Reprinted from

Ninth International Symposium

Machine Processing of

Remotely Sensed Data

with special emphasis on

Natural Resources Evaluation

June 21-23, 1983

Proceedings

Purdue University
The Laboratory for Applications of Remote Sensing
West Lafayette, Indiana 47907 USA

NONPARAMETRIC MINIMUM ERROR RATE FEATURE TRANSFORMATION WITH APPLICATION TO RESOURCE CLASSIFICATION

S. ARUNKUMAR

Indian Institute of Technology/
Computer Science Program
Bombay, India

R. SUPNEKAR

Tata Burroughs Ltd.
Bombay, India

A. ABSTRACT

In resource evaluation, the assumption of an underlying family for the distribution of patterns in a class may not be desirable from the standpoint of validity of the assumption as well as the analysis of information. Dimensionality reduction of the feature space is a desirable and, often, necessary step for economic pattern classification. A novel stochastic gradient algorithm is developed in this paper that yields an asymptotically optimal linear transformation minimizing the Bayes' error rate (probability of misclassification) in the transformed space. The algorithm does not make explicit use of the underlying distribution of the features. The algorithm is generalizable to decision functions other than Bayes' error rate and is adaptive based on supervised samples, available all at one time or collected over time in which case the algorithm is updated at each such point; the evaluation of the decision function is not required at each update. The algorithm, coded in FORTRAN and implemented on a DEC System 1077, has been applied to certain LANDSAT data sets for two and five classes as well as to an eight-class case with features having multivariate normal distribution. The results point to the robustness of the procedure.

I. INTRODUCTION

There is commonly some degree of redundancy in information in a multivariate remote sensing data set. This is particularly true for multispectral data for which a considerable degree of correlation may exist between channels that are spectral neighbors. One can expect, therefore, that a derived data set of reduced dimensionality exists which provides a classification which is at least as good as the one obtained by using the complete set. Reduction of dimensionality is desirable due to the need for data compaction and speed of processing, keeping in mind the important consideration of accuracy in resource classification. In feature space transformation, the objective is to transform the original feature space to a new space, usually in lower dimensions, that is able to contain important information.

Feature transformations that reduce dimensionality are categorized by the criteria used for determining the transformation. These criteria are listed below.

1. Preservation of information, as in principal components analysis.

2. Distance measures, such as the divergence.

3. Measures of separability, other than the above, such as Fisher's linear discriminant.

4. Preserving some structure underlying the data.

5. Probability of misclassification.

No assumption is made regarding the distribution of the features. We assume that labelled (supervised) samples are available for each class and develop a novel procedure for determining a transformation optimizing a decision function which includes the criterion 5.

The organization of the paper is as follows. Notation and a brief literature review are given in section II, the procedure is developed in section III and numerical results are given in section IV. Conclusions are given in section V.

II. NOTATION AND LITERATURE REVIEW

A. NOTATION

We use the following notation.

c number of classes

w_i i^{th} pattern class, $i = 1,\ldots,c$

d dimension before transformation

d' dimension after transformation, $d' \leq d$

T $(d' \times d)$ transformation matrix

p_i a priori probability for class w_i

$\underline{X}$ d- dimensional feature vector, whose generic realization is represented as $\underline{x}$

$\underline{Y}$ d'- dimensional feature vector obtained by transforming $\underline{X}$ by T, with its generic realization represented as $\underline{y}$

$p(\underline{x}/w_i)$ conditional density of $\underline{X}$ given class w_i

$p(\underline{x})$ probability density function of $\underline{X}$; $p(\underline{x}) = \Sigma p_i\, p(\underline{x}/w_i)$

$p(w_i/\underline{x})$ posterior probability for class w_i and equal to $p_i p(\underline{x}/w_i)/p(\underline{x})$

D_{ij} decision functions for classification; $i=1,\ldots,c$; $j=1,\ldots,c$; $i \neq j$

E expectation operator

$u(.)$ unit step function; equal to 1 when the argument is positive and zero otherwise

