

# Appendices

## A.1 Elements of Matrix Algebra

**Definitions of vectors and matrices.**

$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$  is an  $n$ -variate vector column, where  $x_1, x_2, \dots, x_n$  are components of  $\mathbf{X}$ .

$\mathbf{Z} = (x_1, x_2, \dots, x_n)$  is an  $n$ -variate vector row,  $\mathbf{Z} = \mathbf{X}^T$ ;  $T$  denotes a transpose operation; and  $n$  is called a dimensionality of the vectors  $\mathbf{X}$  and  $\mathbf{Z}$ .

$\mathbf{A} = ((a_{ij}))$ , ( $i, j = 1, 2, \dots, n$ ) is an  $n \times n$  quadratic matrix,  $a_{ij}$  are the elements of matrix  $\mathbf{A}$ . If  $a_{ij} = a_{ji}$ , the matrix  $\mathbf{A}$  is symmetric.

**Multiplication of matrices.** Let  $\mathbf{B} = ((b_{ij}))$ , ( $i, j = 1, 2, \dots, n$ ) be another  $n \times n$  quadratic matrix. Then  $\mathbf{AB} = \mathbf{C} = ((c_{ij}))$ , is also an  $n \times n$  quadratic matrix with

elements  $c_{ij} = \sum_{s=1}^n a_{is} b_{sj}$ . If  $\mathbf{Y} = (y_1, y_2, \dots, y_n)^T$  is a  $n$ -variate vector column,

$$\mathbf{Y}^T \mathbf{A} \mathbf{X} = \sum_{j=1}^n \sum_{i=1}^n a_{ij} y_i x_j.$$

**A hyperplane** is  $\mathbf{V}^T \mathbf{X} + v_0 = 0$ , where  $\mathbf{V}$  is a vector column.

**A distance** between a vector and the hyperplane  $\mathbf{V}^T \mathbf{X} + v_0 = 0$  is  $H = \frac{|\mathbf{V}^T \mathbf{X} + v_0|}{\sqrt{\mathbf{V}^T \mathbf{V}}}$ .

**Orthogonal matrices.** If  $\mathbf{Y}^T \mathbf{X} = 0$ , the vectors  $\mathbf{Y}$  and  $\mathbf{X}$  are said to be orthogonal. If, in addition,  $\mathbf{Y}^T \mathbf{Y} = 1$  and  $\mathbf{X}^T \mathbf{X} = 1$ , the vectors  $\mathbf{Y}$  and  $\mathbf{X}$  are said to be orthonormal. A quadratic matrix  $\mathbf{T}$  is said to be orthogonal if  $\mathbf{T} \mathbf{T}^T = \mathbf{D} = ((d_{ij}))$

(diagonal matrix). In the diagonal matrix  $\mathbf{D}$ ,  $d_{ij} = 0$  if  $j \neq i$ . If all  $d_{ii} > 0$ , the matrix  $\mathbf{D}$  is positively defined. Its determinant  $\det(\mathbf{D}) = \prod_{i=1}^n d_{ii} > 0$ . If  $\mathbf{T}\mathbf{T}^T = \mathbf{I}_n$  (the identity matrix), the matrix  $\mathbf{T}$  is said to be orthonormal. The identity matrix has ones on its diagonal, and zeros outside the diagonal. For orthonormal matrix  $\mathbf{T}$  one can write  $\mathbf{T}\mathbf{T}^T = \mathbf{T}^T\mathbf{T} = \mathbf{I}_n$ .

**Singular value decomposition.** The symmetric matrix  $\mathbf{A}$  can be decomposed into a product:  $\mathbf{A} = \mathbf{T}\mathbf{D}\mathbf{T}^T$ , where  $\mathbf{T}$  is an orthonormal  $n \times n$  matrix such that

$$\mathbf{T}^T\mathbf{A}\mathbf{T} = \mathbf{D} = \begin{bmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & d_n \end{bmatrix}. \quad (\text{A1.1})$$

Matrix  $\mathbf{T}$  is called an eigenvectors matrix and diagonal elements of matrix  $\mathbf{D}$  are called eigenvalues of matrix  $\mathbf{A}$ . The matrix  $\mathbf{A}$  is positively defined if all  $d_i > 0$ .

**Inverse of the symmetric quadratic matrix.** Let  $\mathbf{B}$  be a quadratic positively defined matrix, such that  $\mathbf{B}\mathbf{A} = \mathbf{A}\mathbf{B} = \mathbf{I}$ . Then  $\mathbf{B}$  is called an inverse of  $\mathbf{A}$  and denoted by  $\mathbf{B} = \mathbf{A}^{-1}$ . The representation (A1.1) shows that

$$\mathbf{A}^{-1} = (\mathbf{T}\mathbf{D}\mathbf{T}^T)^{-1} = (\mathbf{T}^T)^{-1}\mathbf{D}^{-1}\mathbf{T}^{-1} = \mathbf{T}\mathbf{D}^{-1}\mathbf{T}^T. \quad (\text{A1.2})$$

