{11} & n > 1 \end{cases}$$

(E.25)

$$B_n = \frac{1}{n!} (F^T)^{n-1} Q_{12} \tag{E.26}$$

The following notations are used in the FORTRAN listing:

$$S1 = A_n$$

$$S2 = A_{n-1} \text{ and } (A_n G + B_n)\, t^n$$

$$S3 = \frac{1}{(n-1)!} (F^T)^{n-2}$$

$$FD = \frac{1}{n!} (F^T)^{n-1}$$

Computation of $\tilde{Q}_{11}$:

The integral (E.15) has the series expansion

$$\tilde{Q}_{11}(t) = \sum_{n=1}^{\infty} A_n t^n \tag{E.27}$$

where the coefficients A_n are given by the recursive equations

$$A_n = \begin{cases} Q_n & n = 1 \\ \frac{1}{n}[F^T A_{n-1} + A_{n-1} F] & n > 1 \end{cases} \tag{E.28}$$

The following notations are used in the FORTRAN listing:

$$S1 = A_n$$

$$S2 = A_{n-1} \text{ and } A_n t^n$$

Computation of Φ and Γ:

The functions $\Phi(t)$ and $\Gamma(t)$ have the series expansions

$$\Phi(t) = \sum_{n=0}^{\infty} A_n t^n \tag{E.29}$$

$$\Gamma(t) = \left[\sum_{n=0}^{\infty} B_n t^n \right] G \tag{E.30}$$

where the coefficient A_n is given by the series expansion

$$A_n = \begin{cases} I & n = 0 \\ \frac{1}{n} A_{n-1} F & n > 0 \end{cases} \tag{E.31}$$

and

$$B_n = \frac{1}{n+1} A_n \tag{E.32}$$

The following notations are used in the FORTRAN listing:

$$S1 = A_n$$

$$S2 = A_{n-1}$$

For test purposes, the subroutine considers the particular case when F and Q_{11} are diagonal, i.e.,

$$F = \text{diag.}[\lambda_1, \ldots, \lambda_n] = \text{diag.}[\lambda_i]$$

$$Q_{11} = \text{diag.}[q_1, \ldots, q_n] = \text{diag.}[q_i]$$

Then,

$$\Phi(t) \quad = \quad \text{diag.}\left[e^{\lambda_i t} \right] \qquad \text{(E.33)}$$

$$\Gamma(t) \quad = \quad \text{diag.}\left[\frac{1}{\lambda} \quad e^{\lambda_i t} -1 \right] \; G \qquad \text{(E.34)}$$

$$Q_{11}(t) \quad = \quad \text{diag.}\left[\frac{q_i}{2\lambda_i^{\,2}} \quad e^{2\lambda_i t} - 1 \right] \qquad \text{(E.35)}$$

$$\begin{aligned} Q_{12}(t) \quad = \quad & \text{diag.}\left[\frac{q_i}{2\lambda_i^{\,2}} \quad e^{2\lambda_i t} - 2e^{\lambda_i t} + 1 \right] G \\ & + \text{diag.}\left[\frac{1}{\lambda_i} \quad e^{\lambda_i t} - 1 \right] \; Q_{12} \end{aligned} \qquad \text{(E.36)}$$

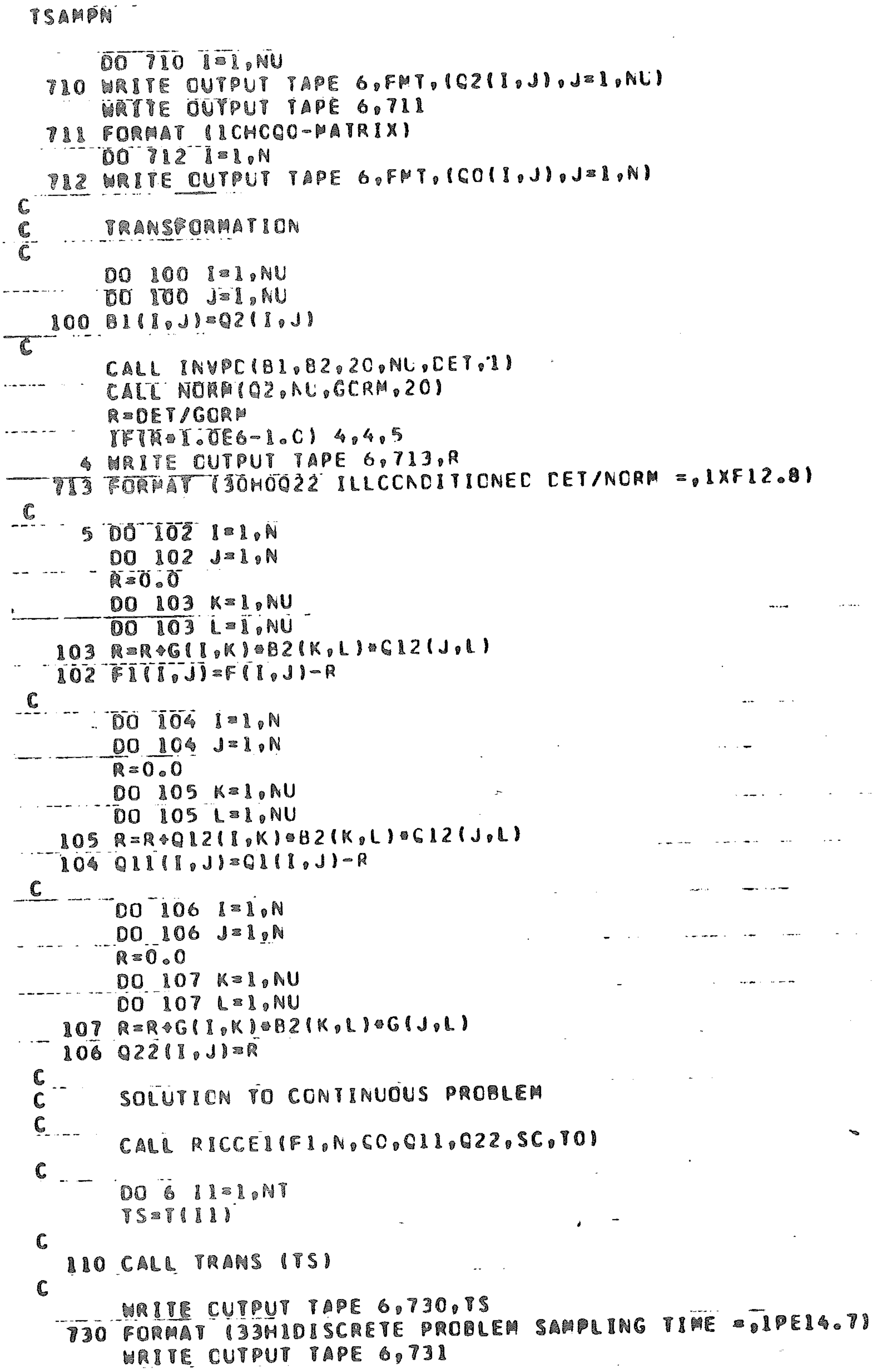
$$\begin{aligned} Q_{22}(t) \quad = \quad & G^T \text{diag.}\left[\frac{q_i}{2\lambda_i^{\,3}} \quad e^{2\lambda_i t} - 4e^{\lambda_i t} + 3 + 2^{\lambda_i t} \right] G \\ & + G^T \text{diag.}\left[\frac{1}{\lambda_i^{\,2}} \quad e^{\lambda_i t} - 1 - \lambda_i t \right] \; Q_{12} \\ & + Q_1{}^T \; \text{diag.}\left[\frac{1}{\lambda_i^{\,2}} \quad e^{\lambda_i t} -1 - \lambda_i t \right] G + Q_{22}\, t \end{aligned}$$

(E.37)

