PROPERTIES AND ORGANIZING CAPABILITY OF STATISTICAL ORGANIZER

Shoichi Noguchi and Juro Oizumi

Research Institute of Electrical Communication
Tohoku Univ., Sendai Japan

Summary

This paper presents some fundamental properties of a statistical organizer that are considered from the view point of system organization and organizing capability. The system performance is evaluated by means of the concept of the information processing capacity under the assumption that the pattern is distributed according to the multivariate normal distribution.

In the two-category problem , the typical statistical organizer, the Bayes organizer and the regression organizer show the same organizing capability in the first-stage approximation and the information processing capacity becomes 1 for each case.

The organizing capability of the organizer depends upon the information which is uthzed for the system organization.

The relationships among them are analyzed theoretically, and if the system is organized by the combination of subcovarlance matrices of rank N (1 = l s) obtained by the samples, the information processing capacity is reduced to max (N1/N) (l - l s).

The relation of the organizing capability between the linear organizer and the nonlinear organizer is obtained theoretically for some cases.

The same considerations are also applied to the typical R category organizers and the information processing capacity becomes R if R times weighting elements arc adopted. Several kinds of the computer simulations are obtained to check the theoretical results. These results coincide well with the theoretical ones.

Introduction

In the self-organizing system, there are two typical organizers: the non-statistical and statistical. The non-statistical organizer is represented by Perceptron type procedure. The properties of this organizer, such as the capability of the information processing and the concrete method for the organization are reported in several papers. Although this system shows high capability in information processing, it has some disadvantages as follows:

(a) If the input patterns are not linearly separable, the organizing procedure does not converge and to check the linear separability, numerous trials are sometimes required.

(b) Even though the patterns are linearly separable, numerous trials and large memories are also sometimes required.

In contrast to the disadvantages of non-statistical method, the statistical method has many merits from the practical point of view.

These merits are summarized as follows:

The system can be organized in an optimal way even when the input patterns are not linearly separable, (b) the organization of the system can be made quickly by the analogue technique.

Several kinds of the statistical organizers have been developed.

However, the capability of the organizer to process information should be further investigated. From this standpoint, we took the theoretical evaluation of the statistical organizer into consideration. We introduced some new concepts of the information processing capacity and the percent information processing capacity, as in the case of the nonstatistical organizers.

The organizing capability of each organizer is evaluated by these concepts.

The problems which are taken into considerations are (1) the organizing capability of the typical organizers, (2) the relation between the organizing capability and the information which is utlized for the system organization, (3) the relation of the organizing capability between the linear organizer and the nonlinear one, and (4) the properties of multi-category organizer.

1. The Properties of the Typical Statistical Organizers!*)

In this section, two typical organizers, the Bayesian organizer and the regression organizer are taken into consideration in order to make clear the general behaviors of the statistical organizer. Our basic assumptions are as follows:

(a) The patterns are classified into two categories, I and II.

(b) Both categories have a member of—jf- sample patterns, respectively, (We can also extend our analysis with slight modification to the various cases of different sample patterns in each category.)

(c) Each pattern is represented by an N dimen-

sional vector $\underline{X}$ and each $\underline{X}$ is distributed according to multivariate normal distribution $N(\underline{\mu}_j, \underline{\Sigma})$ (j = 1, 2) ($\underline{\mu}_j$ is an N dimensional unknown vector and $\underline{\Sigma}$ is an N x N unknown matrix) and
(d) The sample covariance matrix is assumed to be full rank.

1.1. The Bayesian Organizer

In this model, the fundamental principle is to find i which minimizes $C_{\underline{X}(i)}$ when the input pattern $\underline{X}$ is applied;
where

$$C_{\underline{X}(i)} = \sum_{j=1}^{2} \lambda(i|j)P(\underline{X}|j)P(j) \qquad (1.1)$$

and $\lambda(i|j)$ is a loss function assigned to this system.
For simplicity, we assume $\lambda(i|j)$ by

$$\lambda(i|j) = 1 - \delta_{i,j} . \qquad (1.2)$$

Estimating the mean and the covariance matrix by unbiased estimators, one obtains the decision function which decides the region of classification into I, R_1 and into II, R_2 as follows:

$$R_1 ; \underline{X}^t \underline{S}^{-1} \underline{\delta} - \frac{1}{2}(\underline{\bar{X}}^{(1)} + \underline{\bar{X}}^{(2)})^t \underline{S}^{-1} \underline{\delta} \geq 0 \qquad (1.3)$$

$$R_2 ; \underline{X}^t \underline{S}^{-1} \underline{\delta} - \frac{1}{2}(\underline{\bar{X}}^{(1)} + \underline{\bar{X}}^{(2)})^t \underline{S}^{-1} \underline{\delta} < 0 \qquad (1.4)$$

$$\text{where} \quad \underline{\delta} = \underline{\bar{X}}^{(1)} - \underline{\bar{X}}^{(2)} \qquad (1.5)$$

$\underline{\bar{X}}^{(j)}$ (j = 1, 2) are the sample means for each category, and $\underline{S}$ is a sample covariance matrix.

1.2. The Regression Organizer

In this system, we assign the desired output $Y_a^{(j)}$ to each pattern $\underline{X}_a^{(j)}$ ($j = 1, 2$, $a = 1 \sim \frac{M}{2}$) and organize the weight vector $\underline{W}$ and the threshold W_o so as to minimize the following Q function;

$$Q = \sum_j \sum_a \left\{ Y_a^{(j)} - \underline{W}^t \underline{X}_a^{(j)} - W_o \right\}^2 . \qquad (1.6)$$

The optimum values, $\underline{W}$ and W_o are obtained by

$$\underline{W} = \underline{V}^{-1} \underline{U} , \quad W_o = \bar{Y} - \underline{W}^t \underline{\bar{X}} : \qquad (1.7)$$

$$\text{where} \quad \underline{V} = \underline{S} + \frac{M}{4} \underline{\delta}\underline{\delta}^t , \quad \underline{U} = \sum_j \sum_a (Y_a^{(j)} - \bar{Y})(\underline{X}_a^{(j)} - \underline{\bar{X}}) \qquad (1.8)$$

$$\bar{Y} = \frac{1}{M} \sum_j \sum_a Y_a^{(j)} , \quad \underline{\bar{X}} = \frac{1}{M} \sum_j \sum_a \underline{X}_a^{(j)} .$$

and $\underline{S}$ is a sample covariance matrix.
The final decision function which decides the region of classification into I, R_1 and into II, R_2 is obtained by

$$R_1 ; \underline{W}^t \underline{X} + W_o \geq 0 ,$$
$$R_2 ; \underline{W}^t \underline{X} + W_o < 0 . \qquad (1.9)$$

1.3. The Properties of the Bayesian Organizer

If the sample pattern is applied to the decision function, the distribution of this output is approximated by the uninormal distribution:

$$N(\bar{\mu}_j , \bar{T}) \quad (j = 1, 2) \qquad (1.10)$$

where $\bar{\mu}_1 = \frac{1}{2} \underline{\delta}^t \underline{S}^{-1} \underline{\delta}$, $\bar{\mu}_2 = -\frac{1}{2} \underline{\delta}^t \underline{S}^{-1} \underline{\delta}$,

and $\bar{T} = \underline{\delta}^t \underline{S}^{-1} \underline{\delta}$. $\qquad (1.11)$