**Pseudo-inversion.** If  $n-r$  eigenvalues of matrix  $\mathbf{A}$  are equal to zero, the matrix  $\mathbf{A}$  is not positively defined. Its determinant is equal to zero. Let us write the diagonal matrix  $\mathbf{D}$  in a block layout  $\mathbf{D} = \begin{bmatrix} \mathbf{d} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$  in (A1.1), where  $\mathbf{d}$  is an  $r \times r$  diagonal matrix composed from non-zero values of  $\mathbf{D}$ . Thus, we can rewrite Equation (A1.1) in a block layout

$$\mathbf{T}^T\mathbf{A}\mathbf{T} = [\mathbf{T}_1 \ \mathbf{T}_2]^T \mathbf{A} [\mathbf{T}_1 \ \mathbf{T}_2] = \begin{bmatrix} \mathbf{d} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad (\text{A1.3})$$

where we have split the orthogonal matrix  $\mathbf{T} = [\mathbf{T}_1 \ \mathbf{T}_2]$  into  $n \times r$  matrix  $\mathbf{T}_1$  and  $n \times (n-r)$  matrix  $\mathbf{T}_2$ . Then the pseudo-inverse of matrix  $\mathbf{A}$

$$\mathbf{A}^+ = \mathbf{T} \begin{bmatrix} \mathbf{d}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{T}^T = [\mathbf{T}_1 \ \mathbf{T}_2] \begin{bmatrix} \mathbf{d}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} [\mathbf{T}_1 \ \mathbf{T}_2]^T = \mathbf{T}_1 \mathbf{d}^{-1} \mathbf{T}_1^T. \quad (\text{A1.4})$$

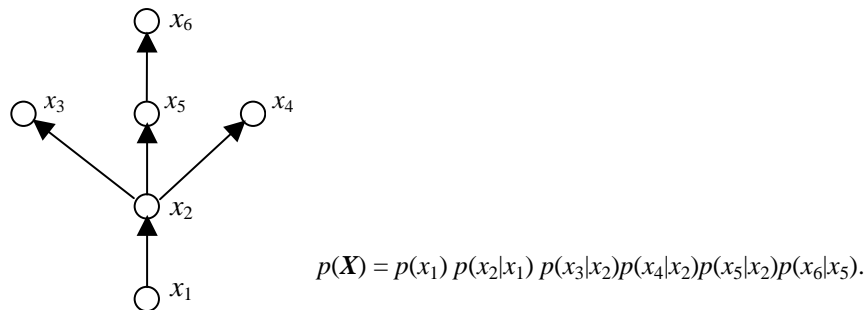
## A.2 The First Order Tree Type Dependence Model

The probability density function of the random vector  $\mathbf{X} = (x_1, x_2, \dots, x_n)$  having the first order tree type dependence between the variables can be written in the following form:

$$p(x_1, x_2, \dots, x_n) = \prod_{j=1}^n p(x_j | x_{m_j}) \quad (0 \leq m_j \leq n) \tag{A2.1}$$

where a sequence  $m_2, \dots, m_n$  constitutes a *graph of connections*,  $\mathbf{m}$  (an unknown permutation of the integers  $1, 2, \dots, n$ ), and  $p(x_i | x_0)$ , by definition, is equal to  $p(x_i)$ . In a general case, the covariance matrix has  $n \times n$  non-zero elements. An inverse of this matrix  $\Sigma^{-1}$  that has to be used in the classifier design, however, has only  $2n-1$  different non-zero elements. It is a result of the assumption that each component of the vector  $\mathbf{X}$  depends directly on only one another component.

To depict the dependence relations graphically, the variable is represented by a point on the plane, and if  $x_i$  and  $x_j$  are two variables such that  $j = m_i$ , they will be joined by an arrow pointing from  $x_j$  to  $x_i$ . Figure A2.1 shows an example of a dependence tree with graph  $\mathbf{m} = (m_2, m_3, m_4, m_5, m_6) = (1, 2, 2, 2, 5)$ .



**Fig. A2.1.** An example of the first order dependence tree.

We present an analytical expression of the density for the multivariate Gaussian case. The conditional density function for two Gaussian distributed variables is

$$p(x_i | x_j) = N(x_i, m_i + (x_j - m_j) \sigma_{ij}^{-1} \sigma_{ii} - \sigma_{ij}^{-1} \sigma_{ij}^2) \tag{A2.2}$$

where  $\sigma_{ij}$  is an element of the covariance matrix  $\Sigma$ .

Using (A2.2) in (A2.1) yields a simple analytical expression for the joint probability density function

$$p(x_1, x_2, \dots, x_n) = \prod_{j=1}^n N(x_j, m_j + (x_{m_j} - m_{m_j}) \sigma_{m_j m_j}^{-1} \sigma_{j m_j}, \sigma_{jj} - \sigma_{m_j m_j}^{-1} \sigma_{j m_j}^2) =$$

$$\frac{1}{(2\pi)^{n/2}} \prod_{j=1}^n (\sigma_{jj} - \sigma_{m_j m_j}^{-1} \sigma_{j m_j}^2)^{-1/2}$$

$$\exp \left( -\frac{1}{2} \sum_{j=1}^n \frac{[(x_j - m_j) - (x_{m_j} - m_{m_j}) \sigma_{m_j m_j}^{-1} \sigma_{j m_j}]^2}{\sigma_{jj} - \sigma_{m_j m_j}^{-1} \sigma_{j m_j}^2} \right) \quad (\text{A2.3})$$

We require that the variables  $x_1, x_2, \dots, x_n$  be ranked in such a way that  $m_j < j$ ,  $j = 2, 3, \dots, n$ . Then density function (A2.3) may be written in the following form

$$p(x_1, x_2, \dots, x_n) = N(\mathbf{X}, \mathbf{M}, \Sigma),$$

where

$$\Sigma^{-1} = (\mathbf{C}^T \mathbf{C}), \quad \mathbf{C} = ((c_{ij})), \quad (\text{A2.4})$$

$$c_{ij} = \begin{cases} \frac{1}{\sqrt{\sigma_{ii}(1-r_{im_i}^2)}} & \text{if } j=i \\ \frac{-r_{im_i}}{\sqrt{\sigma_{m_i m_i}(1-r_{im_i}^2)}} & \text{if } j=m_i \\ 0 & \text{if } j=i \text{ and } j \neq m_i \end{cases} \quad (\text{A2.5})$$