PROGRAM TSAMPN - FORTRAN LISTING

```
 TSAMPN

C     THE PROGRAM SCLVES THE CCNTINUCUS AND THE CISCRETE LINEAR OPTIMAL
C     PROBLEM AND PERFORMS A CCMPARISON FOR THE PURPOSE OF FINCING THE
C     INFLUENCE OF THE SAMPLINC TIME
C
C     SUBROUTINES REGUIREC
C               DYN
C               EXPR
C               INVPD
C               INVS
C               LMAX
C               NCRM
C               RICCE1
C               TRANS
C
      DIMENSION F(20,20),G(20,20),Q1(20,20),C12(20,20),C2(20,20)
      DIMENSION FD(2C,2C),GD(20,20),Q1D(20,20),Q12D(20,20),Q2D(20,20)
      COMMON F,G,Q1,Q12,Q2,FD,GD,Q1C,C12D,Q2D,N,NU
      DIMENSION B1(20,20),B2(2C,20),F1(20,20)
      DIMENSION Q0(20,2C),Q11(20,20),Q22(20,20)
      DIMENSION SD(2C,2C),SC(2C,20),S1(20,20),AL(20,20)
      DIMENSION T(20)
      DIMENSION FMT(3),FST(3),FTT(3)
C
    1 READ INPUT TAPE 5,900,(FMT(I),I=1,3),(FST(I),I=1,3),(FTT(I),I=1,3)
     1,N,NU,NT,CO,TC
  900 FORMAT (9A6,3I2,2E10.5)
      IF(N) 2,2,3
    2 CALL EXIT
C
    3 READ INPUT TAPE 5,FST,((F(I,J),J=1,N),I=1,N),((G(I,J),J=1,NU),I=1,
     1N),((Q1(I,J),J=1,N),I=1,N),((Q2(I,J),J=1,NU),I=1,NU),((Q12(I,J),J=
     21,NU),I=1,N),((Q0(I,J),J=1,N),I=1,N)
C
      READ INPUT TAPE 5,FTT,(T(I),I=1,NT)
C
      WRITE OUTPUT TAPE 6,7C0,TC
  700 FORMAT (23H1CCNTINUOUS PRCBLEM T =,1PE14.7)
      WRITE OUTPUT TAPE 6,7C1
  701 FORMAT ( 9HOF-MATRIX)
      DO 702  I=1,N
  702 WRITE OUTPUT TAPE 6,FMT,(F(I,J),J=1,N)
      WRITE OUTPUT TAPE 6,7C3
  703 FORMAT ( 9HOG-MATRIX)
      DO 704  I=1,N
  704 WRITE OUTPUT TAPE 6,FMT,(G(I,J),J=1,NU)
      WRITE OUTPUT TAPE 6,7C5
  705 FORMAT (11HOQ11-MATRIX)
      DO 7C6 I=1,N
  706 WRITE OUTPUT TAPE 6,FMT,(Q1(I,J),J=1,N)
      WRITE OUTPUT TAPE 6,7C7
  707 FORMAT (11HOQ12-MATRIX)
      DO 708 I=1,N
  708 WRITE OUTPUT TAPE 6,FMT,(Q12(I,J),J=1,NU)
      WRITE OUTPUT TAPE 6,7C9
  709 FORMAT (11HOQ22-MATRIX)
```

PROGRAM TSAMPN - Continued

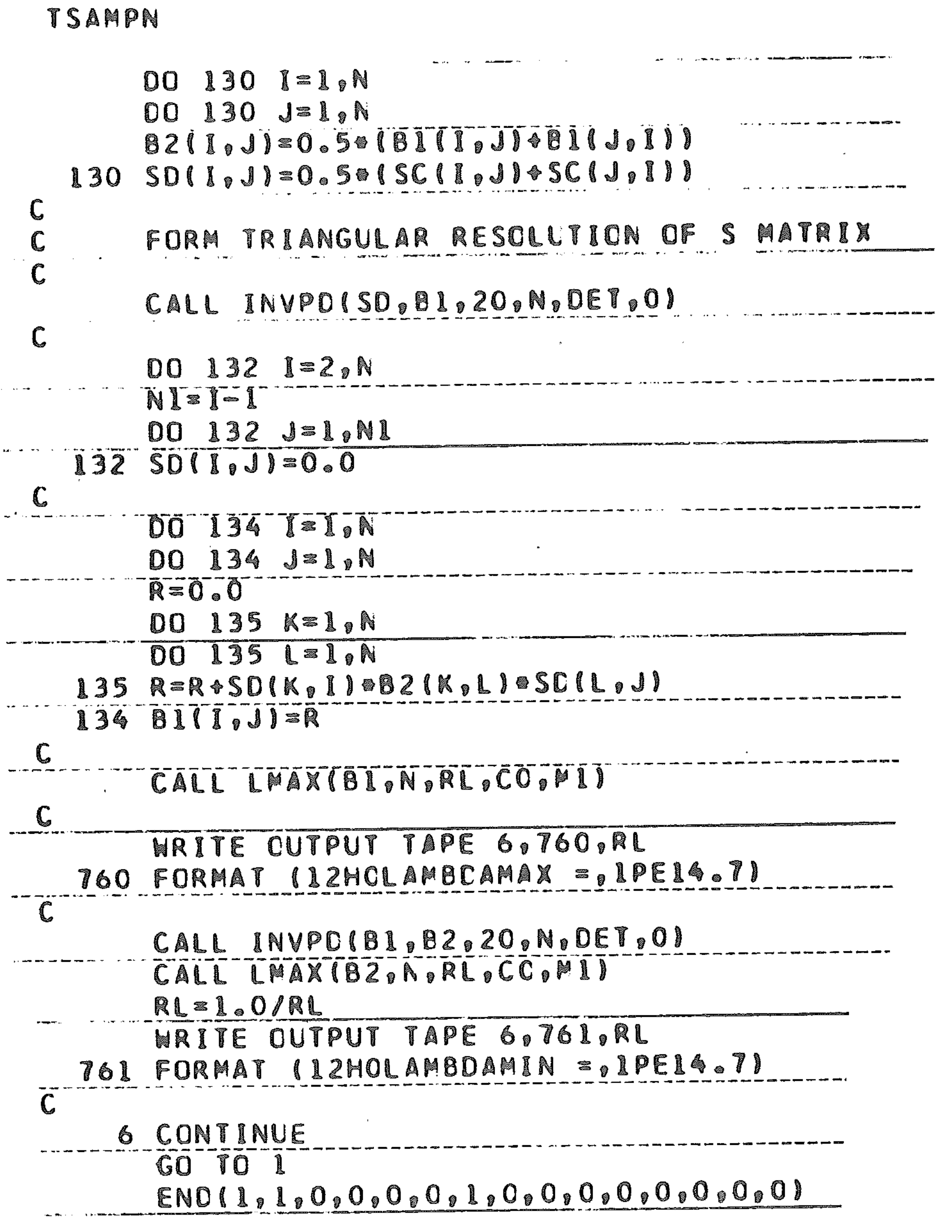


```
TSAMPN

      DO 710 I=1,NU
  710 WRITE OUTPUT TAPE 6,FMT,(Q2(I,J),J=1,NL)
      WRITE OUTPUT TAPE 6,711
  711 FORMAT (1CHCQO-MATRIX)
      DO 712 I=1,N
  712 WRITE OUTPUT TAPE 6,FMT,(CO(I,J),J=1,N)
C
C     TRANSFORMATION
C
      DO 100 I=1,NU
      DO 100 J=1,NU
  100 B1(I,J)=Q2(I,J)
C
      CALL INVPC(B1,B2,20,NL,CET,1)
      CALL NORM(Q2,NL,GCRM,20)
      R=DET/GORM
      IF(R*1.0E6-1.C) 4,4,5
    4 WRITE CUTPUT TAPE 6,713,R
  713 FORMAT (30HOQ22 ILLCCNDITIONEC CET/NORM =,1XF12.8)
C
    5 DO 102 I=1,N
      DO 102 J=1,N
      R=0.0
      DO 103 K=1,NU
      DO 103 L=1,NU
  103 R=R+G(I,K)*B2(K,L)*G12(J,L)
  102 F1(I,J)=F(I,J)-R
C
      DO 104 I=1,N
      DO 104 J=1,N
      R=0.0
      DO 105 K=1,NU
      DO 105 L=1,NU
  105 R=R+Q12(I,K)*B2(K,L)*G12(J,L)
  104 Q11(I,J)=Q1(I,J)-R
C
      DO 106 I=1,N
      DO 106 J=1,N
      R=0.0
      DO 107 K=1,NU
      DO 107 L=1,NU
  107 R=R+G(I,K)*B2(K,L)*G(J,L)
  106 Q22(I,J)=R
C
C     SOLUTION TO CONTINUOUS PROBLEM
C
      CALL RICCE1(F1,N,CO,Q11,Q22,SC,TO)
C
      DO 6 II=1,NT
      TS=T(II)
C
  110 CALL TRANS (TS)
C
      WRITE CUTPUT TAPE 6,730,TS
  730 FORMAT (33H1DISCRETE PROBLEM SAMPLING TIME =,1PE14.7)
      WRITE CUTPUT TAPE 6,731
```