We define a new parameter d_b by

$$d_b = \bar{\mu}_1 - \bar{\mu}_2 = \underline{\delta}^t \underline{S}^{-1} \underline{\delta} . \qquad (1.12)$$

This parameter d_b means the measure of the separation of two categories. The distribution of this parameter is a kind of T^2 distribution derived by Hotelling.[2]
The probability density function of d_b is obtained by

$$f_{N(\lambda)}(d_b) = \sum_{k=0}^{\infty} \frac{e^{-\lambda} \lambda^k}{k!} \frac{\Gamma(\frac{m_1 + m_2}{2} + k)}{\Gamma(\frac{m_1}{2} + k)\Gamma(\frac{m_2}{2})} \cdot$$

$$\frac{1}{4}\left(\frac{d_b}{4}\right)^{\frac{m_1}{2} + k - 1} \left(1 + \frac{d_b}{4}\right)^{-\left(\frac{m_1 + m_2}{2} + k\right)} , \qquad (1.13)$$

where

$$m_1 = N , \quad m_2 = M - N - 1 , \quad \lambda = \frac{1}{2} \underline{D}^t \underline{\Sigma}^{-1} \underline{D} \qquad (1.14)$$

$\underline{D} = \sqrt{\frac{M}{4}} (\underline{\mu}_1 - \underline{\mu}_2)$ and $\Gamma(x)$ is a gamma function.

The relation between the correct recognition probability R_b and d_b is given by

$$R_b = \Phi\left(\frac{1}{2} \sqrt{d_b}\right) \qquad (1.15)$$

where $\Phi(x)$ is an error function.

1.4. The Properties of the Regression Organizer

If the input patterns are applied to this system, the distribution of the output is approximated by the uninormal distribution $N(\tilde{\mu}_j , \tilde{T})$. ($j = 1, 2$); The assignment of the desired output $Y_a^{(j)}$ is an important problem. However putting $Y_a^{(1)} = 1$ and $Y_a^{(2)} = -1$ proved to be an optimum assignment.

In this case, $\tilde{\mu}_j$ and $\tilde{T}$ are given by

$$\tilde{\mu}_1 = \frac{M}{4}\underline{\delta}^t \underline{V}^{-1}\underline{\delta} \quad , \quad \tilde{\mu}_2 = -\frac{M}{4}\underline{\delta}^t \underline{V}^{-1}\underline{\delta}$$

$$\text{and} \quad \tilde{T} = \frac{M}{8}\underline{\delta}^t \underline{V}^{-1}\underline{\delta}(2 - \frac{M}{2}\underline{\delta}^t \underline{V}^{-1}\underline{\delta}). \qquad (1.16)$$

A new parameter d_r is defined by

$$d_r = \frac{M}{2}\underline{\delta}^t \underline{V}^{-1}\underline{\delta} \quad . \qquad (1.17)$$

The correct recognition probability is obtained by

$$R_r = \oint (\sqrt{\frac{d_r}{2 - d_r}}) \quad . \qquad (1.18)$$

After some manipulations, it is proved that d_b is the unique eigenvalue of the following equation,

$$| \underline{\delta}_o \underline{\delta}_o^t - \lambda \underline{S}_o | = 0 \quad , \qquad (1.19)$$

and d_r,

$$| \underline{\delta}_o \underline{\delta}_o^t - f\underline{B} | = 0 \quad , \qquad (1.20)$$

where

$$\underline{\delta}_o = \sqrt{\frac{M}{4}}\,\underline{\delta} \quad , \quad \underline{S}_o = (M - 2)\underline{S} \qquad (1.21)$$

and $\underline{B} = \underline{S}_o + \underline{\delta}_o \underline{\delta}_o^t$.

We obtain the important relation between d_b and d_r by

$$d_b \simeq \frac{4d_r}{2 - d_r} \quad . \qquad (1.22)$$

Substituting this relation in (1.15), we obtain the following probability density function of d_r, when $Y_a^{(1)} = 1$ and $Y_a^{(2)} = -1$;

$$f(d_r) = \sum_{k=o}^{\infty} \frac{e^{-\lambda}(\lambda)^k \; 2^{(1 - k - \frac{m_1 + m_2}{2})}}{k! \; B(\frac{m_1}{2} + k, \; \frac{m_2}{2})} \cdot$$

$$d_r^{\frac{m_1}{2} + k - 1} \; (2 - d_r)^{\frac{m_2}{2} - 1} \qquad (1.23)$$

where B(x) is a Beta function.
This is a kind of noncentral Beta distribution.

1.5. The Organizing Capability and the Information Processing Capacity of the Organizer

The probability density function of the correct classification is determined theoretically as the function of M, N, $(\mu_1 - \mu_2)$ and Σ for each organizer. But in order to have an insight in the gross, we simplify the problem by considering the representative parameters of the distribution.

We define $\bar{R}_b$ and $\bar{R}_r$ by

$$\bar{R}_b = \oint (\bar{d}_b) \quad \text{and} \quad \bar{R}_r = \oint (\bar{d}_r) \quad , \qquad (1.24)$$

where $\bar{d}_b$ and $\bar{d}_r$ are the mean value of each distribution d_b and d_r, respectively.

We define two concepts, the information processing capacity $I_n^{(b)}$ and percent processing capacity $I_{n\,\underline{(a\,\%)}}^{(b)}$ of the Bayesian organizer by the following equations:

$$I_n^{(b)} \triangleq (\frac{M}{N}), \quad \text{when } \bar{R}_b = 1 \qquad (1.25)$$

and

$$I_{n\,\underline{(a\,\%)}}^{(b)} \triangleq (\frac{M}{N}), \quad \text{when } \bar{R}_b = a\,\% \quad . \qquad (1.26)$$

This means the pattern processing capacity for one dimension. The same definition is also adopted in other organizers as discussed in the following section.