$$r_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{jj} \sigma_{ii}}}.$$

**Example.** Let  $\sigma_{21} = 0.7$ ;  $\sigma_{32} = -0.3$ ;  $\sigma_{42} = 0.4$ ;  $\sigma_{52} = 0.2$ ;  $\sigma_{65} = -0.6$ , and all variances  $\sigma_{11} = \sigma_{22} = \sigma_{33} = \sigma_{44} = \sigma_{55} = \sigma_{66} = 1$ . Then

$$\mathbf{C} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ -0.9802 & 1.4003 & 0 & 0 & 0 & 0 \\ 0 & 0.3145 & 1.0483 & 0 & 0 & 0 \\ 0 & -0.4364 & 0 & 1.0911 & 0 & 0 \\ 0 & -0.2041 & 0 & 0 & 1.0206 & 0 \\ 0 & 0 & 0 & 0 & 0.7500 & 1.2500 \end{bmatrix},$$

$$\Sigma^{-1} = \begin{bmatrix} 1.9608 & -1.3725 & 0 & 0 & 0 & 0 \\ -1.3725 & 2.2918 & 0.3297 & -0.4762 & -0.2083 & 0 \\ 0 & 0.3297 & 1.0989 & 0 & 0 & 0 \\ 0 & -0.4762 & 0 & 1.1905 & 0 & 0 \\ 0 & -0.2083 & 0 & 0 & 1.6042 & 0.9375 \\ 0 & 0 & 0 & 0 & 0.9375 & 1.5625 \end{bmatrix}, \text{ and}$$

$$\Sigma = \begin{bmatrix} 1 & 0.700 & -0.210 & 0.280 & 0.140 & -0.084 \\ 0.700 & 1 & -0.300 & 0.400 & 0.200 & -0.120 \\ -0.210 & -0.300 & 1 & -0.120 & -0.06 & 0.0360 \\ 0.280 & 0.400 & -0.120 & 1 & 0.080 & -0.048 \\ 0.140 & 0.200 & -0.06 & 0.080 & 1 & -0.600 \\ -0.084 & -0.120 & 0.0360 & -0.048 & -0.600 & 1 \end{bmatrix}.$$

When the graph  $\mathbf{m} = (m_2, \dots, m_n)$  is known, we estimate the inverse covariance matrix from the sample covariance matrix as  $\hat{\Sigma}^{-1} = \hat{\mathbf{C}}^T \hat{\mathbf{C}}$ , where  $\hat{\mathbf{C}}$  is determined by Equation (A2.5) with the elements  $\sigma_{ij}$  substituted by their corresponding sample estimates.

To estimate the graph  $\mathbf{m} = (m_2, m_3, \dots, m_n)$ , it is suggested that one uses a stepwise algorithm developed by Kruskal (1956) for the construction of trees of maximum total branch weight. Let  $\{|\hat{r}_{12}|, |\hat{r}_{13}|, |\hat{r}_{14}|, \dots, |\hat{r}_{n-1 n}|\}$  be the absolute values of the sample correlation coefficients between the variables. Then, the first step is to select a branch with the greatest weight  $|\hat{r}_{st}|$ , while the  $i$ -th step ( $2 \leq i \leq n-1$ ) is to choose a branch with the greatest weight  $|\hat{r}_{vu}|$  that is different from all the branches selected during the previous steps and does not form a cycle with the previously selected branches. If the multivariate normal density may be represented by the branch  $\mathbf{m}$  model, then the sample estimate  $\hat{\mathbf{m}}$

of the graph asymptotically (as the sample size  $N_1, N_2 \rightarrow \infty$  and the dimensionality  $n \rightarrow \infty$ ) converges to the true graph  $\mathbf{m}$ .

When the above model is used to design the linear discriminant function (for the two category case when it is assumed that  $\Sigma_2 = \Sigma_1$ ) then one has to estimate  $4n - 1$  parameters:  $2n$  components of the mean vectors,  $n$  variances  $\sigma_{ii}$  and  $n - 1$  covariances, as well as  $n - 1$  numbers that compose the graph  $\mathbf{m}$ .

### A.3 Temporal Dependence Models

Let the components  $x_1, x_2, \dots, x_{n-1}, x_n$  of the multivariate vector be measurements differing in time or in space, and assume they are a stationary random process. Then the covariance matrix has a following structure

$$\Sigma = \begin{bmatrix} \delta_1 & \delta_2 & \delta_3 & \dots & \delta_{n-1} & \delta_n \\ \delta_2 & \delta_1 & \delta_2 & \dots & \delta_{n-2} & \delta_{n-1} \\ \delta_3 & \delta_2 & \delta_1 & \dots & \delta_{n-3} & \delta_{n-2} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \delta_{n-1} & \delta_{n-2} & \delta_{n-3} & \dots & \delta_1 & \delta_2 \\ \delta_n & \delta_{n-1} & \delta_{n-2} & \dots & \delta_2 & \delta_1 \end{bmatrix}. \quad (\text{A3.1})$$

We see only  $n$  parameters  $\delta_1, \delta_2, \delta_3, \dots, \delta_n$  that describe the dependence between the variables and have to be estimated from the training-set. Let  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N$  be  $N$  training-set vectors,  $\mathbf{X}_j = (x_{j1}, x_{j2}, \dots, x_{jn})^T$ . Then, the sample estimate

$$\hat{\delta}_l = \frac{1}{N(n-l)} \sum_{j=1}^N \sum_{t=1}^{n-l} (x_{jt} - \bar{x}_t)(x_{j,t+l} - \bar{x}_{t+l}).$$

A number of special models, such as autoregression, moving average, ARMA, and circular, reduce the number of parameters even more.