PROGRAM TSAMPN - Continued

```
 TSAMPN

  731 FORMAT (1CHCFC-MATRIX)
      DO 732 I=1,N
  732 WRITE OUTPUT TAPE 6,FMT,(FD(I,J),J=1,N)
      WRITE OUTPUT TAPE 6,733
  733 FORMAT (10HCGC-MATRIX)
      DO 734 I=1,N
  734 WRITE OUTPUT TAPE 6,FMT,(GD(I,J),J=1,NU)
      WRITE OUTPUT TAPE 6,735
  735 FORMAT (12HCQ11C-MATRIX)
      DO 736 I=1,N
  736 WRITE CUTPUT TAPE 6,FMT,(C1D(I,J),J=1,N)
      WRITE CUTPUT TAPE 6,737
  737 FORMAT (12HCQ12C-MATRIX)
      DO 738 I=1,N
  738 WRITE OUTPUT TAPE 6,FMT,(C12D(I,J),J=1,NU)
      WRITE CUTPUT TAPE 6,739
  739 FORMAT (12HOQ22C-MATRIX)
      DO 740  I=1,NU
  740 WRITE OUTPUT TAPE 6,FMT,(C2D(I,J),J=1,NU)
C
C     SCLUTICN TO DISCRETE PROBLEM
C
      DO 112 I=1,N
      DO 112 J=1,N
  112 SD(I,J)=QO(I,J)
      T1=0.0
C
  113 DO 114 I=1,N
      DC 114 J=1,N
  114 S1(I,J)=0.5*(SC(I,J)+SD(J,I))
C
      CALL DYN (FD,GC,N,NU,C1D,C12D,Q2D,S1,SD,AL)
C
      T1=T1+TS
      IF(T1-T0+0.5*TS) 113,115,115
C
  115 DC 116 I=1,N
      DO 116 J=1,N
  116 B1(I,J)=SC(I,J)-SC(I,J)
C

      WRITE CUTPUT TAPE 6,75C
  750 FORMAT (9HOS-MATRIX)
      DC 751 I=1,N
  751 WRITE OUTPUT TAPE 6,FMT,(SC(I,J),J=1,N)
      WRITE CUTPUT TAPE 6,752
  752 FORMAT (14HOSTILDE-MATRIX)
      DU 753 I=1,N
  753 WRITE OUTPUT TAPE 6,FMT,(SD(I,J),J=1,N)
      WRITE OUTPUT TAPE 6,754
  754 FORMAT (9HOT-MATRIX)
      DO 755 I=1,N
  755 WRITE OUTPUT TAPE 6,FMT,(B1(I,J),J=1,N)
C
C     SYMMETRIZATION OF MATRICES
C
```

PROGRAM TSAMPN - Continued

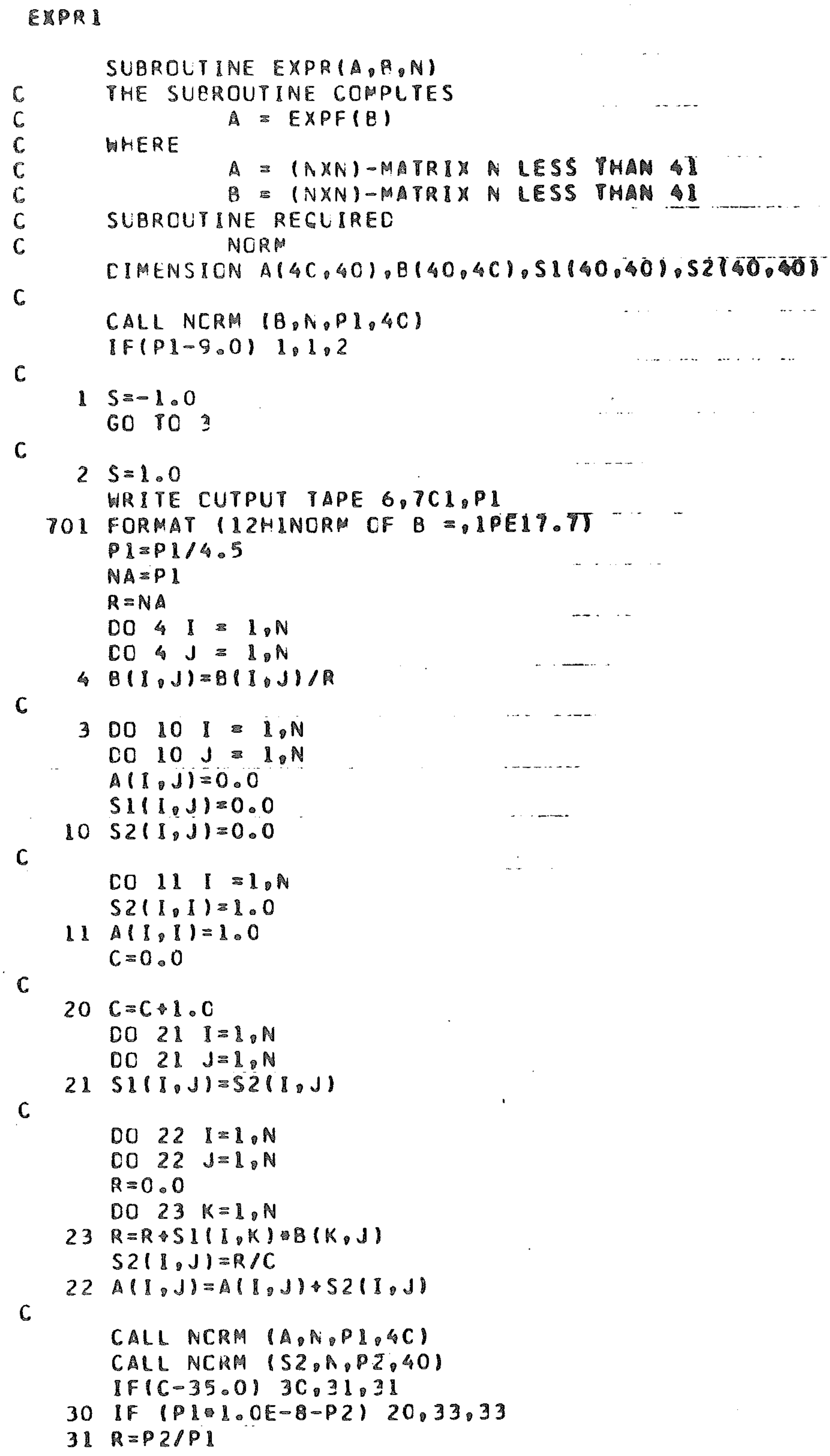


```
TSAMPN
      DO 130 I=1,N
      DO 130 J=1,N
      B2(I,J)=0.5*(B1(I,J)+B1(J,I))
  130 SD(I,J)=0.5*(SC(I,J)+SC(J,I))
C
C     FORM TRIANGULAR RESOLLTION OF S MATRIX
C
      CALL INVPD(SD,B1,20,N,DET,0)
C
      DO 132 I=2,N
      N1=I-1
      DO 132 J=1,N1
  132 SD(I,J)=0.0
C
      DO 134 I=1,N
      DO 134 J=1,N
      R=0.0
      DO 135 K=1,N
      DO 135 L=1,N
  135 R=R+SD(K,I)*B2(K,L)*SC(L,J)
  134 B1(I,J)=R
C
      CALL LMAX(B1,N,RL,CO,M1)
C
      WRITE OUTPUT TAPE 6,760,RL
  760 FORMAT (12HOLAMBCAMAX =,1PE14.7)
C
      CALL INVPD(B1,B2,20,N,DET,0)
      CALL LMAX(B2,N,RL,CC,M1)
      RL=1.0/RL
      WRITE OUTPUT TAPE 6,761,RL
  761 FORMAT (12HOLAMBDAMIN =,1PE14.7)
C
    6 CONTINUE
      GO TO 1
      END(1,1,0,0,0,0,1,0,0,0,0,0,0,0,0)
```

PROGRAM DYN - FORTRAN LISTING

```
 DYN
       SUBROUTINE CYN(F,G,N,NU,C1,Q12,Q2,S1,S,AL)
C      THE PRCGRAM  COMPLTES S AND AL FROM
C              AL=((GT)S1G+Q2)-1)((GT)S1F+Q12T)
C              S=(FT)S1F-ALT(C12T+(GT)S1F)+Q1
C      SUBROUTINE RECUIRED
C              INVPD
C
       DIMENSION F(20,20),G(20,20),Q1(20,20),Q12(20,20),Q2(20,20),
      1S1(20,20),S(20,20),AL(2C,2C)
       DIMENSION S2(2C,20),S3(2C,20)
C
C      COMPUTATION OF L
C
       DO 10 I=1,NU
       DO 10 J=1,NU
       R=0.0
       CO 11 K=1,N
       DO 11 L=1,N
    11 R=R+G(K,I)*S1(K,L)*G(L,J)
    10 S2(I,J)=R+Q2(I,J)
C
       CALL NORM(S2,NU,GORM,20)
       CALL INVPD(S2,S3,20,NU,DET,1)
C
       R=DET/GORM
       IF(R*1.CE6-1.0) 1,2,2
     1 WRITE CUTPUT TAPE 6,7C0,R
   700 FORMAT (51HOTHE MATRIX S2 IN SUBRCUTINE DYN ILLCONDITIONED R =,1XF
      110.8)
C
     2 DO 12 I=1,NU
       DO 12 J=1,N
       R=0.0
       DO 13 K=1,N
       DO 13 L=1,N
    13 R=R+G(K,I)*S1(K,L)*F(L,J)
    12 S2(I,J)=R+Q12(J,I)
C
       DO 14 I=1,NU
       DO 14 J=1,N
       R=0.0
       DO 15 K=1,NU
    15 R=R+S3(I,K)*S2(K,J)
    14 AL(I,J)=R
C
C      COMPUTATICN OF S
C
       CO 16 I=1,N
       CO 16 J=1,N
       R=0.0
       DO 17 K=1,N
       DO 17 L=1,N
    17 R=R+F(K,I)*S1(K,L)*F(L,J)
       DO 18 K=1,NU
    18 R=R-AL(K,I)*S2(K,J)

 DYN
    16 S(I,J)=R+Q1(I,J)
C
       RETURN
       END(1,1,0,0,0,C,1,0,0,C,C,0,0,0,0)
```