1.6. The Organizing Capability of the Bayesian and the Regression Organizer

By the definition of $\bar{R}_b$ and $\bar{R}_r$, these are obtained finally as follows:

$$\bar{R}_b = \bar{R}_r = \oint (\sqrt{(1 + \frac{2\lambda}{N}) \; / \; (\frac{M}{N} - 1)}) \quad . \qquad (1.27)$$

As the typical cases, we consider the following problems.
(1) The pattern is distributed according to $N(\underline{0}, \underline{\Sigma})$.
(2) The pattern is distributed according to $N(\frac{1}{2}\Delta\underline{I}, \underline{\Sigma}_1)$ for I categories and to $N(-\frac{1}{2}\Delta\underline{I}, \underline{\Sigma}_1)$ for II, respectively;

where $\underline{0} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$, $\underline{I} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$ and $\underline{\Sigma}_1 = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & & & \vdots \\ \vdots & & & \rho \\ \rho & \cdots & \rho & 1 \end{bmatrix}$. (1.28)

For the first case, we obtain

$$\bar{d}_b = \frac{4N}{M - N - 3} \simeq \frac{4}{\frac{M}{N} - 1} \quad \text{and} \quad \bar{d}_r = \frac{2N}{M} \quad (M - N > 3) \qquad (1.29)$$

$$\text{and} \quad \bar{R}_b = \bar{R}_r = \oint (\sqrt{1 \; / \; (\frac{M}{N} - 1)} \quad . \qquad (1.30)$$

These results imply that

$$\bar{R}_b = 1 \quad \text{and} \quad \bar{R}_r = 1 \quad \text{if} \quad M \simeq N \qquad (1.31)$$

The information processing capacity and the percent information processing capacity are given by

$$I_n^{(b)} = I_n^{(r)} = 1 \qquad (1.32)$$

and

$$I_n^{(b)}(99\ \%) = I_n^{(r)}(99\ \%) \simeq 1 + 0.184 \quad . \qquad (1.33)$$

For the second case, we obtain

$$\bar{R}_b = \bar{R}_r = \Phi\left(\sqrt{\frac{(1 + \frac{M}{4}\frac{\Delta^2}{1 + (N-1)\rho})}{\frac{M}{N} - 1}}\right) \qquad (1.34)$$

and

$$I_n^{(b)}(99\ \%) = I_n^{(r)}(99\ \%) \simeq 1 + 0.184\,(1 + \frac{M}{4} \cdot \frac{\Delta^2}{1 + (N-1)\rho}) \quad . \qquad (1.35)$$

1.8. Computer Simulations

In our simulations, the normal random patterns distributed according to $N(\underline{0}, \underline{II})$ are adopted first. ($\underline{II}$ is an unit matrix). Every point on the curve is obtained by the mean value of the ten trials.

Figures (1) and (2) show the theoretical curves and the experimental results of the mean recognition rate vs. M/N for the Bayesian organizer and the regression organizer, respectively.

The results of the computer simulations show good agreement with those of the theoretical curves. The same results are also obtained when the patterns are distributed according to $N(\frac{1}{2}\Delta\underline{I}, \underline{\Sigma}_1)$ for I category and to $N(-\frac{1}{2}\Delta\underline{I}, \underline{\Sigma}_1)$ for II, respectively.

2. The Relation between the Organizing Capability and the Total Information Which is Utlized for the System Organization

We consider the effect on the organizing capability by the information which is utlized for the system organization. In this section, the Bayesian decision rule is adopted as the basic principle for the system organization, so the mean values and the covariance matrices of the sampled patterns are only utlized for the organization. The decision function is organized so as to be the admissible linear function. [3]

The assumptions for the input pattern are the same as in the Section 1 except the assumption for the covariance matrices. Consider the typical case in which the covariance matrices $\underline{\Sigma}_1$ and $\underline{\Sigma}_2$ have the following relation.

We put

$$\underline{\mu}^{(r)} \triangleq \begin{bmatrix} \underline{\mu}_{(1)}^{(r)} \\ \vdots \\ \underline{\mu}_{(m)}^{(r)} \end{bmatrix} \qquad (2.1), \qquad \underline{\Sigma}_1 = \begin{bmatrix} \underline{\Sigma}_{11} & & \\ & \underline{\Sigma}_{12} & \underline{A} \\ \underline{B} & & \ddots \\ & & & \underline{\Sigma}_{1m} \end{bmatrix} \qquad (2.2)$$

and

$$\underline{\Sigma}_2 = \begin{bmatrix} C_1\underline{\Sigma}_{11} & & \\ & C_2\underline{\Sigma}_{12} & \underline{C} \\ \underline{D} & & \ddots \\ & & & C_m\underline{\Sigma}_{1m} \end{bmatrix} \qquad (2.3)$$

where each C_i is some constant. The rank of $\underline{\Sigma}_{1i}$ and $\underline{\Sigma}_{2i}$ are equal to N_i.

In this section, the system is organized with the partial information such as the unbiased estimators of $\underline{\mu}_{(i)}^{(r)}$ and $\underline{\Sigma}_{1i}$, respectively. These are denoted by $\underline{\bar{X}}_{(i)}^{(r)}$ and $\underline{S}_{1i}$. The information contained in the matrices $\underline{A}$, $\underline{B}$, $\underline{C}$ and $\underline{D}$ are not utlized.

The admissible linear decision function $g(\underline{X})$ is obtained as follows:

$$g(\underline{X}) = \underline{W}^t\underline{X} + W_o \qquad (2.4)$$

where

$$\underline{W} = (t_1\underline{S}_1 + t_2\underline{S}_2)^{-1} \quad , \quad \underline{\delta} = (\underline{\bar{X}}^{(1)} - \underline{\bar{X}}^{(2)})$$
$$W_o = -\underline{W}^t\underline{\bar{X}}^{(1)} + t_1\underline{W}^t\underline{S}_1\underline{W}, \quad t_1 + t_2 = 1 \qquad (2.5)$$

$$\underline{\bar{X}}^{(r)} = \begin{bmatrix} \underline{\bar{X}}_{(1)}^{(r)} \\ \vdots \\ \underline{\bar{X}}_{(m)}^{(r)} \end{bmatrix}, \quad \underline{S}_1 = \begin{bmatrix} \underline{S}_{11} & & 0 \\ & \ddots & \\ 0 & & \underline{S}_{1m} \end{bmatrix} \qquad (2.6)$$

$$\underline{S}_2 = \begin{bmatrix} C_1\underline{S}_{11} & & 0 \\ & \ddots & \\ 0 & & C_m\underline{S}_{1m} \end{bmatrix}$$

$$\underline{\bar{X}}_{(i)}^{(r)} = \frac{2}{M}\sum_{j=1}^{M/2} \underline{X}_{ij}^{(r)} \qquad (r = 1, 2) \quad ,$$

and

$$\underline{S}_{1i} = \frac{1}{M-2}\left\{\sum_{j=1}^{M/2}(\underline{X}_{(i)j}^{(1)} - \underline{\bar{X}}_{(i)}^{(1)})(\underline{X}_{(i)j}^{(1)} - \underline{\bar{X}}_{(i)}^{(1)})^t + (\frac{1}{C_i})\sum_{j=1}^{M/2}(\underline{X}_{(i)j}^{(2)} - \underline{\bar{X}}_{(i)}^{(2)})(\underline{X}_{(i)j}^{(2)} - \underline{\bar{X}}_{(i)}^{(2)})^t\right\} \qquad (2.7)$$

The output of $g(\underline{X})$ is approximated by the uni-normal distribution with the mean $\underline{W}^t\underline{X}_1 + W_o$ and the covariance matrix $\underline{W}^t\underline{S}_1\underline{W}$, if $\underline{X}$ belongs to I. The average correct recognition probability in this system is given by $\Phi(t_1\sqrt{\underline{W}^t\underline{S}_1\underline{W}})$ for the pattern belonging to I, and by $\Phi(t_2\sqrt{\underline{W}^t\underline{S}_2\underline{W}})$ for the pattern belonging to II.