In the *autoregression model*, we have  $q + 1$  independent parameters, and here the dependence among variables is determined by the equation

$$x_t + a_1 x_{t-1} + \dots + a_r x_{t-r} = v_t, \quad (\text{A3.2})$$

where random variables  $v_t, v_{t-1}, \dots$  are supposed to be mutually independent and identically distributed  $N(0, 1)$ . In this model, the last  $p-r$  variables  $x_r, x_{r+1}, \dots, x_{n-1}$  are linearly dependent on the previous ones. The inverse covariance matrix has a simple form, and can be calculated analytically (Kligiene, 1977)

$$\Sigma^{-1} = \begin{pmatrix} \kappa_{11} & \dots & \kappa_{1r} & \kappa_r & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \kappa_{r1} & \dots & \kappa_{rr} & \kappa_1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ \kappa_r & \dots & \kappa_1 & \kappa_0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & \kappa_0 & \dots & 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & \kappa_0 & 0 & 0 & \dots & 0 \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 & \kappa_0 & \kappa_1 & \dots & \kappa_r \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 & \kappa_1 & \kappa_{rr} & \dots & \kappa_{r1} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 & \kappa_r & \kappa_{1r} & \dots & \kappa_{11} \end{pmatrix}, \tag{A3.3}$$

where  $\kappa_l = \sum_{k=0}^{r-l} a_k a_{k+l}$ ,  $\kappa_{st} = \sum_{k=0}^{\min(s,t)-1} a_k a_{k+|s-t|}$ ,  $s, t = 1, 2, \dots, r$ .

In the *moving average model*, we have  $q + 1$  independent parameters, and here the dependence among variables is determined by the equation

$$x_t = \mu_t + b_0 v_t + b_1 v_{t-1} + \dots + b_q v_{t-q}, \tag{A3.4}$$

where the variables  $v_t, v_{t-1}, \dots$  are supposed to be mutually independent and identically distributed  $N(0, 1)$ . In this model, we have

$$\delta_{q+1} = \delta_{q+2} = \delta_{q+3} = \dots = \delta_n = 0; \quad q < n.$$

The *circular* covariance matrix has  $n/2$  independent parameters and has form

$$\Sigma = \sigma_0 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \dots & \rho_2 & \rho_1 \\ \rho_1 & 1 & \rho_1 & \rho_2 & \dots & \rho_3 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 & \dots & \rho_4 & \rho_3 \\ \rho_3 & \rho_2 & \rho_1 & 1 & \dots & \rho_5 & \rho_4 \\ \rho_4 & \rho_3 & \rho_2 & \rho_1 & \dots & \rho_6 & \rho_5 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \rho_1 & \rho_2 & \rho_3 & \rho_4 & \dots & \rho_1 & 1 \end{pmatrix}. \tag{A3.5}$$

This matrix can be transformed into a canonical form by means of the following orthonormal matrix  $\mathbf{L} = ((l_{mn}))$  with elements

$$l_{ms} = \sqrt{n} \left( \cos \frac{2\pi}{n} (m-1)(s-1) + \sin \frac{2\pi}{n} (m-1)(s-1) \right), \quad (\text{A3.6})$$

$$\text{such that } \mathbf{L}^T \mathbf{\Sigma} \mathbf{L} = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_n \end{pmatrix}.$$

Note, the transformation matrix  $\mathbf{L}$  does not depend on  $\mathbf{\Sigma}$ .

The model of the *additive noise* is typical to physical measurements of the same origin where all variables are influenced by the same systematic Gaussian error  $N(0, \sigma_1^2)$ . The original and the inverse covariance matrices are determined only by two parameters:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 + \sigma_2^2 & \sigma_1^2 & \sigma_1^2 & \dots & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 + \sigma_2^2 & \sigma_1^2 & \dots & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 + \sigma_2^2 & \dots & \sigma_1^2 \\ \dots & \dots & \dots & \dots & \dots \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 & \dots & \sigma_1^2 + \sigma_2^2 \end{pmatrix},$$

$$\mathbf{\Sigma}^{-1} = \begin{pmatrix} a & b & b & \dots & b \\ b & a & b & \dots & b \\ b & b & a & \dots & b \\ \dots & \dots & \dots & \dots & \dots \\ b & b & b & \dots & a \end{pmatrix}, \quad (\text{A3.7})$$

where

$$a = ((n-1)\sigma_1^2 + \sigma_2^2) / (\sigma_2^2(n\sigma_1^2 + \sigma_2^2)),$$

$$b = -\sigma_1^2 / (\sigma_2^2(n\sigma_2^2 + \sigma_2^2)).$$

## A.4 Pikelis Algorithm for Evaluating Means and Variances of the True, Apparent and Ideal Errors in Model Selection

Let  $\hat{\epsilon}_{v_1}, \hat{\epsilon}_{v_2}, \dots, \hat{\epsilon}_{v_M}$  be  $M$  “inaccurate” (e.g. validation set) performance estimates, and  $\hat{\epsilon}_{t_1}, \hat{\epsilon}_{t_2}, \dots, \hat{\epsilon}_{t_M}$  be  $m$  “accurate” (e.g. large test set) performance estimates, corresponding to  $M$  variants (models). We investigate all possible collections composed of  $m$  models selected from a pool of  $M$  models ( $m < M$ ). We need to estimate the mean values  $E$  and variances  $V$  of the true, apparent, and ideal errors when we select the best variant from an arbitrary collection composed of  $m$  randomly chosen vectors  $(\hat{\epsilon}_{v_j}, \hat{\epsilon}_{t_j})$ .