$h(.)$ smoothing function

B. LITERATURE REVIEW

The principal components[1] transformation minimizes the loss function corresponding to the extent to which we can predict $\underline{X}$ given $\underline{Y}$. Although the transformation is optimal with respect to fitting the data, it is not necessarily optimal for separating the classes. Decell and Mayekar[2] give an iterative procedure for obtaining T such that the divergence in transformed space is minimized. For $c=2$ and $d'=1$, Fisher's linear discriminant[3] tries to maximize the separation in the transformed space. Using the Fisher ratio as the optimality criterion and selecting features for their discriminatory potential, Foley and Sammon[4] derive an algorithm for extracting a set of features for a 2-class problem. Also for $c=2$, Fukanaga and Koontz[5] suggest a preliminary transformation such that the eigen vectors which best fit class 1 are the poorest for class 2. Odell et. al.[6] develop an explicit expression for a transformation T for which using the Bayes' classification procedures $\underline{x}$ is assigned to w_i if and only if $T\underline{x}$ is assigned to w_i, where $d' \leq d$ is the smallest integer for which this equivalence holds. T is calculated directly in terms of known class means μ_i and covariance matrices Σ_i. Lissack and Fu[7] describe a 2-class feature extracting transformation when the error rate is to be minimized; explicit results are given for Gaussian distributions. Nonlinear mapping algorithms attempt to preserve some structure underlying the data. Sammon[8] gives an iterative procedure for a good match between samples in d-space with their transformations in d'-space; a variation to reduce computational requirement was given by White[9], where the Hamming metric is used in place of the Euclidean metric. A relaxation method for nonlinear mappings has been proposed by Chang and Lee[10]. A nonlinear transformation to make features in the transformed space linearly separable has been proposed by Calvert and Young[11]. In respect of criterion 5, de Figueiredo[12] suggests a procedure, specialized by Starks et. al.[13], based on mathematical programming when the distribution of the features are known.

III. NONPARAMETRIC FEATURE TRANSFORMATION

We now consider the problem of finding a linear transformation that minimizes certain decision functions including the error rate, when the underlying distributions are unknown. Our approach leads to a procedure that is optimal in the large sample sense.

A. EXPRESSION FOR THE ERROR PROBABILITY

Given real-valued decision functions $\{D_{ij}\}$, for a feature vector $\underline{X} = \underline{x} \in R_d$, the d-dimensional Euclidean space, the classification is according to the following rule: $\underline{x}$ is assigned to class w_i if $D_{ij}(\underline{x}) > 0, \forall j \neq i$.

The probability P_e of misclassification can then be represented as:

$$P_e(\{D_{ij}\}) = 1 - \sum_{i=1}^{c} \int_{R_d} p(w_i/\underline{x}) \left[\prod_{\substack{j=1 \\ j \neq i}}^{c} u(D_{ij}(\underline{x})) \right] p(\underline{x})\, d\underline{x} \qquad (1)$$

The choice of

$$D_{ij}(\underline{x}) = p(w_i/\underline{x}) - p(w_j/\underline{x}) \qquad (2)$$

yields the Bayes' error probability and hence is optimal for the error rate criterion.

We are interested in reducing the dimensionality of the feature vector $\underline{X}$. Let T be a (d'xd) matrix and $\underline{Y} = T\underline{X}$ be the transformed feature vector. If we wish to minimize the Bayes' error rate with respect to T, the decision functions are given by

$$D_{ij}(\underline{x}) = p(w_i/T\underline{x}) - p(w_j/T\underline{x}) \qquad (3)$$

where $p(w_i/T\underline{x})$ represents the posterior probability given transformed feature values. In the classification problem, the posterior probabilities are unknown. We are given a sequence of independent pairs $(\underline{x}_1,\rho_1), (\underline{x}_2,\rho_2),\ldots,(\underline{x}_n,\rho_n),\ldots$ where $\{\underline{x}_i\}$ are randomly selected points from R_d with probability density $p(\underline{x})$ and $\{\rho_i\}$ give the class label of the respective samples. Note that

$$P(\rho_n = i / \underline{X} = \underline{x}) = p(w_i/\underline{x}). \qquad (4)$$

B. TWO-CLASS SPECIALIZATION

Let

$$D_B(\underline{x}) = p(w_1/\underline{x}) - p(w_2/\underline{x}). \qquad (5)$$

The error probability (1) becomes

$$P_e = \frac{1}{2} - \frac{1}{2} \int_{R_d} D_B(\underline{x}) \operatorname{sgn} D_{12}(\underline{x}) p(\underline{x}) d\underline{x} \qquad (6)$$

Also, if the training sequence is given such that $\rho_n = 1$ if $\underline{x}_n \in w_1$ and $\rho_n = -1$ if $\underline{x}_n \notin w_1$, we have using (5) that

$$E(\rho_n / \underline{X}_n = \underline{x}) = D_B(\underline{x}) \qquad (7)$$

making it possible to estimate D_B from the training samples. Besides independence, this is the only property required of the sequence $(\underline{x}_n, \rho_n)$.