Putting $t_i\sqrt{\underline{W}^t\underline{S}_i\underline{W}} = y_i \quad (i = 1, 2)$, it is proved after some calculations that y_i^2 is proportionally distributed according to the sum of the noncentral F distributions.

The mean value $\bar{y}_1^2$ of y_1^2 and $\bar{y}_2^2$ of y_2^2 are obtained, respectively, as follows:

$$\bar{y}_1^2 = \sum_{i=1}^{m} \frac{2(M-2)(1+C_i)t_1^2}{M(t_1+C_it_2)^2} \cdot \frac{N+\psi_i}{M-N_i-3} \qquad (2.8)$$

and

$$\bar{y}_2^2 = \sum_{i=1}^{m} \frac{2(M-2)}{M} \cdot \frac{(1+C_i)C_it_2^2}{(t_1+C_it_2)^2} \cdot \frac{N+\psi_i}{M-N_i-3}, \qquad (2.9)$$

where

$$\psi_i = \frac{M}{2(1+C_i)} \underline{V}_i^t \underline{\Sigma}_{1i}^{-1} \underline{V}_i \quad \text{and} \quad \underline{V}_i = \underline{\mu}_1^{(i)} - \underline{\mu}_2^{(i)}. \qquad (2.10)$$

By the definition of the information processing capacity, I_n, in this organizer becomes:

$$I_n = \max(N_i/N) \qquad (2.11)$$

The mean recognition ability $\bar{R}_D$ is approximately given as follows if $M, N \gg 1$:

$$\bar{R}_D = \frac{1}{2}\left[\Phi\sqrt{(\bar{y}_1^2)} + \Phi\sqrt{(\bar{y}_2^2)}\right] \qquad (2.12)$$

As typical examples, consider the following cases.

(i) $t_1 = t_2$, $\underline{V}_i = \underline{0}$ $(i = 1, 2)$, $\underline{\Sigma}_i$ are the same as Eq. (2.2) and (2.3) and $N_1 = N_2 = \cdots = N_m = \frac{N}{m}$:

and

(ii) $t_1 = t_2$, $\underline{\Sigma}_1 = \underline{\Sigma}_2 = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{pmatrix}$,

$\underline{\mu}_1 - \underline{\mu}_2 = (\Delta, \cdots, \Delta)^t$ and $N_1 = N_2 = \cdots = N_m = \frac{N}{m}$.

We obtain $\bar{y}_1^2$ and $\bar{y}_2^2$ for each case as follows:

$$\text{(i)} \quad \bar{y}_1^2 = \frac{1}{\frac{M}{N} - \frac{1}{m}} \cdot \frac{1}{m} \sum_{i=1}^{m} \frac{2}{(1+C_i)} \qquad (2.13)$$

$$\text{and} \quad \bar{y}_2^2 = \frac{1}{\frac{M}{N} - \frac{1}{m}} \left(\frac{1}{m} \sum_{i=1}^{m} \frac{2C_i}{(1+C_i)}\right) \qquad (2.14)$$

$$\text{(ii)} \quad \bar{y}_1^2 = \bar{y}_2^2 = \left(1 + \frac{M}{4} \cdot \frac{2}{m + (N-m)\rho}\right) \Big/ \left(\frac{M}{N} - \frac{1}{m}\right) \qquad (2.15)$$

$$\text{and} \quad \bar{R}_D = \Phi\left(\sqrt{\left(1 + \frac{M}{4} \cdot \frac{2}{m + (N-m)}\right) / \left(\frac{M}{N} - \frac{1}{m}\right)}\right). \qquad (2.16)$$

The results obtained by these analyses show the very interesting fact that if the distributions of both patterns have the covariance matrices as Eq. (2.2) and (2.3), the organizing capability of this system is limited to $\max(N_i/N)$. The relation between the organizing capability and the total information which is utilized to organize the system is obtained by Eq. (2.16) for the case (ii). This result implies that the information processing capacity decreases to $\frac{1}{m}$ and the curve $\bar{R}_D$ vs. M/N decreases more loosely with the increase of m.

The minimum distance classifier is the most extreme case stated above. In this organizer, the information contained in the covariance matrix is not utilized at all and the decision is based only upon the difference of the means of two distributions. The information processing capacity becomes zero in this organizer.

The theoretical curves for $\bar{R}_D$ vs. M/N are given in Fig. 3 for the typical cases of m with computer simulations.

3. The Relation of the Organizing Capability between the Linear Organizer and the Nonlinear Organizer[4]

As is well known, the optimum Bayesian organizer becomes the quadratic form if the patterns belong to $N(\underline{\mu}_j, \underline{\Sigma}_j)$ $(j = 1, 2)$ and the covariance matrices are different.

Because the general consideration of the organizing properties, when $\underline{\Sigma}_1 \neq \underline{\Sigma}_2$, is very complicated, we treat the simple case where $\lambda^2\underline{\Sigma}_2 = \underline{\Sigma}_1$ as a first step.

The relation of the organizing capability between the linear organizer and the nonlinear organizer are made clear in the following discussions.

We investigate the characteristics of the linear admissible organizer and the optimum nonlinear organizer separately and then summarize them.

The assumptions for the pattern are the same as in the Section 1 except that the covariance matrices are different. We assume that $\Sigma_2 = \Sigma$ and that λ^2 is known.

3.1. The Properties of the Admissible Linear Organizer

Assigning the loss function $\lambda(i \mid j)$ as $1 - \delta_{ij}$ and minimizing the system loss, the admissible linear decision function $g(\underline{X})$ is obtained as follows;

$$g(\underline{X}) = \underline{\delta}^t \underline{S}^{-1} \underline{X} - \frac{1}{\lambda + 1} \underline{\delta}^t \underline{S}^{-1} (\bar{\underline{X}}^{(1)} + \lambda \bar{\underline{X}}^{(2)}) \quad (3.1)$$

where

$$\underline{\delta} = \bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)}$$

and

$$\underline{S} = \frac{1}{M-2} \left\{ \frac{1}{\lambda^2} \sum_{a=1}^{M/2} (\underline{X}_a^{(1)} - \bar{\underline{X}}^{(1)})(\underline{X}_a^{(1)} - \bar{\underline{X}}^{(1)})^t + \sum_{a=1}^{M/2} (\underline{X}_a^{(2)} - \bar{\underline{X}}_a^{(2)})(\underline{X}_a^{(2)} - \bar{\underline{X}}^{(2)})^t \right\} \quad (3.2)$$

Considering that the output of $g(\underline{X})$ is distributed approximately according to uninormal distribution, the recognition probability R_L is given as follows:

$$R_L = \Phi\left(\frac{1}{\lambda + 1} \sqrt{\hat{d}_b}\right)$$

where $$\hat{d}_b = \underline{\delta}^t \underline{S}^{-1} \underline{\delta} \quad . \quad (3.3)$$