1. Input data:  $M$  – a total number of variants considered in the experiment,  
 $m$  – a number of variants in a collection under consideration  
 $M$  – two-variate vectors  $(\hat{\epsilon}_{v_j}, \hat{\epsilon}_{t_j}), j = 1, 2, \dots, M$ . Note,  $m \ll M$ .

2. Rank  $M$  values  $\hat{\epsilon}_{v_1}, \hat{\epsilon}_{v_2}, \dots, \hat{\epsilon}_{v_M}$  in increasing order and find order numbers  $i_1, i_2, \dots, i_M$  of the ranked data.

3. Calculate a number  $v_j$  of cases when the value  $\hat{\epsilon}_{v_j}$  was the smallest one in each of  $\mathcal{J} = C_M^m = \frac{M!}{(M-m)!m!}$  possible collections composed of  $r$  vectors from  $m$  ones:  $v_1 = m/M$ . The following values of  $v_2, v_3, \dots$  are found using the recursive formula:

$$v_{j+1} = v_j * (M-r-j+1) / (M-j), \quad j = 1, 2, \dots, M-m+1.$$

4. Calculate the means and variances:

$$E\epsilon_{\text{apparent}} = \sum_{j=1}^{M-m+1} v_j \hat{\epsilon}_{v_j}, \quad V\epsilon_{\text{apparent}} = \sum_{j=1}^{M-m+1} v_j (\hat{\epsilon}_{v_j} - E\epsilon_{\text{apparent}})^2,$$

$$E\epsilon_{\text{true}} = \sum_{j=1}^{M-m+1} v_j \hat{\epsilon}_{v_j}, \quad V\epsilon_{\text{true}} = \sum_{j=1}^{M-m+1} v_j (\hat{\epsilon}_{v_j} - E\epsilon_{\text{true}})^2.$$

The mean and the variance of the ideal error  $E\epsilon_{\text{ideal}}$  can be found in an analogous way. We recommend applying this algorithm for a set of values of  $m$  ( $m \ll M$ ).

## A.5 Matlab Codes (the Non-Linear SLP Training, the First Order Tree Dependence Model, and Data Whitening Transformation)

```

% A main program to test for the data whitening by the
% first order tree dependence model and subsequent
% training the nonlinear SLP
% This program generates two p-variate Gaussian classes
% with covariance matrix C having a linear tree
% dependence model structure
% The SLP is trained twice: in original and
% transformed feature spaces.

% AUTHOR: Sarunas Raudys <raudys@das.mii.lt>

% Example:    dimensionality:  p=30;
% training & validation sizes  nm=30;nv=500;
% N of iterations:             iter=500;
% regularization:              lambda=0.001;
% correlation:                 ro=0.5;
% increase in learning step:   gama=1.01 (Section 4.6.6)
CI=eye(p);CI1=ro*eye(p+1);
CI=CI+CI1(2:p+1,1:p)+CI1(1:p,2:p+1);C=inv(CI);
[u,d]=eig(C);G=sqrt(real(d))*real(u);

A= randn(nm,p)*G; % training-data
B=randn(nm,p)*G+ones(nm,1)*[ones(1,10),-ones(1,p-10)];
Av=randn(nv,p)*G; % validation-data
Bv=randn(nv,p)*G+ones(nv,1)*[ones(1,10),-ones(1,p-10)];

% conditions E1, E2 (see Section 4.1.2.1):
M=mean([A;B]);A=A-ones(nm,1)*M;B=B-ones(nm,1)*M;
Av=Av-ones(nv,1)*M;Bv=Bv-ones(nv,1)*M;
CM=0.5*(cov(A)+cov(B));Wstart=zeros(1,p+1);

% training SLP in original feature space:
[W,etest]=slp(A,B,iter,0.1,0.00,Wstart,Av,Bv,gama);

[covt,covtI]=gentree(CM);% search for tree structure
[u,d,v]=svd(covt+lambda);% data whitening
T=u*inv(sqrtm(d));a=A*T;b=B*T;av=Av*T;bv=Bv*T;
% training SLP in transformed (whitened) feature space:
[W,et]=slp(a,b,iter,0.1,0.00,Wstart,av,bv,gama);
figure(1);plot([1:iter],etest,'r-',[1:iter],et,'g-');

%      Try to PRINT:
%disp(covtI(1:7,1:7))a part of inverse of structurised CM
% disp ([min(etest),min(et)]) % minima of
% generalisation error without and with
% whitening data transformation
+++++

```

**Single layer perceptron**

```

% Non-linear single-layer perceptron for 2 classes
% Author S.Raudys <raudys@das.mii.lt>
% Inputs:
% A, B - training sets from two classes
% rows contains observations, columns - features
% n - number of training iterations
% step - learning step
% target - target value for class A (sigmoid transfer
% function)
% Wstart - initial weight vector
% Aval, Bval - validation set from classes A, B;
% rows contains observations, columns - features
% gama - coefficient to change the learning speed
% after each training iteration, step=step * gama
% Outputs: W - weight after the last iteration
% etest - test error after each iteration

function
[W,etest]=slp(A,B,iter,step,target,Wstart,Av,Bv,gama)

[ma k] = size(A);[mb k] = size(B);
[mav k] = size(Av);[mbv k] = size(Bv);

W=Wstart;
stepz=step/(ma+mb);
ta = (1-target) * ones(ma,1);tb = target * ones(mb,1);
oa = ones(ma,1);ob = ones(mb,1);
oav = ones(mav,1);obv = ones(mbv,1);
A=[A,oa];B=[B,ob];Av=[Av,oav];Bv=[Bv,obv];

    for i=1:iter    da = A * W';
db = B * W';
e = (sum(da<0) + sum(db>=0))/(ma+mb);
fa = oa./(oa+exp(-da));
fb = ob./(ob+exp(-db));
za = ((ta-fa).* (fa - fa.*fa))'*A;
zb = ((tb-fb).* (fb - fb.*fb))'*B;
W = W + stepz * (za + zb);
etest(i)=
size([find(W*Av'<0),find(W*Bv'>0)],2)/(mav+mbv);
stepz=stepz*gama;
    end
return