The decision function D_T' in the transformed space is given by

$$D_T'(T\underline{x}) = p(w_1/T\underline{x}) - p(w_2/T\underline{x}) \qquad (8)$$

Note that for each T, we have a different decision function and

$$E(\rho_n / T\underline{X}_n = \underline{y}) = D_T'(\underline{y}) \qquad (9)$$

so that D_T' may be estimated from another sequence $\{(T\underline{X}_n, \rho_n)\}$.

C. PROCEDURE FOR TWO CLASSES

For convenience, we write, when necessary, the matrix T as a vector $\underline{t}$ of length dd'; i.e.,

$$t_k = T_{ij}, \text{ for } k = (i-1)d + j, \quad 1 \leq i \leq d', \quad 1 \leq j \leq d$$

Let

$$J(\underline{t}) = P_e(D_T')$$
$$= \frac{1}{2} - \frac{1}{2} \int_{R_d} D_B(T\underline{x}) \operatorname{sgn} D_T'(T\underline{x}) p(\underline{x}) d\underline{x} \qquad (10)$$

We minimize $J(\underline{t})$ by a variant of the stochastic gradient algorithm (Cf. Fritz and Gyorfi[4]). If $p(\underline{x})$ is continuous, $J(\underline{t})$ is differentiable at $t \neq 0$. Denote this gradient by $\underline{U}(\underline{t})$ $(\underline{t} \neq 0)$. We have that $J(\alpha\underline{t}) = J(\underline{t})$ and $\alpha\underline{U}(\alpha\underline{t}) = \underline{U}(\underline{t})$ if $\alpha > 0$. We minimize J on the unit sphere

$$S_{dd'} = \left\{ \underline{t} / \underline{t} \in R_{dd'}, \quad \|\underline{t}\| = 1 \right\}$$

where $\|\cdot\|$ denotes the Euclidean norm. The algorithm is:-

$$\underline{b}_{n+1} = \underline{t}_n - \gamma_n \underline{w}_{n+1}$$

$$\underline{t}_{n+1} = \underline{b}_{n+1} / \|\underline{b}_{n+1}\| \qquad (11)$$

$$\underline{w}_{n+1} = \underline{Z}_{n+1} - \underline{t}_n \langle \underline{t}_n, \underline{Z}_{n+1} \rangle$$

$$\underline{Z}_{n+1} = -\rho_{n+1} \lambda_n h(\lambda_n D'_{T_n}(T_n \underline{X}_{n+1})) \nabla D'_{T_n}(T_n \underline{X}_{n+1})$$

where $\langle .,. \rangle$ is the inner product in $R_{dd'}$, $\{\gamma_n\}$ is a sequence of positive real numbers, $\underline{Z}_{n+1}$ is an estimator for $\underline{U}(\underline{t}_n)$ based on the training sequence and h is a smoothing function(taken as the univariate normal density with mean 0 and variance 1). Our approach to estimate $\underline{U}$ is to use a smoothing version of the Kiefer-Wolfowitz scheme. If the constants γ_n and λ_n satisfy

$$\gamma_n \geq 0, \lambda_n > 0 \text{ and } \lambda_n \to \infty \text{ as } n \to \infty,$$

$$\sum_{n=0}^{\infty} \gamma_n = \infty, \sum_{n=0}^{\infty} \gamma_n^2 \lambda_n^2 < \infty, \sum_{n=0}^{\infty} \gamma_n \lambda_n^{-1} < \infty$$

and $\underline{U}$ satisfies the uniform Lipshitz condition, then it can be shown that the algorithm converges to a stationary point. The algorithm may have to be started at several initial states to locate the global minimum.

D. EXTENSIONS TO c CLASSES

Extending the procedure to an arbitrary number of classes is quite straight-forward. In the training sequence of pairs $(\underline{X}_n, \rho_n)$, ρ_n now represents the class. This implies that

$$P(\rho_n = m / \underline{X}_n = \underline{x}) = p(w_m / \underline{x}) \qquad (12)$$

Analogous to (11), define

$$\underline{Z}_{n+1} = -\sum_m \sum_{i \neq m} \lambda_n h(\lambda_n D'_{mi}(\underline{Y}_{n+1})) \nabla D'_{mi}(\underline{Y}_{n+1}) \prod_{j \neq m} u(D'_{mj}(\underline{Y}_{n+1})) \qquad (13)$$

Under conditions similar to the case of c = 2, it can be shown that $\underline{Z}_{n+1}$ in (13) is an appropriate approximation to the gradient of J and the algorithm converges.