PROGRAM EXPR1 - FORTRAN LISTING

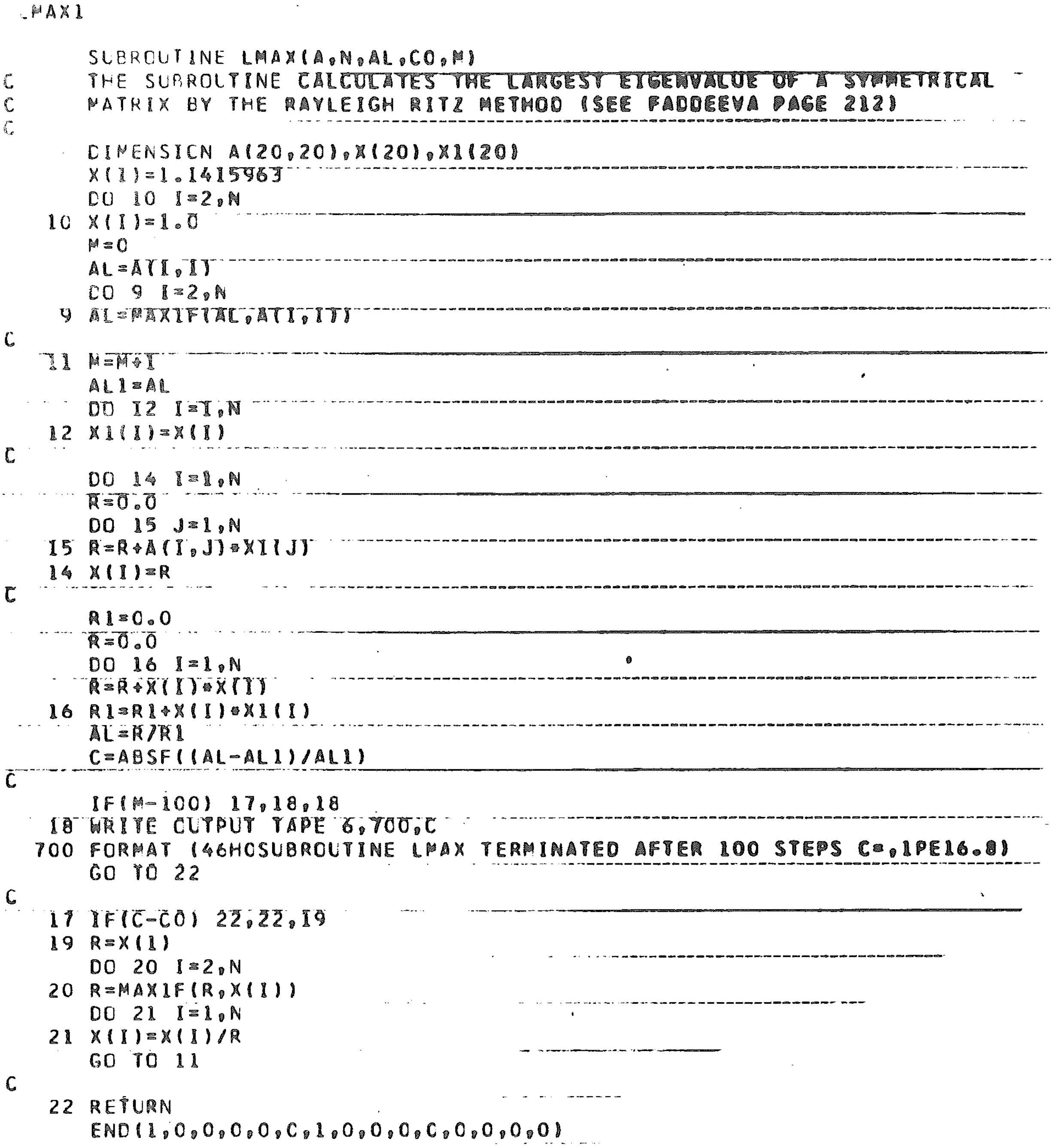


```
 EXPR1

       SUBROUTINE EXPR(A,B,N)
C      THE SUBROUTINE COMPLTES
C              A = EXPF(B)
C      WHERE
C              A = (NXN)-MATRIX N LESS THAN 41
C              B = (NXN)-MATRIX N LESS THAN 41
C      SUBROUTINE REGUIRED
C              NORM
       CIMENSION A(4C,40),B(40,4C),S1(40,40),S2(40,40)
C
       CALL NCRM (B,N,P1,4C)
       IF(P1-9.0) 1,1,2
C
     1 S=-1.0
       GO TO 3
C
     2 S=1.0
       WRITE CUTPUT TAPE 6,7C1,P1
   701 FORMAT (12H1NORM CF B =,1PE17.7)
       P1=P1/4.5
       NA=P1
       R=NA
       DO 4 I = 1,N
       CO 4 J = 1,N
     4 B(I,J)=B(I,J)/R
C
     3 DO 10 I = 1,N
       CO 10 J = 1,N
       A(I,J)=0.0
       S1(I,J)=0.0
    10 S2(I,J)=0.0
C
       CO 11 I =1,N
       S2(I,I)=1.0
    11 A(I,I)=1.0
       C=0.0
C
    20 C=C+1.C
       DO 21 I=1,N
       DO 21 J=1,N
    21 S1(I,J)=S2(I,J)
C
       DO 22 I=1,N
       DO 22 J=1,N
       R=0.0
       DO 23 K=1,N
    23 R=R+S1(I,K)*B(K,J)
       S2(I,J)=R/C
    22 A(I,J)=A(I,J)+S2(I,J)
C
       CALL NCRM (A,N,P1,4C)
       CALL NCRM (S2,N,P2,40)
       IF(C-35.0) 3C,31,31
    30 IF (P1*1.0E-8-P2) 20,33,33
    31 R=P2/P1
```

PROGRAM EXPR1 - Continued

```
 EXPR1

       WRITE CUTPUT TAPE 6,7CC,R
   700 FORMAT (45HOSLBRCLTINE EXPR TERMINATED AFTER 35 TERMS R=,1XF1C.8)
C
    33 IF(S) 35,35,4C
C
    4C CO 41 I=1,N
       CC 41 J=1,N
    41 B(I,J)=A(I,J)
C
       NA=NA-1
       CO 44 K=1,NA
       DO 43 I=1,N
       DO 43 J=1,N
    43 S1(I,J)=A(I,J)
C
       DO 44 I=1,N
       DO 44 J=1,N
       R=0.0
       CO 45 L=1,N
    45 R=R+B(I,L)*S1(L,J)
    44 A(I,J)=R
C
    35 RETURN
       END(1,1,0,0,0,0,1,0,0,C,C,0,0,0,0)
```

PROGRAM LMAX1 - FORTRAN LISTING

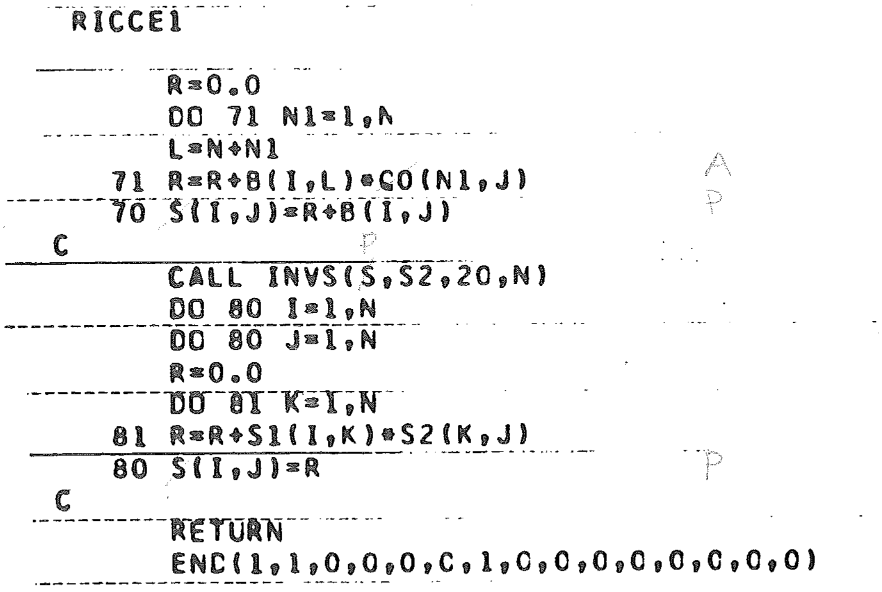


```
.MAX1

      SUBROUTINE LMAX(A,N,AL,CO,M)
C     THE SUBROUTINE CALCULATES THE LARGEST EIGENVALUE OF A SYMMETRICAL
C     MATRIX BY THE RAYLEIGH RITZ METHOD (SEE FADDEEVA PAGE 212)
C
      DIMENSION A(20,20),X(20),X1(20)
      X(1)=1.1415963
      DO 10 I=2,N
   10 X(I)=1.0
      M=0
      AL=A(1,1)
      DO 9 I=2,N
    9 AL=MAX1F(AL,A(I,I))
C
   11 M=M+1
      AL1=AL
      DO 12 I=1,N
   12 X1(I)=X(I)
C
      DO 14 I=1,N
      R=0.0
      DO 15 J=1,N
   15 R=R+A(I,J)*X1(J)
   14 X(I)=R
C
      R1=0.0
      R=0.0
      DO 16 I=1,N
      R=R+X(I)*X(I)
   16 R1=R1+X(I)*X1(I)
      AL=R/R1
      C=ABSF((AL-AL1)/AL1)
C
      IF(M-100) 17,18,18
   18 WRITE OUTPUT TAPE 6,700,C
  700 FORMAT (46H0SUBROUTINE LMAX TERMINATED AFTER 100 STEPS C=,1PE16.8)
      GO TO 22
C
   17 IF(C-CO) 22,22,19
   19 R=X(1)
      DO 20 I=2,N
   20 R=MAX1F(R,X(I))
      DO 21 I=1,N
   21 X(I)=X(I)/R
      GO TO 11
C
   22 RETURN
      END(1,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0)
```