```

```

+++++

% Structurisation of the covariance matrix by
% the first-order tree-type dependence model
% A main program
% Author: Ausra Saudargiene, ausrsaud@takas.lt
% Department of Data Analysis, Institute of Mathematics
% and Informatics, Akademijos 4, 2600 Vilnius

% Input: C-covariance matrix
% Outputs: covgen-structured CM, covgenI-inverse of CM

function [covtree,covtreeI]=gentree(C);

[M,S,SS,num]=treev(C);%estimation of the graph,
% weights of the branches, and the new order of
% features

[covtree,covtreeI]=treecov(M,S,SS,num);
%calculation of the tree-type covariance matrix
return

+++++

% First-order tree-type dependence model
% Estimation of the graph, weights of the branches,
% and the new order of the features
% Author: A.Saudargiene, ausrsaud@takas.lt

% Input: C-covariance matrix
% Output:
% M - graph of the tree
% S - weights of the branches (covariances)
% SS - variances, num - initial order of the features

function [M,S,SS,num]=treev(C);
[p,ms]=size(C);
%Regularization, if covariance matrix is singular
alfa=0.01;%regularization constant
if det(C)<1e-10, C=C+alfa*eye(p,p); end

%Correlation coefficients
for j=1:p
for i=1:p
A0(i,j)=C(i,j)/sqrt(C(i,i)*C(j,j));
end
end

A=triu(A0);
Mn(1)=0; M(1)=0;S(1)=0;
k=1;

```

```

M_all=zeros((p*p-p)/2,1);

%Estimating graph of the tree
for ind=2:p % number of branches
    max_A=0;
%finding max value for i-th branch
    for i=1:p
        for j=i+1:p
            ski=0;skj=0;sk=0;
% current A(i,j): checking for common points with the
% previously selected branches
% finding max value
            if abs(A(i,j))>abs(max_A)
                for l=1:k
                    if i==M_all(l) ski=1; end
                    if j==M_all(l) skj=1; end
                end
                sk=ski+skj;
                if ind==2, sk=1; end
                if sk==1
                    max_A=A(i,j);ki=i;kj=j;
                    if ski==1
                        Mn(ind)=ki; num(ind)=kj;
                    else
                        Mn(ind)=kj; num(ind)=ki;
                    end
                end
            end
        end
    end
end

S(ind)=C(ki,kj);
A(ki,kj)=0;
M_all(k)=ki;k=k+1;M_all(k)=kj;k=k+1;

%Finding initial point
if ind==2
    if M_all(1)<M_all(2)
        Mn(2)=M_all(1);
num(1)=M_all(1);num(2)=M_all(2);
    else
        Mn(2)=M_all(2);
num(1)=M_all(2);num(2)=M_all(1);
    end
end
end

% Changing the order of the features
for i=1:p
    for j=1:p
        if Mn(i)==num(j) M(i)=j; end
    end
end

```

```

        C1(i,j)=C(num(i),num(j));
    end
end
SS=diag(C1)';

return
+++++
% First-order tree-type dependence model
% Calculation of the covariance matrix (CM)
% Author: A.Saudargiene, ausrsaud@takas.lt
% Inputs: are outputs of treev.m:
% Outputs: covgen - CM,covgenI - inverse of CM
function [covgen,covgenI]=treecov(M,S,SS,num)

[s1,s2]=size(M);p=s2;
for i=2:p
    r(i)=S(i)/sqrt(SS(i)*SS(M(i)));
end
for i=1:p
    for j=1:p
        C(i,j)=0;
        if j==i
            C(i,j)=1/sqrt(SS(i)*(1-r(i)^2));
        end
        if j==M(i)
            C(i,j)=-r(i)/sqrt(SS(M(i))*(1-r(i)^2));
        end
    end
end
end
KI=C'*C;d=diag(C);
K=inv(KI);

%Initial order of the features
for i=1:p
    k=find(num==i);
    numret(i)=k;
end

for i=1:p
    for j=1:p
        covgen(i,j)=K(numret(i),numret(j));
        covgenI(i,j)=KI(numret(i),numret(j));
    end
end
end
return

```

## References

- Abramson N, Braverman D (1962) Learning to recognize patterns in a random environment. *IRE Transactions Information Theory* IT-8(5):58–63.
- Abend K, Harley TJ (1969) Comments “On mean accuracy of statistical pattern recognizers”. *IEEE Transactions on Information Theory* IT-15:420–421.
- Aivazian S, Buchstaber V, Yeniukov I, Meshalkin L (1989) *Applied Statistics: Classification and Reduction of Dimensionality*. Reference edition. Finansy i statistika, Moscow (in Russian).
- Allais DC (1966) The problem of too many measurements in pattern recognition. *IEEE International Convention Record* 7:124–30.
- Amari S (1967) A theory of adaptive pattern classifiers. *IEEE Transactions Electronic Computers* EC-16:299–307.
- Amari S (1987) Differential-geometrical methods in statistics. *Lecture notes in statistics*, 28. Springer, New York.
- Amari S (1993) A universal theorem on learning curves. *Neural Networks* 6:161–66.
- Amari S, Fujita N, Shinomoto S (1992) Four types of learning curves. *Neural Computation* 4:605–18.
- Amari S, Murata N (1993) Statistical theory of learning curves under entropy loss criterion. *Neural Computation* 5:140–53.
- An G (1996) The effects of adding noise during back propagation training on generalization performance. *Neural Computation* 8:643–74.
- Anderson TW (1951) Classification by multivariate analysis. *Psychometrika* 16:31–50.
- Anderson TW (1958) *An Introduction to Multivariate Statistical Analysis*. John Wiley, New York.