IV. NUMERICAL RESULTS

A. DATA

The algorithm developed in the preceding section was coded and run on the following three data sets:

1. 4-channel, 5-class Landsat data of the Malaprabha dam site (Site M);

2. 4-channel, 5-class Landsat data of Chandrapur district (Site C); and

3. 12-band, 8-class data given by Starks, et. al.[13] (Site D).

In the first two cases, supervised samples were available but not enough in number (only 45 for some classes). In case 3, only the mean and covariance matrices were available. Hence in each case, samples were generated assuming a multivariate Gaussian distribution.

B. IMPLEMENTATION

We note that our algorithm does not require the computation of error at any stage. However, programs were written to estimate/approximate P_e.

Each data set was split into three sets S_D, S_Z and S_e:

1. S_Z was used to compute the estimator $\underline{Z}_n$;

2. S_D was used to compute the functions D'_T or D_{ij}, which are essentially frequencies;

3. S_e was used to compute P_e.

The samples in S_D and S_e are class-segregated because we need only $p(w_i / T\underline{x})$. We describe below the steps for c=2.

C. COMPUTATION OF T

1. Choose an initial transformation such that $\|\underline{t}\| = 1$.

2. For each $\underline{x}_n \in S_Z$, do steps 3 and 4 below.

3. Evaluate $D_T'(y_n)$ and its gradient $\nabla D_T'(y_n)$.

4. Find $\underline{z}_n$ and update T_n using (11).

Since the algorithm in general converges to a local minimum, the procedure has to be restarted with different initial $\underline{t}$.

D. COMPUTATION OF $D_T'(\underline{y})$

Since $D_T'(\underline{y}) = E(\rho_n / T\underline{X}_n = \underline{y})$, $D_T'(\underline{y})$ can be obtained by counting the number of samples in a small hypercube containing $\underline{y}$; the width of the hypercube is kept as a parameter in the program.

E. COMPUTATION OF THE GRADIENT OF D_T'

The gradient of D_T' at any $\underline{x}_n \in S_Z$ is computed by finite difference approximation.

F. COMPUTATION OF ERROR

The error estimation program classifies each sample of S_e and counts the number of misclassifications. A quasi-Newton procedure was also used for error evaluation. For two classes, assuming Gaussian distribution, the exact error is easy to evaluate.

G. APPLICATIONS

Our procedure was applied to two-class case corresponding to Sites M, C and D, five classes for Sites M and C and eight classes for Site D. The exact and approximate error computation procedures were compared when d=12, d'=1 and S_e =100. Some typical results, rounded to two decimal places, are given in the following tables. Unless otherwise stated, $J(\underline{t})$ is the estimated value.

Table 1. Two-class Results for Site C
$|S_Z| = 50$; d = 4; d' = 1

n	0	50	0	50
$\underline{t}_n$	0	-0.01	-0.71	-0.49
	0	0.16	0.71	0.84
	0	0.34	0.00	0.16
	1	0.93	0.00	0.17
$J(\underline{t}_n)$ exact	0.29	0.19	0.38	0.27

Table 2. Two-class Results for Site M
$|S_Z| = 50$; d = 4; d' = 1

n	0	50	0	50
$\underline{t}_n$	1	0.97	0	0.06
	0	0.00	0	0.15
	0	0.23	0	-0.12
	0	-0.12	1	0.98
$J(\underline{t}_n)$ exact	0.06	0.00	0.00	0.00

Table 3. Two-class Results for Site D
$|S_Z| = 50$; d = 12; d' = 1

n	0	50	0	50	0	50
$\underline{t}_n$	1	0.87	0	-0.03	0.29	0.27
	0	-0.17	0	-0.05	0.29	0.31
	0	0.18	0	-0.10	0.29	0.42
	0	-0.10	0	0.05	0.29	0.34
	0	-0.20	0	-0.07	0.29	0.25
	0	-0.13	1	0.91	0.29	0.25
	0	-0.02	0	-0.04	0.29	0.30
	0	-0.09	0	-0.11	0.29	0.32
	0	-0.21	0	-0.10	0.29	0.34
	0	-0.05	0	-0.08	0.29	0.30
	0	-0.11	0	-0.26	0.29	0.03
	0	-0.22	0	-0.21	0.29	0.15
$J(\underline{t}_n)$ exact	0.27	0.03	0.26	0.03	0.07	0.01
$J(\underline{t}_n)$ Quasi-Newton	-	0.00	-	0.00	-	0.00