PROGRAM RICCE - FORTRAN LISTING

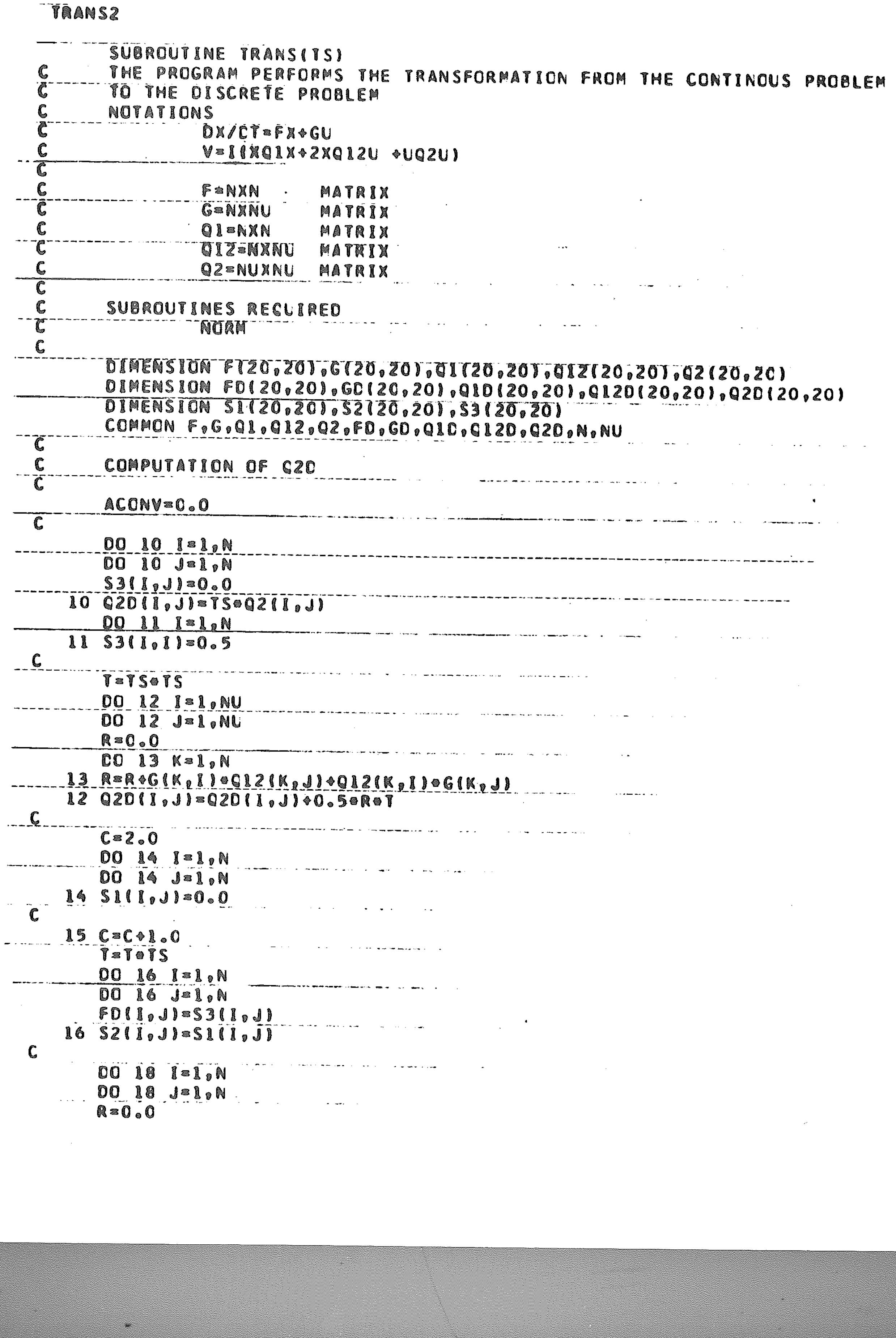
```
 RICCE1

       SUBROUTINE RICCE1(F,N,QC,C1,Q2,S,T)
C      THE PRCGRAM COMPLTES THE SOLUTION TO THE RICCATIEQUATION
C              DS/CT=(FT)S+SF-SC2S+Q1
C      BY USING THE EXPONENTIAL SERIES FOR THE CANONICAL EQUATICN
C
C      SUBROUTINES RECUIREC
C              EXPR
C              INVPD
C              INVS
       DIMENSION F(20,20),Q0(20,20),Q1(20,20),Q2(20,20),S(20,20)
       DIMENSION S1(20,20),S2(2C,20),A(40,40),B(40,40)
C
C      COMPUTATION OF A-MATRIX
C
       CO 90 I=1,N
       DO 90 J=1,N
    90 A(I,J)=F(I,J)
C
       DO 20 I=1,N
       DO 20 J=1,N
       K=N+J
    20 A(I,K)=-Q2(I,J)
C
       DO 30 I=1,N
       DO 30 J=1,N
       K=I+N
    30 A(K,J)=-Q1(I,J)
C
       DO 40 I=1,N
       DO 40 J=1,N
       K=N+I
       L=N+J
    40 A(K,L)=-F(J,I)
C
C
C      COMPUTATICN OF EXP(A*T)
C
       M=N+N
       DO 50 I=1,M
       DO 50 J=1,M
    50 A(I,J)=-T*A(I,J)
C
       CALL EXPR(B,A,M)
C
       DO 60 I=1,N
       DO 60 J=1,N
       R=0.0
       K=N+I
       DO 61 N1=1,N
       L=N+N1
    61 R=R+B(K,L)*Q0(N1,J)
    60 S1(I,J)=B(K,J)+R
C
       DO 70 I=1,N
       DC 70 J=1,N
```

PROGRAM RICCE - Continued



```
 RICCE1

        R=0.0
        DO 71 N1=1,N
        L=N+N1
     71 R=R+B(I,L)*CO(N1,J)
     70 S(I,J)=R+B(I,J)
 C
        CALL INVS(S,S2,20,N)
        DO 80 I=1,N
        DO 80 J=1,N
        R=0.0
        DO 81 K=1,N
     81 R=R+S1(I,K)*S2(K,J)
     80 S(I,J)=R
 C
        RETURN
        ENC(1,1,0,0,0,C,1,C,0,0,0,0,C,0,0)
```