- Anderson TW (1971) *The Statistical Analysis of Time Series*. John Wiley, New York.
- Anderson TW, Bahadur RR (1962) Classification into two multivariate normal distributions with different covariance matrices. *Annals of Mathematical Statistics* 33: 420–31.
- Anoulova S, Fisher P, Polt S, Simon HU (1992) PAB-Decisions for boolean and real-valued features. *Fifth Annual Workshop on Computational Learning Theory* 5:353–62. Morgan Kaufman, San Mateo, CA.
- Atick JJ, Redlich AN (1990) Towards a theory of early visual processing. *Neural Computation* 2:308–20.
- Babu CC, Chen WC (1971) An optimal algorithm for pattern classification. *International Journal of Control* 3:577–86.
- Barsov DA (1982) Optimal ridge estimates of the covariance matrix in highdimensional discriminant analysis. *Theory of Probability and Applications* N4:820–21.
- Bishop CM (1995) *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.
- Boser B, Guyon I, Vapnik V (1992) A training algorithm for optimal margin classifiers. *Fifth Annual Workshop on Computational Learning Theory* 5: 144–52. Morgan Kaufman, San Mateo, CA.
- Bourland H, Kamp Y (1988) Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics* 59:291–4.
- Box GEF, Jenkins GM (1976) *Time Series Analysis. Forecasting and control*. Holden-Day, San Francisco, CA.
- Brailovskij VL (1964) An object recognition algorithm with many parameters and its applications. *Engineering Cybernetics* N2:22–30 (in Russian).
- Breiman L, Friedman J, Olshen R, Stone C (1984). *Classification and Regression Trees*. Chapman & Hall, New York,
- Broomhead DS, Lowe D (1988) Multivariable functional interpolation and adaptive networks. *Complex Systems* 2: 321–55.
- Bryant J, Guseman LE Jr (1979) Distance preserving linear feature selection. *Pattern Recognition* 11:347–52.
- Bryson AE, Ho YC (1969) *Applied Optimal Control*. Blaisdell, New York.

- Buchsbaum G, Gottshalk A (1983) Trichromacy opponent colours coding and optimum colour information transmission in the retina. *Proceedings of Royal Society*, London, 220: 1221-3.
- Bunke O, Fisher K (1983) Some fundamentals and procedures of distribution free and discrete discriminant analysis. *Preprint N 51, Neue Folge*. Humbolt Universitat zu Berlin, Sektion Matematik, Berlin.
- Chandrasekaran B, Jain AK (1979) On balancing decision functions. *Journal of Cybernetics and Information Science* 2(2):12-15.
- Cheng B, Titterington D (1994) Neural networks: a review from a statistical perspective. *Statistical Science* 9: 2-54.
- Cherkassky V, Mulier F (1998) *Learning from Data: Concepts, theory, and methods*. John Wiley, New York.
- Chow CK, Liu CN (1968) Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory* IT-14:462-7.
- Cibas T, Fogelman-Soulie F, Gallinari P, Raudys S (1996) Variable selection with neural networks. *Neurocomputing* 12:223-48.
- Cochran WD, Hopkins C (1961) Some classification problems with multivariate qualitative data. *Biometrics* 17:11-31.
- Cortes C, Vapnik V (1995) Support-vector networks. *Machine Learning* 20: 273-97.
- Cover TM (1965) Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers* EC-14:325-34.
- Cybenko G (1989) Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems* 2:303-14.
- Day NE (1969a) Estimating the components of a mixture of normal distributions. *Biometrika* 56:463-74.
- Day NE (1969b) Linear and quadratic discrimination in pattern recognition. *IEEE Transactions on Information Theory* IT-15:419-20.
- Day NE, Kerridge DF (1967) A general maximum likelihood discriminant. *Biometrics* 23:313-23.

- Deev AD (1970) Representation of statistics of discriminant analysis and asymptotic expansions in dimensionalities comparable with sample size. *Reports of Academy of Sciences of the USSR* 195(4):756–62 (in Russian).
- Deev AD (1972) Asymptotic expansions for distributions of statistics  $W, M, W^*$  in discriminant analysis. In Y Blagoveschenskij (editor), *Statistical Methods of Classification*, 31: 6–57. Moscow University Press, Moscow (in Russian).
- Deev AD (1974) Discriminant function designed on independent blocks of variables. *Engineering Cybernetics* N12:153–6 (in Russian).
- Denoeux T, Langelle R (1993) Initialising back-propagation with prototypes. *Neural Networks*, 6(3): 351–361.
- Devijver PA, Kittler J (1982) *Pattern Recognition: A statistical approach*. Prentice-Hall, London.
- Devroye L, Giorfi L, Lugosi G (1996) *A Probabilistic Theory of Pattern Recognition*, Springer, New York.
- Di Pillo PJ (1979) Biased discriminant analysis: evaluation of the optimum probability of misclassification. *Communications in Statistics: Theory and methods* A8(14):1447–57.
- Dong DW (1993) Associative decorrelation dynamics in visual cortex. *Lawrence Berkeley Laboratory Technical Report LBL-34491*. Berkeley, CA.
- Dong DW (1994) Associative decorrelation dynamics: a theory of self organization and optimization in feedback networks. In G Tesauro, DS Touretzky and TK Leen (editors), *Advances in Neural Information Processing Systems*, 7:925–32. MIT Press, Cambridge, MA.
- Do-Tu H, Installe M (1978) Learning algorithms for nonparametric solution to the minimum error classification problem. *IEEE Transactions on Computers* C-27:648–59.
- Duda RO, Hart PE, Stork DG (2000) *Pattern Classification and Scene Analysis*. 2nd edition. John Wiley, New York.
- Duin RPW (1978) *On the Accuracy of Statistical Pattern Recognizers*. Ph.D. dissertation. Delft University of Technology, Delft.
- Duin RPW (1993) Nearest neighbour interpolation for error estimation and classifier optimisation. In KA Hogda, B Braathen, K Heia (editors) *Proceedings of the 8th Scandinavian Conference on Image Analysis*, 5–6. Tromso, Norway.