Table 4. Five-class Results for Site C

$|S_Z| = 100$; $|S_D| = 250$; $|S_e| = 250$; d=4

n	0	100	100	100
d'	-	1	2	3
$\underline{t}_n$	0.5	0.36	0.53	0.14
	0.5	0.65	-0.40	-0.09
	0.5	0.43	0.20	0.17
	0.5	0.52	0.25	-0.33
	0		0.53	-0.30
	0		-0.40	0.22
	0		0.20	-0.50
	0		0.25	-0.18
	0			-0.30
	0			-0.22
	0			-0.41
	0			-0.18
$J(\underline{t}_n)$	0.22	0.18	0.17	0.17

Table 5. Five-class Results for Site M

$|S_Z| = 100$; $|S_D| = 250$; $|S_e| = 250$; d=4

n	0	100	100	100
d'	-	1	2	3
$\underline{t}_n$	0.5	0.47	0.15	-0.23
	0.5	0.50	0.08	-0.06
	0.5	0.54	-0.68	0.16
	0.5	0.48	-0.09	0.65
	0		0.15	0.45
	0		0.08	-0.02
	0		-0.68	-0.01
	0		-0.09	0.20
	0			0.45
	0			-0.02
	0			-0.01
	0			0.20
$J(\underline{t}_n)$	0.20	0.19	0.16	0.11

H. DISCUSSIONS

1. When reducing the dimensions from d to d', we would expect the final error to be non-decreasing as d' decreases.

2. For a fixed collection of training samples, the error rate will be higher as more parameters, viz. $\underline{t}$, have to be determined.

Table 6. Eight-class Comparative Results

$|S_Z| = 100$; $|S_D| = 400$; $|S_e| = 400$

d = 12; d' = 1

	n=0	n=100	$\underline{t}$ (Starks, et. al.)
$\underline{t}_n$	1	0.07	-0.44
	0	-0.40	-0.39
	0	0.40	0.05
	0	0.08	-0.24
	0	0.06	-0.08
	0	0.16	-0.24
	0	0.04	-0.13
	0	0.28	0.44
	0	0.38	0.52
	0	0.31	0.04
	0	-0.51	-0.08
	0	-0.25	0.22
$J(\underline{t}_n)$	0.51	0.26	0.27

Table 7. Comparison of the Error estimating Procedure with exact values for a multi-variate normal case reduced to one dimension.

d = 12; d' = 1; S_e = 100

n	$J(\underline{t}_n)$ "EXACT"	$J(\underline{t}_n)$ APPROXIMATE
1	0.2087122	0.13
5-6	0.2181920	0.34
10-11	0.0720590	0.08
17-18	0.0126058	0.01
20-24	0.0230671	0.02
29-40	0.0132714	0.06
41-51	0.0118969	0.08
52-80	0.0036274	0.03
81-100	0.0022369	0.01

3. The requirement for training samples would be more for lower d'.

4. The algorithm converges almost surely. The value of $J(\underline{t}_n)$ need not decrease at every step.

5. Approximations to D_T' and $\nabla D_T'$ are sensitive to the size of hypercube and difference parameter respectively.

V. CONCLUSIONS

In this paper, we have constructively proposed a nonparametric feature transformation for dimensionality reduction, important in resource evaluation.

Although the literature describes many transformations, both linear and nonlinear, only a few of these are determined using the error rate criterion. Of these, Odell's[6] transformation is restricted to Gaussian classes and that Lissack and Fu[7] deals with subset selection for two classes. The methodology of de Figueiredo[12] requires knowledge of the underlying multivariate distribution function and, in general, difficult to implement and the methodology has been specialized by Starks, et. al.[13] to Gaussian classes for reducing to one dimension.

The salient features of our procedure are the following.

1. It is not restricted to two classes.

2. It is not restricted to any specific distribution.

3. It is non-parametric. The algorithm is convergent, almost surely. The convergence characteristics have been studied when the supervised samples available are of the order of fifty to a few hundred.

4. It can be extended to minimize the error rate with respect to decision functions other than the Bayesian decision function; i.e, the functions $\{D_{ij}\}$ may be changed so long as they satisfy suitable conditions and can be obtained from samples.