After calculations, it is proved that $\frac{M-N-1}{N} \cdot \left(\frac{1}{M-2} \frac{M}{2} \frac{1}{\lambda^2+1}\right) \hat{d}_b$ is distributed according to noncentral F distribution and the mean of $\hat{d}_b$ is obtained as follows;

$$E\left\{ f(\hat{d}_b) \right\} \simeq 2(\lambda^2 + 1) \frac{1 + \frac{\Psi}{N}}{\frac{M}{N} - 1} \quad (M, N \gg 1) \quad (3.4)$$

where

$$\Psi = \frac{1}{\lambda^2 + 1} \frac{M}{2} (\mu_1 - \mu_2)^t \Sigma^{-1} (\mu_1 - \mu_2) \, . \quad (3.5)$$

The mean recognition ability $\bar{R}_L$ is obtained as follows;

$$\bar{R}_L = \Phi\left(\sqrt{2 \frac{\lambda^2 + 1}{(\lambda + 1)^2} \frac{1 + \frac{\Psi}{N}}{\frac{M}{N} - 1}} \right) \quad (3.6)$$

The analysis given here shows that the influence of λ is slight; $\bar{R}_L$ becomes minimum at $\lambda = 1$ and shows a slight increase with a change of λ.
The relation between $\bar{R}_L$ and M/N when $\Psi = 0$ is given in Fig. 4 with some computer simulations.

3.2. The Properties of the Optimum Nonlinear Organizer

Consider the properties of the optimum nonlinear organizer, assuming the same input patterns as in Section 3.1 :
We consider the case $\lambda > 1$, but the same results are also true for $\lambda < 1$.
The basic decision function is obtained as follows:

$$g(\underline{X}) = \frac{\lambda^2 - 1}{2\lambda^2} \left(\underline{X} + \frac{1}{\lambda^2 - 1} \bar{\underline{X}}^{(1)} - \frac{\lambda^2}{\lambda^2 - 1} \bar{\underline{X}}^{(2)} \right)$$

$$\underline{S}^{-1} \left(\underline{X} + \frac{1}{\lambda^2 - 1} \bar{\underline{X}}^{(1)} - \frac{\lambda^2}{\lambda^2 - 1} \bar{\underline{X}}^{(2)} \right)$$

$$- \frac{1}{2} \frac{1}{\lambda^2 - 1} (\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)})^t \underline{S}^{-1} (\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)}) - N \log \lambda \, . \quad (3.7)$$

where $\bar{\underline{X}}^{(j)}$ $(j = 1, 2)$ are the sample means and $\underline{S}$ is the same as Eq. (3.2).
The approximate distribution of $g(\underline{X})$ is theoretically obtained for both cases when $\underline{X}$ belongs to I or to II, respectively. These are expressed as the sum of the distribution functions.
Substituting the mean values for these functions, the mean recognition ability $\bar{R}_N$ of this organizer is obtained.
This is the first stage approximation of the capability of this organizer.
After troublesome derivations, $\bar{R}_N$ is obtained as follows;

$$\bar{R}_N = \frac{1}{2} + \frac{1}{2} \sum_{k=0}^{\infty} \frac{e^{-\frac{\bar{\Psi}_1}{2}} \left(\frac{\bar{\Psi}_1}{2}\right)^k}{k!} \cdot \int_0^{\bar{k}_2} \frac{1}{2\Gamma(\frac{N}{2} + k)} \left(\frac{y}{2}\right)^{\frac{N}{2}+k-1} e^{-\frac{y}{2}} dy$$

$$- \frac{1}{2} \sum_{k=0}^{\infty} \frac{e^{-\frac{\bar{\Psi}_2}{2}} \left(\frac{\bar{\Psi}_2}{2}\right)^k}{k!} \cdot \int_0^{\bar{k}_1} \frac{1}{2\Gamma(\frac{N}{2} + k)} \left(\frac{y}{2}\right)^{\frac{N}{2}+k-1} e^{-\frac{y}{2}} dy \quad (3.8)$$

where

$$\bar{\Psi}_1 = \frac{2\lambda^2(\lambda^2 + 1)}{(\lambda^2 - 1)^2} \frac{1 + \frac{\Psi}{N}}{\frac{M}{N} - 1} \, ,$$

$$\bar{\Psi}_2 = \frac{2(\lambda^2 + 1)}{(\lambda^2 - 1)^2} \frac{1 + \frac{\Psi}{N}}{\frac{M}{N} - 1}$$

$$\bar{C} = \frac{\lambda^2 + 1}{\lambda^2 - 1} \frac{1 + \frac{\Psi}{N}}{\frac{M}{N} - 1} + N \log \lambda \qquad (3.9)$$

$$\bar{K}_1 = \frac{2}{\lambda^2 - 1} \bar{C} \qquad \bar{K}_2 = \frac{2\lambda^2}{\lambda^2 - 1} \bar{C}$$

and Ψ is the same as Eq. (3.5).
When $\Psi = 0$, the relation between $\bar{R}_L$ and M/N is obtained in Fig. 5 with the computer simulations.

3.3. The Relation of the Organizing Capability between the Linear Organizer and the Nonlinear Organizer

By the analyses obtained in Sections 3.1 and 3.2, the organizing capability of the nonlinear organizer also shows the same information processing capacity, $I_n = 1$. But it must be mentioned, however, that the number of the weighting element which is required to organize the system is $N(N+1)/2$. In the derivation of I_n, it is important to mention that the dimension of the input pattern N is adopted instead of the number of the weighting element.
The improvement of the organizing capability compared to the linear case is quite poor when λ^2 is near one, although this situation is improved rapidly with the increase or the decrease of λ^2.
The theoretical results and the computer simulations show that the organizing capability of the linear and the nonlinear organizer are nearly the same if $0.8 \leq \lambda^2 \leq 1.2$.
From these considerations, it is important to develop the useful nonlinear organizer for the case where $\lambda^2 \simeq 1$.

4. The Organizer for R Category Problem

The organizer decribed in the previous section is chiefly suitable fitted for the two-category problem.
Because of its complexity for the final decision, it is not always suitable for the multi-category classification. In order to improve this point, two multi-category organizers, the modified fisher organizer and the modified regression organizer, are considered.
The organizing capability of these organizers are evaluated theoretically and experimentally.
The basic assumption for the input patterns is the same as for Section 1 except that the number of the category is increased to R and the number of the pattern belonging to each category is M/R. The set of the category are composed of $\omega_1, \omega_2, \ldots, \omega_R$.
It should be mentioned that the organization of the classifier described in this section has no relation to the distribution but, for simplicity, the case $\underline{\Sigma}_1 = \underline{\Sigma}_2 = \ldots = \underline{\Sigma}_R = \underline{\Sigma}$ is considered.