PROGRAM TRANS2 - FORTRAN LISTING

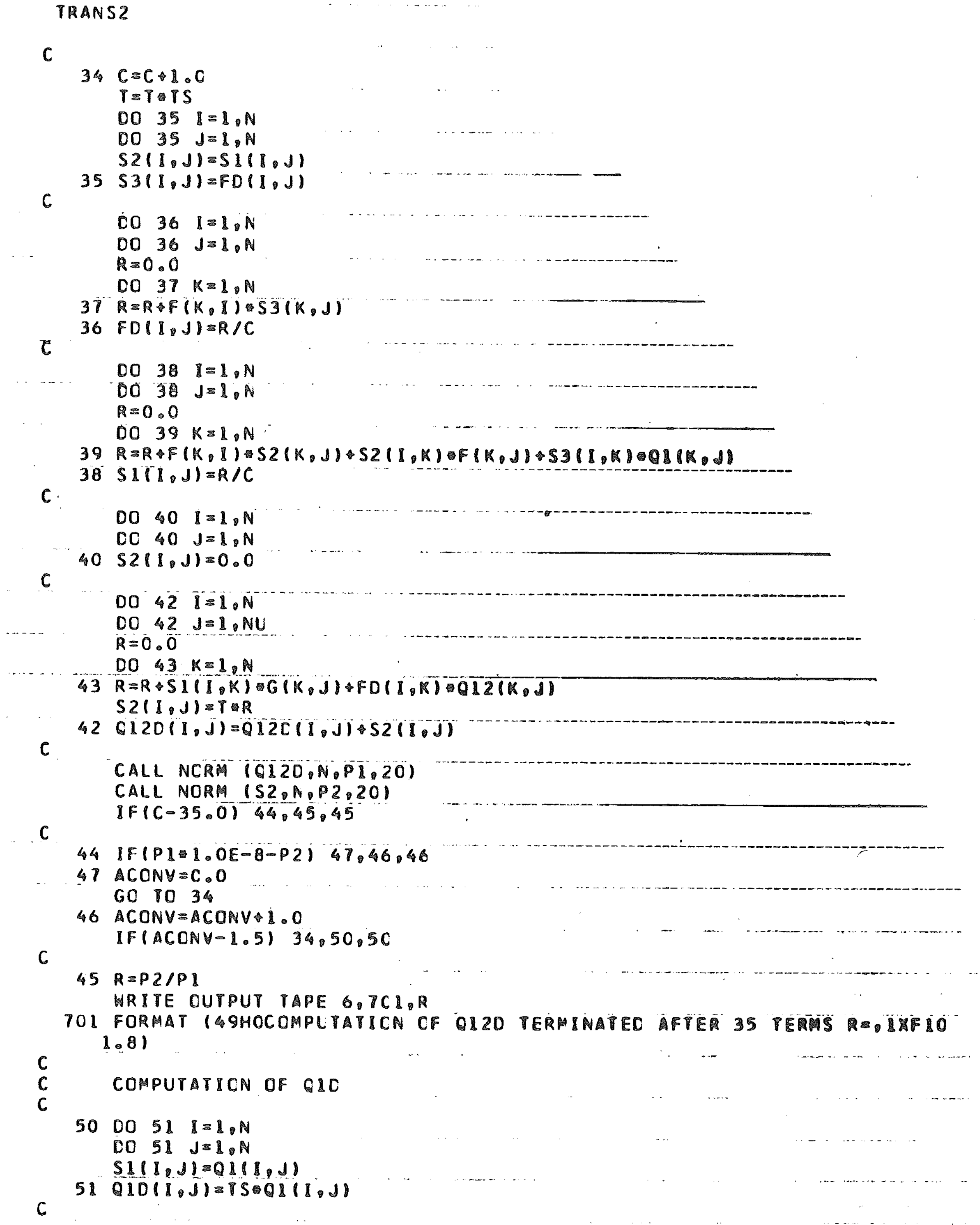


```
 TRANS2

      SUBROUTINE TRANS(TS)
C     THE PROGRAM PERFORMS THE TRANSFORMATION FROM THE CONTINOUS PROBLEM
C     TO THE DISCRETE PROBLEM
C     NOTATIONS
C               DX/CT=FX+GU
C               V=I(XQ1X+2XQ12U +UQ2U)
C
C               F=NXN        MATRIX
C               G=NXNU       MATRIX
C               Q1=NXN       MATRIX
C               Q12=NXNU     MATRIX
C               Q2=NUXNU     MATRIX
C
C     SUBROUTINES RECUIRED
C               NORM
C
      DIMENSION F(20,20),G(20,20),Q1(20,20),Q12(20,20),Q2(20,20)
      DIMENSION FD(20,20),GD(20,20),Q1D(20,20),Q12D(20,20),Q2D(20,20)
      DIMENSION S1(20,20),S2(20,20),S3(20,20)
      COMMON F,G,Q1,Q12,Q2,FD,GD,Q1C,Q12D,Q2D,N,NU
C
C     COMPUTATION OF Q2C
C
      ACONV=0.0
C
      DO 10 I=1,N
      DO 10 J=1,N
      S3(I,J)=0.0
   10 Q2D(I,J)=TS*Q2(I,J)
      DO 11 I=1,N
   11 S3(I,I)=0.5
C
      T=TS*TS
      DO 12 I=1,NU
      DO 12 J=1,NU
      R=0.0
      DO 13 K=1,N
   13 R=R+G(K,I)*Q12(K,J)+Q12(K,I)*G(K,J)
   12 Q2D(I,J)=Q2D(I,J)+0.5*R*T
C
      C=2.0
      DO 14 I=1,N
      DO 14 J=1,N
   14 S1(I,J)=0.0
C
   15 C=C+1.0
      T=T*TS
      DO 16 I=1,N
      DO 16 J=1,N
      FD(I,J)=S3(I,J)
   16 S2(I,J)=S1(I,J)
C
      DO 18 I=1,N
      DO 18 J=1,N
      R=0.0
```

PROGRAM TRANS2 - Continued

```
TRANS2

      DO 19 K=1,N
   19 R=R+F(I,K)*FD(K,J)
   18 S3(I,J)=R/C
C
      DO 20 I=1,N
      DO 20 J=1,N
      R=0.0
      CO 21 K=1,N
   21 R=R+F(K,I)*S2(K,J)+S2(I,K)*F(K,J)+Q1(I,K)*FD(K,J)+FC(K,I)*Q1(K,J)
   20 S1(I,J)=R/C
C
      CO 22 I=1,NU
      DO 22 J=1,NU
      R=0.0
      DO 23 K=1,N
      CO 23 L=1,N
   23 R=R+G(K,I)*S1(K,L)*G(L,J)+G(K,I)*S3(L,K)*Q12(L,J)+Q12(K,I)*S3(K,L)
     1*G(L,J)
      S2(I,J)=T*R
   22 Q2D(I,J)=Q2D(I,J)+S2(I,J)
C
C
      CALL NCRM (Q2D,NU,P1,20)
      CALL NORM(S2,NU,P2,20)
      IF (C-35.0) 24,25,25
C
   24 IF (P1*1.0E-8-P2) 27,26,26
   27 ACONV=C.0
      GO TO 15
   26 ACONV=ACONV+1.C
      IF(ACONV-1.5) 15,30,3C
C
   25 R=P2/P1
      WRITE CUTPUT TAPE 6,7CC,R
  700 FORMAT (48HCCCMPUTATICN CF Q2C TERMINATED AFTER 35 TERMS R=,1XF10.
     18)
C
C     COMPUTATICN OF Q12D
C
      ACONV=0.0
C
   30 DO 31 I=1,N
      CO 31 J=1,N
      Q12C(I,J)=C.0
      S1(I,J)=0.0
   31 FD(I,J)=0.0
C
      CO 32 I=1,N
   32 FD(I,I)=1.0
C
      DO 33 I=1,N
      CO 33 J=1,NU
   33 Q12D(I,J)=TS*Q12(I,J)
C
      C=1.0
      T=TS
```

PROGRAM TRANS2 - Continued



```
 TRANS2

C
   34 C=C+1.0
      T=T*TS
      DO 35 I=1,N
      DO 35 J=1,N
      S2(I,J)=S1(I,J)
   35 S3(I,J)=FD(I,J)
C
      DO 36 I=1,N
      DO 36 J=1,N
      R=0.0
      DO 37 K=1,N
   37 R=R+F(K,I)*S3(K,J)
   36 FD(I,J)=R/C
C
      DO 38 I=1,N
      DO 38 J=1,N
      R=0.0
      DO 39 K=1,N
   39 R=R+F(K,I)*S2(K,J)+S2(I,K)*F(K,J)+S3(I,K)*Q1(K,J)
   38 S1(I,J)=R/C
C
      DO 40 I=1,N
      CC 40 J=1,N
   40 S2(I,J)=0.0
C
      DO 42 I=1,N
      DO 42 J=1,NU
      R=0.0
      DO 43 K=1,N
   43 R=R+S1(I,K)*G(K,J)+FD(I,K)*Q12(K,J)
      S2(I,J)=T*R
   42 Q12D(I,J)=Q12C(I,J)+S2(I,J)
C
      CALL NCRM (Q12D,N,P1,20)
      CALL NORM (S2,N,P2,20)
      IF(C-35.0) 44,45,45
C
   44 IF(P1*1.0E-8-P2) 47,46,46
   47 ACONV=C.0
      GO TO 34
   46 ACONV=ACONV+1.0
      IF(ACONV-1.5) 34,50,5C
C
   45 R=P2/P1
      WRITE CUTPUT TAPE 6,7C1,R
  701 FORMAT (49HOCOMPLTATICN CF Q12D TERMINATED AFTER 35 TERMS R=,1XF10
     1.8)
C
C     COMPUTATICN OF Q1C
C
   50 DO 51 I=1,N
      CO 51 J=1,N
      S1(I,J)=Q1(I,J)
   51 Q1D(I,J)=TS*Q1(I,J)
C
```

PROGRAM TRANS2 - Continued

```
TRANS2
      T=TS
      C=1.0
C
   53 C=C+1.C
      T=T*TS
      DO 52 I=1,N
      DO 52 J=1,N
   52 S2(I,J)=S1(I,J)
C
      DO 54 I=1,N
      DO 54 J=1,N
      R=0.0
      DO 55 K=1,N
   55 R=R+F(K,I)*S2(K,J)+S2(I,K)*F(K,J)
      S1(I,J)=R/C
   54 Q1D(I,J)=Q1D(I,J)+T*S1(I,J)
C
      CALL NORM(Q1D,N,P1,20)
      CALL NORM(S1 ,N,P2,20)
      IF(C-35.0) 56,57,57
   56 IF(P1*1.0E-8-P2*T) 53,60,6C
   57 R=P2*T/P1
      WRITE OUTPUT TAPE 6,7C2,R
  702 FORMAT (48HOCCMPUTATICN CF Q1D TERMINATED AFTER 35 TERMS R=,1XF10.
     18)
C
C     COMPUTATICN OF FC AND GC
C
   60 DO 61 I=1,N
      DO 61 J=1,N
      FC(I,J)=0.0
      GD(I,J)=0.0
      S1(I,J)=0.0
   61 S3(I,J)=0.0
C
      DO 62 I=1,N
      FD(I,I)=1.0
      S1(I,I)=1.0
   62 S3(I,I)=TS
C
      C=C.0
      T=1.0
C
   63 C=C+1.0
      T=TS*T
C
      CO 64 I=1,N
      DO 64 J=1,N
   64 S2(I,J)=S1(I,J)
C
      CO 66 I=1,N
      DO 66 J=1,N
      R=C.0
      DO 67 K=1,N
   67 R=R+S2(I,K)*F(K,J)
      S1(I,J)=R/C
```