5. It can be used to reduce the feature space to desired dimensions.

6. The procedure can be used in the design and operational phases, updating the transformation as and when a labelled sample is obtained.

7. The value of the error rate need not be computed at each step for the determination of the transformation.

VI. ACKNOWLEDGEMENTS

This work was supported in part by the Centre for Studies in Resources Engineering of the Indian Institute of Technology, Bombay. The authors wish to thank Professor R.K.Katti, Head, CSRE, for this support and Dr G.Venkatachalam and Mr P.R.Saraph for their input. Thanks are also due to Dr S.P.Mudur of the Tata Institute of Fundamental Research, Bombay and Dr G.Nagaraja of the IIT, Bombay for their constructive criticisms on an earlier draft of this paper.

VII. REFERENCES

1. Y.S.Chien and K.S.Fu, "On the generalized Karhunen - Loeve expansion", IEEE Trans. Inform. Theory, vol. IT-13, pp. 518-520, 1967.

2. H.P.Decell and S.U.Mayekar, "Feature combinations and the divergence criterion", Comput. Math. Applics, vol. 3, pp.71-76 1977.

3. R.A.Fisher, "The use of multiple measurements in taxonomic problems", Ann. Eugenics, vol. 7, pp. 179-188, 1936.

4. D.H.Foley and J.W.Sammon, "Optimal set of discriminant vectors", IEEE Trans. Comput., vol. C-24 pp. 281-289, 1975.

5. K.Fukunaga and W.L.G.Koontz, "Application of the K-L expansion to feature selection and ordering", IEEE Trans. Comput., vol. C-19, pp. 311-318, 1970.

6. P.L.Odell, K.A.Coberly and H.P.Decell, "Linear dimension reduction and Bayes' classification", University of Houston, Report No. 6, 1978.

7. T.Lissack and K.S.Fu, "Parametric feature extraction through error minimization applied to medical diagnosis", IEEE Trans. Syst., Man, Cybern., vol. SMC-6, pp. 605-611, 1976.

8. J.W.Sammon, "A nonlinear mapping for data structure analysis", IEEE Trans. Comput., vol. C-18, pp. 401-409, 1969.

9. I.White, "Comments on a nonlinear mapping for data structure analysis", IEEE Trans. Comput., vol. C-21, pp. 220-221, 1972.

10. C.L.Chang and R.C.J.Lee, "A heuristic relaxation method for nonlinear mapping in cluster analysis", IEEE Trans. Syst., Man, Cybern., vol. SMC-3, pp. 197-200, 1973.

11. T.W.Calvert and T.Y.Young, "Randomly generated nonlinear transformation for pattern recognition", IEEE Trans. Syst. Sci., Cybern., vol. SSC-5, pp. 266-273, 1969.

12. R.J.P. de Figueiredo, "Design of optimal feature extractors by mathematical programming technique", in Pattern Recognition and Artificial Intelligence, C.H.Chen (ed.), Acad. Press, 1976.

13. S.A.Starks, R.J.P. de Figueiredo and D.L.Van Rooy, "An algorithm for single linear feature extraction from several Gaussian classes", Intl. J. Comp. Info. Sci., vol. 6, pp. 41-54, 1977.

14. J.Fritz and L.Gyorfi, "On the minimization of classification error probability in statistical pattern recognition", Problems Control Inform. Theory, vol. 5, pp. 371-382, 1975.

VIII. AUTHORS' BIOGRAPHICAL DATA

Subramani Arunkumar received the B.Tech.(Hons.) degree from the Indian Institute of Technology, Bombay, M.A.Sc. from the University of Toronto and the M.S. and Ph.D degrees from the University of California at Berkeley. He was a professor at the University of California, Los Angeles prior to joining IIT, Bombay where he is Professor of Computer Science and in-charge of the inter-disciplinary program in Industrial Management. He has been a Visiting Professor at the University of California, Berkeley and the Bell Laboratories, New Jersey. The current areas of his research include layout and testing in VLSI, intelligent systems and applications, computer networks and systems modelling and optimization.

Rajendra Supnekar received the M.Sc. and M.Tech. degrees from the Indian Institute of Technology, Bombay. During the period of his graduate work for the M.Tech. degree, he was a research scholar of the Centre for Studies in Resources Engineering, IIT, Bombay. He is presently involved in software development with Tata Burroughs, Ltd., Bombay.