4.1. The Modified Fisher Organizer

In this system, the classifier organizes the decision function for each category.
Thus, there are R organizers. The input patterns are assumed to be distributed according to $N(\underline{\mu}_j, \underline{\Sigma}_j)$, respectively. $(j = 1 \sim R)$. The linear function for each categorizer is organized to make the following function F_j maximum under the constraint such that $\underline{W}_j^t \underline{S} \underline{W}_j = C$:
where

$$F_j = \frac{\left\{ E\ (\underline{W}_j^t \underline{X}) \right\}^2}{V\ (\underline{W}_j^t \underline{X})} \quad (\underline{X} \in \omega_j) \qquad (4.1)$$

Using the method of Lagrange multiplier, and after some calculations, the decision function $g_j(\underline{X})$ for each category is obtained as follows:

$$g_j(\underline{X}) = \left| \bar{\underline{X}}^{(j)} \underline{S}_j^{-1} (\underline{X} - \bar{\underline{X}}^{(j)}) \right| \quad (j = 1 \sim R) \qquad (4.2)$$

where $$\bar{\underline{X}}^{(j)} = \frac{R}{M} \sum_{a=1}^{M/R} \underline{X}_a^{(j)}$$

and

$$\underline{S}_j = \frac{R}{M - R} \sum_{a=1}^{M/R} (\underline{X}_a^{(j)} - \bar{\underline{X}}^{(j)})(\underline{X}_a^{(j)} - \bar{\underline{X}}^{(j)})^t . \qquad (4.3)$$

By this organization, the pattern is classified into the category ω_s if $g_i(\underline{X})$ $(i = 1 \sim R)$ shows the minimum value at $i = S$.
The correct recognition probability of this organizer is obtained by calculating the conditional probability $P(i \mid 1)$ such that

$$P(i \mid 1) = P_r \left\{ |g_1| > |g_i| \mid \underline{X} \in \omega_1 \right\} . \qquad (4.4)$$

After some calculations, $P(i \mid 1)$ is approximated by the following equation :

$$P(i \mid 1) \simeq 2 \Phi\left(\sqrt{\frac{d_{1i}}{2}} \right) \left(1 - \Phi\sqrt{\frac{d_{1i}}{2}} \right) \qquad (4.5)$$

where

$$d_{1i} = \frac{R(1 + P_i)}{\left(\frac{M}{N} - R\right)} - \frac{\underline{\mu}_1^t \underline{\Sigma}^{-1} \underline{\mu}_i}{R} \left(\frac{M}{N} - R \right)$$

and

$$P_i = \frac{M}{NR} \underline{\mu}_i^t \underline{\Sigma}^{-1} \underline{\mu}_1 \qquad (4.6)$$

Thus the average recognition ability $\overline{R}_F$ is approximated by

$$\overline{R}_F \simeq 1 - \frac{2}{R} \sum_{\substack{i \\ i \neq j}} \sum_{j} \left(\Phi\left(\sqrt{\frac{d_{ij}}{2}}\right)\right)\left(1 - \Phi\left(\sqrt{\frac{d_{ij}}{2}}\right)\right) \quad (4.7)$$

As the most typical case, we consider a two-category problem, assuming that both patterns belong to the same distribution $N(\underline{0}, \underline{\Sigma})$.
In this case, we obtain

$$\overline{R}_F \simeq 2\Phi\left(\sqrt{\frac{1}{\frac{M}{N} - 2}}\right) - 1 \quad \left(\frac{M}{N} \simeq 2\right). \quad (4.8)$$

By the definition of the information processing capacity, we obtain

$$I_n = \left(\frac{M}{N}\right)\overline{R}_{F=1} = 2 \quad . \quad (4.9)$$

and

$$I_{n(99\,\%)} = \left(\frac{M}{N}\right)\overline{R}_{F=0.99} = 2.150 \quad . \quad (4.10)$$

The analysis obtained here can also be extended to the R-category problem and we obtain as follows:

$$I_n = \left(\frac{M}{N}\right)\overline{R}_{F=1} = R \quad .$$

It should be mentioned that the information processing capacity for one weighting element is also one in this organizer, as in other cases discussed in the previous sections. The theoretical curve for this example is given in Fig. 6.

4.2. Modified Regression Classifier

We have already discussed the regression organizer for a two-category problem in the Section 1. This method can be straightly extended to an R-category problem.
We set the optimal super surface for each ω_j so that $|\underline{W}_j{}^t\underline{X} + W_{0j}|$ becomes minimum on the average under the constraint such that $|\underline{W}_j^t\underline{X}_j| = C$.
After some calculations, it is proved that the same decision functions are obtained for each category as shown by Eq. (4.2).
From this consideration, the results obtained in the previous section are also true for this organizer.

5. Conclusion

The properties and the organizing capability of the statistical organizer are analysed systematically by means of the information processing capacity and percent information processing capacity.
So far as these concepts are concerned, we found that the Bayesian organizer and the regression organizer show the same organizing capability in the first stage approximation and the information processing capacity becomes 1 for both cases.
The relation between the organizing capability and the total information which is utilized for the system organization is studied for some typical cases.
As one of these results, we obtain the interesting fact that if the distributions of both patterns belong to $N(\underline{\mu}_j, \underline{\Sigma})$ $(j = 1, 2)$, the information processing capacity is limited to $\max(N_i/N)$, where N_i is the rank of the subcovariance matrix which is utilized to organize the classifier.
The relation of the organizing capability between the linear classifier and the nonlinear classifier is obtained when the patterns are distributed according to $N(\underline{\mu}_1, \lambda^2\underline{\Sigma})$ and $N(\underline{\mu}_2, \underline{\Sigma})$ respectively.
The effect of λ^2 on the organizing capability is discussed. So long as the admissible linear organizer is adopted, the variation of the organizing capability is slight with the change of λ.
Although the optimal nonlinear Bayesian organizer shows a higher organizing capability than the linear organizer, the improvement of its capability is poor as long as λ^2 is near 1. This situation can be improved rapidly, however, with the increase or the decrease of λ^2.
For the R-category problem, two typical organizers - the modified fisher organizer and the modified regression organizer - are studied.
The organizing capability of these organizers are also evaluated by means of the information processing capacity.
Both organizers show the same organizing capability for the R-category problem with $I_n = R$.
This implies that the organizers can treat $R \cdot N$ patterns correctly. But the number of the weighting element required for the system organization is also $R \cdot N$. In order to check the theoretical results, several kinds of computer simulation are made for the organizing capability and the information processing capacity for each organizer using the random patterns. These results coincide well with our theoretical results.

Reference

(1) S. Noguchi, K. Nagasawa and J. Oizumi "Fundamental Consideration on Statistical Recognizer" Proceeding of the Hawaii International Conference on System Science 1968 - Jan.

(2) T. W. Anderson " An Introduction to Multivariate Statistical Analysis" John Wiley & Sons, Inc..