PROGRAM TRANS2 - Continued

```
 TRANS2

       FC(I,J)=FC(I,J)+T*S1(I,J)
    66 S3(I,J)=S3(I,J)+T*S1(I,J)*TS/(C+1.0)
C
       CALL NCRM (FC,N,P1,20)
       CALL NCRM (S1,N,P2,20)
       IF(C-35.0) 68,69,69
    68 IF(P1*1.0E-8-P2*T) 63,70,7C
    69 R=P2*T/P1
       WRITE CUTPUT TAPE 6,7C3,R
   7U3 FORMAT (54HOCCMPLTATICN CF FD AND GC TERMINATED AFTER 35 TERMS R=,
      11XF10.8)
    70 DC 72 I=1,N
       DO 72 J=1,NU
       R=0.0
       DO 73 K=1,N
    73 R=R+S3(I,K)*G(K,J)
    72 GD(I,J)=R
C
       RETURN
       END(1,1,0,C,C,C,1,0,0,0,C,C,C,C,0)
```

APPENDIX F

Program SAMPAS:

This program integrates the differential equations

$$\frac{d}{dt} \sum (t;t_1) = A \sum (t;t_1) \qquad , \sum (t_1;t_1) = I \qquad (F.1)$$

$$\frac{d}{dt} E\,(t;t_1) = A \;\; E(t;t_1) + B \sum (t;t_1) \quad , E\,(t_1;t_1) = 0 \qquad (F.2)$$

where

$$A = \begin{pmatrix} F^* & -G\,Q_{22}^{-1}\,G^T \\ -Q_{11}{}^* & -F^{*T} \end{pmatrix} \qquad (F.3)$$

$$B = \begin{pmatrix} F^*\,G\,Q_{22}^{-1}\,G^T\,Q_{11}{}^* & F^*\,G\,Q_{22}^{-1}\,G^T\,F^{*T} \\ -Q_{11}{}^*\,G\,Q_{22}^{-1}\,G^T\,Q_{11}{}^* & -Q_{11}{}^*\,G\,Q_{22}^{-1}\,G^T\,F^{*T} \end{pmatrix} \qquad (F.4)$$

$$F^* = F - G\,Q_{22}^{-1}\,Q_{21} \qquad (F.5)$$

$$Q_{11}{}^* = Q_{11} - Q_{12}\,Q_{22}^{-1}\,Q_{21} \qquad (F.6)$$

and forms the matrices

$$S = [\Sigma_2 + \Sigma_{22}\,Q_o]\;[\Sigma_{11} + \Sigma_{12}\,Q_o]^{-1} \qquad (F.7)$$

$$C = \{E_2 + E_{22}\,Q_o - S[E_{11} + E_1\;Q_o]\}\;[\Sigma_{11} + \Sigma_{12}\,Q_o]^{-1} \qquad (F.8)$$

which are used in the asymptotic formulas of Section II. The program also computes the largest eigenvalue of the matrix $S^{-1}C$.

The integration of the differential equations (F.1) and (F.2) is done by using the exponential series.

The program has two sets of data cards:

1. The first card gives the formats FMT and FST for the print-out of the results and the data cards to follow. Further, the card gives N, NV, specifying the dimensions of F and G. The number C 0 gives the accuracy desired in the computation of LAMBDAMAX, and T 0 is the length of the time interval over which the optimization is performed.
2. The second set of data cards gives the elements of the matrices F, G, Q_{11}, Q_{12}, Q_{22}, and Q_o as specified by the format FST.

The program uses the subroutines EXP, FORMA, INVPD, LMAX, and NORM. The subroutine FORMA is described below, all others were described in Appendix E.

SUBROUTINE FORMA:

Given F, G, Q_{11}, Q_{12}, and Q_{22}, this subroutine forms the matrix

$$\begin{pmatrix} A & 0 \\ B & A \end{pmatrix}$$

as defined by equations (F. 1) through (F. 6).

PROGRAM SAMPAS - FORTRAN LISTING

```
 SAMPAS
C      THE PRCGRAM EVALUATES THE ASYMPTOTIC FORMULAS  WHICH GIVES THE
C      INFLUENCE OF THE SAMPLING RATE  FCR A LINEAR STATIONARY PLANT
C
C      SUBRCUTINES RECLIRED
C              EXP
C              FCRMA
C              INVPD
C              INVS
C              LMAX
C              NCRM
C
       DIMENSICN F(16,16),G(16,16),Q1(16,16),Q12(16,16),Q2(16,16)
       DIMENSICN Q0(16,16),S(16,16),S1(16,16),S2(16,16),C(16,16)
       DIMENSION A(64,64),B(64,64)
       DIMENSICN FMT(3),FST(3)
       COMMON A,F,G,Q1,Q12,Q2,N,NU
C
     1 READ INPUT TAPE 5,900,(FMT(I),I=1,3),(FST(I),I=1,3),N,NU,TO,CO
   900 FORMAT (6A6,2I2,2E10.5)
       IF(N) 2,2,3
     2 CALL EXIT
C
     3 READ INPUT TAPE 5,FST,((F(I,J),J=1,N),I=1,N),((G(I,J),J=1,NU),I=1,
      1N),((Q1(I,J),J=1,N),I=1,N),((Q2(I,J),J=1,NU),I=1,NU),((Q12(I,J),J=
      21,NU),I=1,N),((Q0(I,J),J=1,N),I=1,N)
C
       WRITE CUTPUT TAPE 6,700,TO
   700 FORMAT (23H1CCNTINUCUS PRCBLEM T =,1PE14.7)
       WRITE OUTPUT TAPE 6,7C1
   701 FCRMAT ( 9HCF-MATRIX)
       DO 702  I=1,N
   702 WRITE CUTPUT TAPE 6,FMT,(F(I,J),J=1,N)
       WRITE CUTPUT TAPE 6,7C3
   703 FCRMAT ( 9HCG-MATRIX)
       DO 704  I=1,N
   704 WRITE OUTPUT TAPE 6,FMT,(G(I,J),J=1,NU)
       WRITE OUTPUT TAPE 6,7C5
   705 FORMAT (11HCQ11-MATRIX)
       DO 706 I=1,N
   706 WRITE CUTPUT TAPE 6,FMT,(Q1(I,J),J=1,N)
       WRITE CUTPUT TAPE 6,7C7
   707 FCRMAT (11HCQ12-MATRIX)
       DO 708 I=1,N
   708 WRITE CUTPUT TAPE 6,FMT,(Q12(I,J),J=1,NU)
       WRITE OUTPUT TAPE 6,7C9
   709 FORMAT (11HCQ22-MATRIX)
       DO 710 I=1,NU
   710 WRITE CUTPUT TAPE 6,FMT,(Q2(I,J),J=1,NU)
       WRITE CUTPUT TAPE 6,711
   711 FORMAT (1CHCQC-MATRIX)
       DO 712 I=1,N
   712 WRITE CUTPUT TAPE 6,FMT,(Q0(I,J),J=1,N)
C
C      CALCULATICN CF SIGMATILCE AND E
C
```

PROGRAM SAMPAS - Continued

```
 SAMPAS

      CALL FCRMA
C
      M=N+N
      M1=M+M
      DO 4 I=1,M1
      DO 4 J=1,M1
    4 A(I,J)=-TC*A(I,J)
C
      CALL EXP(B,A,M1)
C
C
C     CCMPUTC S
C
      DO 1C I = 1,N
      DC 10 J = 1,N
      R = 0.C
      K = N + I
      DO 11 N1 = 1,N
      L = N + N1
   11 R = R + B(K,L)*CO(N1,J)
   1C S1(I,J) = B(K,J) + R
C
      DO 12 I = 1,N
      DC 12 J = 1,N
      R = C.C
      DO 13 N1 = 1,N
      L = N + N1
   13 R = R + B(I,L)*CO(N1,J)
   12 S(I,J) = R + B(I,J)
C
      CALL INVS(S,S2,16,N)
C
   15 DC 16 I = 1,N
      DO 16 J = 1,N
      R = 0.C
      DO 17 K = 1,N
   17 R = R + S1(I,K)*S2(K,J)
   16 S(I,J) = R
C
C     CCMPUTC C
C
      DC 18 I=1,M
      DC 18 J=1,M
      K=I+M
   18 A(I,J)=E(K,J)
C
      DO 20 I = 1,N
      DO 20 J = 1,N
      R = 0.C
      DO 21 K = 1,N
      L = N + K
   21 R=R+A(I,L)*CC(K,J)
   2C F(I,J)=A(I,J)+R
C
      DC 22 I = 1,N
      DC 22 J = 1,N
```