(3) T. W. Anderson and R. P. Bahadur
"Classification into Two Multivariate Normal Distributions With Different Covariance Matrices" Ann. of Math. Stat. vol. 33. 1962

(4) M. Sato, S. Noguchi and J. Oizumi
"Classification Capability of Statistical Linear Recognition Function (I)" The Record of Electrical and Communication Engineering Conversazione, Tohoku Univ., Sendai, Japan Vol. 37, No. 2, 1968

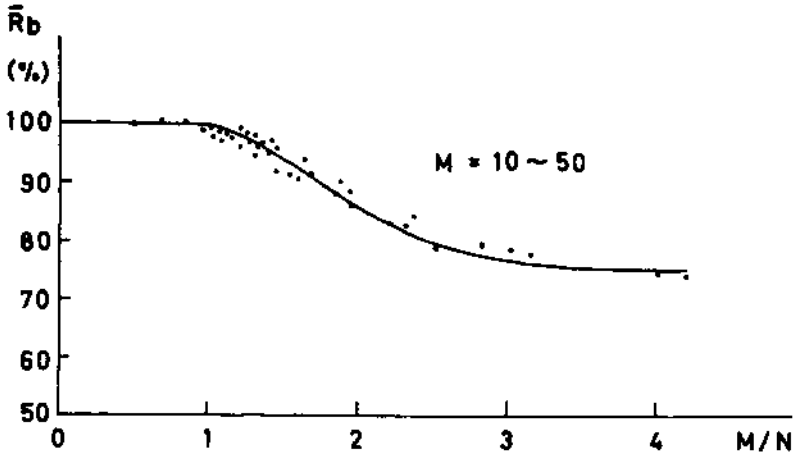

Fig. 1 The theoretical curve and the experimental results of $\bar{R}_b$ vs. M/N

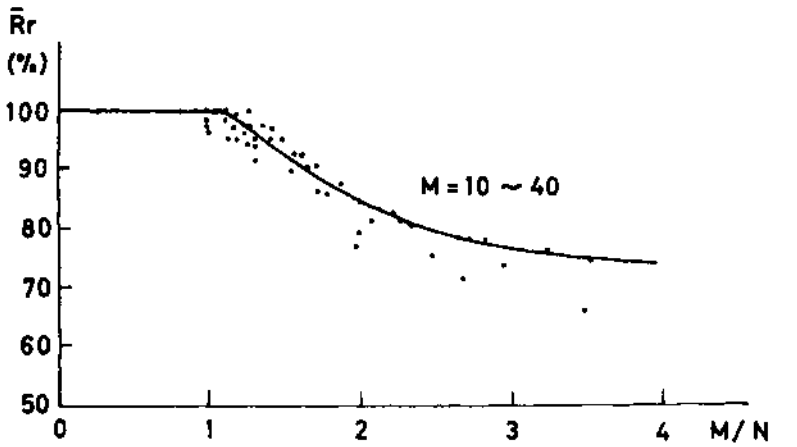