PROGRAM SAMPAS - Continued

```
 SAMPAS

       K = N + I
       R = 0.0
       DO 23 N1 = 1,N
       L = N + N1
    23 R = R + A(K,L)*CC(N1,J)-S(I,N1)*F(N1,J)
    22 G(I,J) = A(K,J) + R
C
       DO 24 I = 1,N
       DO 24 J = 1,N
       R = 0.0
       DO 25 K = 1,N
    25 R = R + G(I,K)*S2(K,J)
    24 C(I,J) = R/12.0
C
       WRITE OUTPUT TAPE 6,715
   715 FORMAT (9HOS-MATRIX)
       DO 716 I = 1,N
   716 WRITE OUTPUT TAPE 6,FMT, (S(I,J),J = 1,N)
       WRITE OUTPUT TAPE 6,717
   717 FORMAT (12HOC-MATRIX/12)
       DO 718 I = 1,N
   718 WRITE OUTPUT TAPE 6,FMT,(C(I,J), J = 1,N)
C
C      SYMMETRIZATION OF MATRICES
C
       DO 30 I = 1,N
       DO 30 J = 1,N
       F(I,J) = 0.5 * (C(I,J) + C(J,I))
    30 G(I,J) = 0.5 * (S(I,J) + S(J,I))
C
C      FORM TRIANGULAR RESOLUTION OF S MATRIX
C
       CALL INVPD (G,S1,16,N,DET,0)
C
       DO 32 I = 2,N
       N1 = I - 1
       DO 32 J = 1,N1
    32 G(I,J) = 0.0
C
       DO 34 I = 1,N
       DO 34 J = 1,N
       R = 0.0
       DO 35 K = 1,N
       DO 35 L = 1,N
    35 R=R+G(K,I)*F(K,L)*G(L,J)
    34 C(I,J) = R
C
       CALL LMAX(C,N,RL,CO,M2)
C
       WRITE OUTPUT TAPE 6,760,RL
   760 FORMAT (16HCLAMBDANULL/12 =,1PE14.7)
       GO TO 1
       END(1,0,0,0,0,0,1,0,0,0,0,0,0,0,0)
```

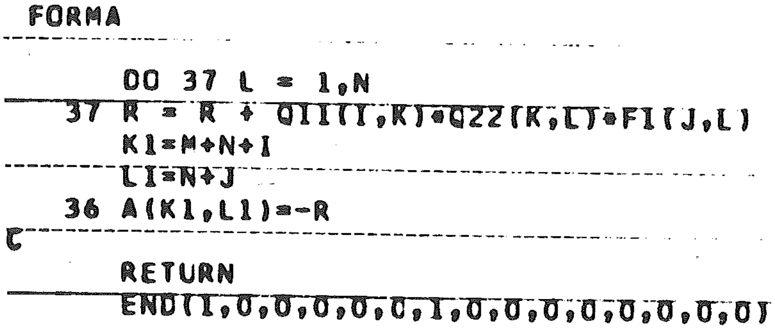
PROGRAM FORMA - FORTRAN LISTING

```
 FORMA

       SUBROUTINE FORMA
C
       DIMENSION F(16,16),G(16,16),Q1(16,16),Q12(16,16),Q2(16,16)
       DIMENSION B1(16,16),B2(16,16),F1(16,16),Q11(16,16),Q22(16,16)
       DIMENSION A(64,64)
       COMMON A,F,G,Q1,Q12,Q2,N,NU
C
       DO 10 I = 1,NU
       DO 10 J = 1,NU
    10 B1(I,J) = Q2(I,J)
C
       CALL NORM (B1,NU,GORM,16)
       CALL INVPD (B1,B2,16,NU,DET,1)
       R = DET/GORM
       IF(R*1.0E6-1.0) 11,11,12
    11 WRITE OUTPUT TAPE 6,700,R
   700 FORMAT (30H0Q22 ILLCONDITIONED DET/NORM =,1XF10.8)
C
    12 DO 14 I = 1,N
       DO 14 J = 1,N
       R = 0.0
       DO 15 K = 1,NU
       DO 15 L = 1,NU
    15 R = R+G(I,K)*B2(K,L)*Q12(J,L)
    14 F1(I,J) = F(I,J)-R
C
       DO 16 I = 1,N
       DO 16 J = 1,N
       R = 0.0
       DO 17 K = 1,NU
       DO 17 L = 1,NU
    17 R = R + Q12(I,K)*B2(K,L)*Q12(J,L)
    16 Q11(I,J) = Q1(I,J) - R
C
       DO 18 I = 1,N
       DO 18 J = 1,N
       R = 0.0
       DO 19 K = 1,NU
       DO 19 L = 1,NU
    19 R = R + G(I,K)*B2(K,L)*G(J,L)
    18 Q22(I,J) = R
C
C      FORM DIAGONAL SUBMATRICES OF A
C
       M=N+N
       DO 20 I = 1,N
       DO 20 J = 1,N
       K=M+I
       L=M+J
       A(K,L)=F1(I,J)
    20 A(I,J) = F1(I,J)
C
       DO 21 I = 1,N
       DO 21 J = 1,N
       K=J+N
```

PROGRAM FORMA - Continued

```
 FORMA

      K1=I+M
      L1=K+M
      A(K1,L1)=-Q22(I,J)
   21 A(I,K) = -Q22(I,J)
C
      DO 22 I = 1,N
      DO 22 J = 1,N
      K=N+I
      K1=M+K
      L1=M+J
      A(K1,L1)=-Q11(I,J)
   22 A(K,J)=-Q11(I,J)
C
      DO 23 I = 1,N
      DO 23 J = 1,N
      K = N + I
      L = N + J
      K1=K+M
      L1=L+M
      A(K1,L1)=-F1(J,I)
   23 A(K,L)=-F1(J,I)
C
C     FORM SUBMATRIX A(2,1)
C
      DO 30 I = 1,N
      DO 30 J = 1,N
      R = 0.0
      DO 31 K = 1,N
      DO 31 L = 1,N
   31 R = R + F1(I,K)*Q22(K,L)*Q11(L,J)
      K1=I+M
   30 A(K1,J)=R
C
      DO 32 I = 1,N
      DO 32 J = 1,N
      R = 0.0
      DO 33 K = 1,N
      DO 33 L = 1,N
   33 R = R + F1(I,K)*Q22(K,L)*F1(J,L)
      K1=I+M
      L1=J+N
   32 A(K1,L1)=R
C
      DO 34 I = 1,N
      DO 34 J = 1,N
      R = 0.0
      DO 35 K = 1,N
      DO 35 L = 1,N
   35 R = R + Q11(I,K)*Q22(K,L)*Q11(L,J)
      K1=M+N+I
   34 A(K1,J)=-R
C
      DO 36 I = 1,N
      DO 36 J = 1,N
       R = 0.0
      DO 37 K = 1,N
```

PROGRAM FORMA - Continued



```
 FORMA

       DO 37 L = 1,N
    37 R = R + Q11(I,K)*Q22(K,L)*F1(J,L)
       K1=M+N+I
       L1=N+J
    36 A(K1,L1)=-R
C
       RETURN
       END(1,0,0,0,0,0,1,0,0,0,0,0,0,0,0)
```

REFERENCES:

1. Bellman, R.: Adaptive Control Processes, A Guided Tour, Princeton University Press, Princeton, N. J., 1961.

2. Bertram, J. E., Sarachik, P. E.: "On Optimal Computer Control", Proc. IFAC Congress, Moscow, 1960.

3. Gunckel, T. L., Franklin, G. F.: "A General Solution for Linear Sampled Data Control", Journal of Basic Engineering, 1963 (to appear).

4. Henrici, P.: Discrete Variable Methods in Ordinary Differential Equations, John Wiley & Sons, New York - London, 1961.

5. Joseph, P. D., Tou, J. T.: "On Linear Control Theory", AIEE Trans., Vol. 80, Part II, 1961.

6. Kalman, R. E.: "Contributions to the Theory of Optimal Control", Bol. Soc. Mat. Mex., 1961.

7. Kalman, R. E.: "New Methods and Results in Linear Prediction and Filtering Theory", RIAS Report TR 61-1, 1961. (Also in Proceedings of the First Symposium on Engineering Applications of Random Function Theory and Probability, Bogdanoff, J. L., editor, John Wiley & Sons, 1963.)

8. Kalman, R. E., Koepcke, R. E.: "Optimal Control of Linear Sampling Control Systems using Generalized Performance Indexes", Trans. Amer. Soc. Mech. Engrs. 80, 1958, p. 1820.

9. Penrose, R.: "A Generalized Inverse for Matrices", Proc. Cambridge Phil. Soc., 51, 1955, pp. 406-413.

10. Ragazzini, J. R., Franklin, G. F.: Sampled-Data Control Systems, McGraw-Hill Book Co., Inc., New York, 1958.