Fig. 2 The theoretical curve and the experimental results of $\bar{R}_r$ vs. M/N

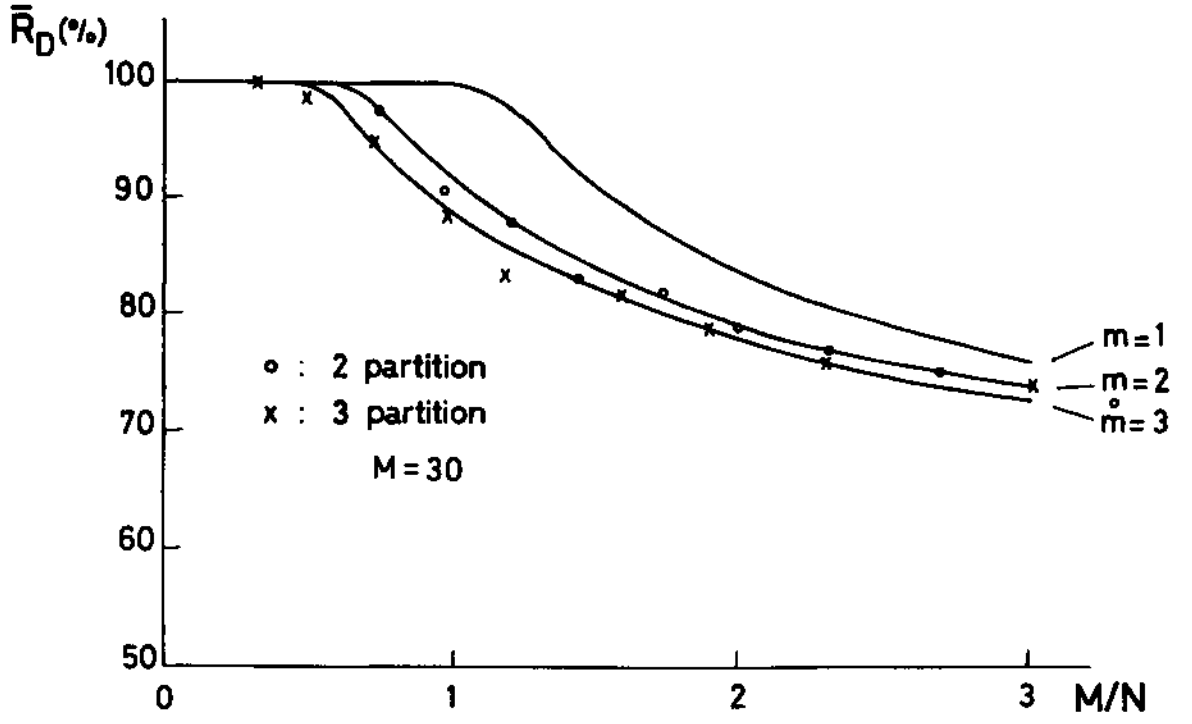

Fig. 3 The relation between $\bar{R}_D$ and M/N

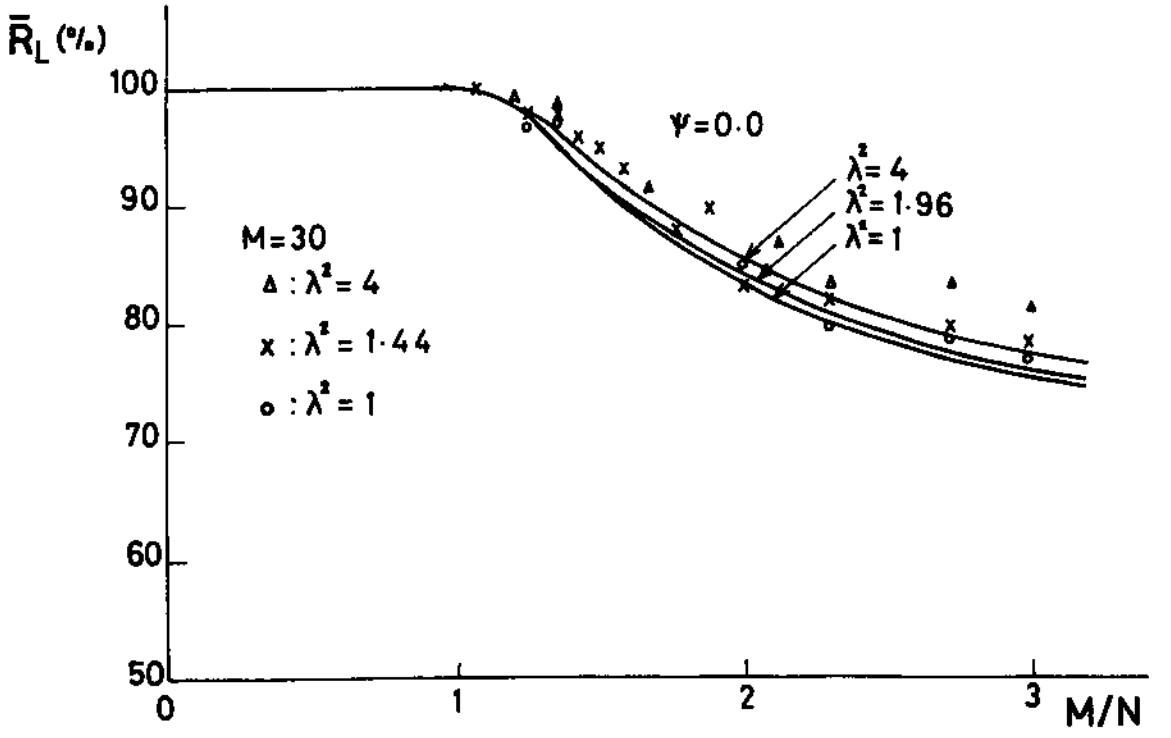

Fig. 4 The relation between $\bar{R}_L$ and M/N ($\Sigma_1=\lambda^2\Sigma_2$) and the computer simulations

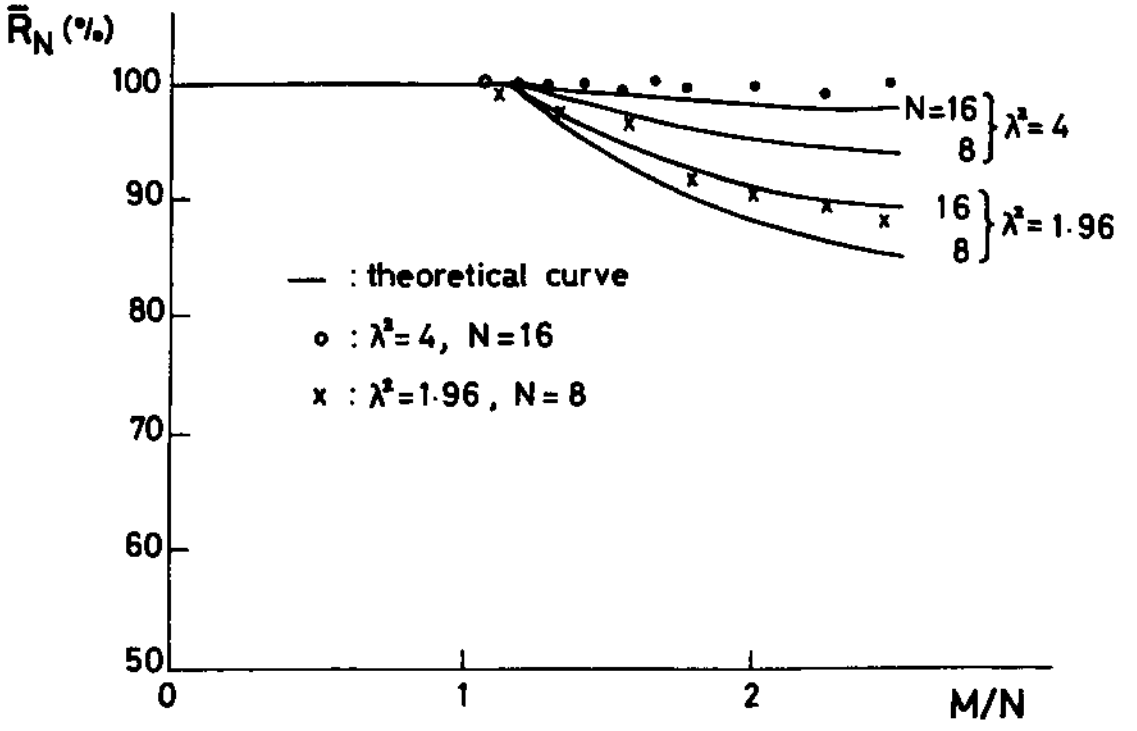

Fig. 5 The relation between $\bar{R}_N$ and M/N and the computer simulations

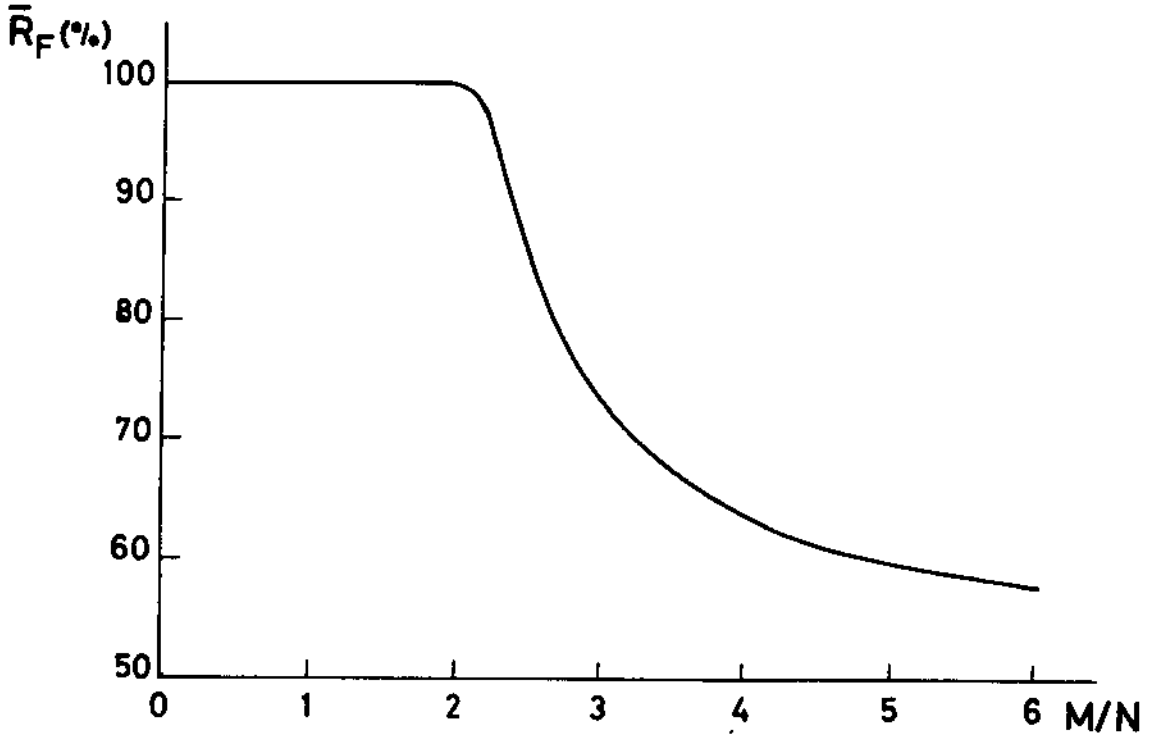

Fig. 6 The relation between $\bar{R}_F$ and M/N