- Duin RPW (1995) Small sample size generalization. In G. Borgefors (editor), *Proceedings of the 9th Scandinavian Conference on Image Analysis*, 2:957-64.
- Duin RPW (1996) A note on comparing classifiers. *Pattern Recognition Letters* 17:529-36.
- Efron B (1975) The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association* 70:892-8.
- Efron B (1982) *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM, Philadelphia.
- Estes SE (1965) *Measurement Selection for Linear Discriminant used in Pattern Classification*. PhD dissertation. Stanford University, Stanford, CA.
- Fisher RA (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7(2):179-188.
- Fix E, Hodges JLLr (1951) Discriminatory analysis, nonparametric discrimination: consistency properties. *Report No. 4, Project 21-49-004*. USAF School of Aviation Medicine, Randolph Field, TX.
- Fogelman-Soulie F, Vinnert E, Lamy B (1993) Multimodular neural network architectures: application in optical character and human face recognition. *Pattern Recognition and Artificial Intelligence* 5:721-55.
- Foley DH, Sammon JWW (1975) An optimal set of discriminant vectors. *IEEE Transactions on Computers* C-24: 281-9.
- Friedman JM (1989) Regularized discriminant analysis. *Journal of American Statistical Association* 84:165-75.
- Fukunaga K (1990) *Introduction to Statistical Pattern Recognition*. 2nd edition. Academic Press, New York.
- Fukunaga K, Hayes RR (1989) The reduced Parzen classifier. *IEEE Transactions Pattern Analysis and Machine Intelligence* PAMI-11:423-5.
- Fukunaga K, Kessel D (1971) Estimation of classification errors. *IEEE Transactions on Computers* C-20:151-7.
- Fukunaga K, Kessel D (1973) Nonparametric Bayes error estimation using unclassified samples. *IEEE Transactions on Information Theory* IT-19:434-9.
- Geisser S (1964) Posterior odds for multivariate normal classifications. *Journal of the Royal Statistical Society Series B* 21(1):69-76.

- Gelsema ES, Eden G (1980) Mapping algorithms in ISPAHAN. *Pattern Recognition* 12(3):27–36.
- Geman SL, Bienenstock E, Doursat R (1992) Neural networks and bias/variance dilemma. *Neural Computation* 4:1–58.
- Glick N (1978) Additive estimators for probabilities of correct classification. *Pattern Recognition* 10(3):211–22.
- Glucksman H (1966) On improvement of a linear separation by extending the adaptive process with a stricter criterion. *IEEE Transactions on Electronic Computers* EC-15:941–4.
- Goldin SV, Poplavskij NN (1970) Methods to increase a robustness of discriminant function. *Mathematical Methods in Oil Geology and Geophysics. Proceedings of Siberian NIGNI* 36:129–55. Tiumen (in Russian).
- Güler C, Sankur B, Kahya Y, Skurichina M, Raudys S (1996) Classification of respiratory sound patterns by means of cooperative neural networks. In: G.Ramponi, G.L.Sicuranza, S. Carrato, S.Marsi (editors), *Proceedings of 8th European Signal Processing Conference* (isbn 88-86179-83-9). Edizioni Lint, Trieste.
- Gyorgyi G, Tishby N (1990) Statistical theory of learning a rule. In K Thueemann and R Koeberle (editors), *Neural Networks and Spin Glasses*, 31–6. World Scientific, Singapore.
- Griskevicius D, Raudys S (1979) On the expected probability of the classification error of the classifier for discrete variables. In S Raudys (editor), *Statistical Problems of Control*, 38:95–112. Institute of Mathematics and Informatics, Vilnius (in Russian).
- Gupta AK (1977) On the equivalence of two Classification Rules. *Biometrical Journal* 19(5): 365–7.
- Halkaer S, Winter O (1996) The effect of correlated input data on the dynamics of learning. In MC Mozer, MI Jordan and T Petsche (editors), *Advances in Neural Information Processing Systems*, 9:169–75. MIT Press, Cambridge, MA.
- Han CP (1968) A note on discrimination in the case of unequal covariance matrices. *Biometrika* 55:586–7.
- Han CP (1970) Distribution of discriminant function in circular models. *Annals of Institute of Statistical Mathematics* 22(1):117–25.
- Hand DJ (1982) *Kernel Discriminant Analysis*. Research Studies Press, Chichester.

Hand DJ (1986) Recent advances in error rate estimation. *Pattern Recognition Letters* 22:335–46.

Harley TJ (1963) Pseudoestimates versus pseudo-inverses for singular sample covariance matrices. Section 2 in *Report No 5, Contract DA-36-039-SC-90742, AD427172* (Sept, 1963); also MS thesis, Moore School of Electrical Engineering, University of Pennsylvania (1965).

Hausler D, Kearns M, Seung HS, Tishby N (1994) Rigorous learning curves from statistical mechanics. *Proceedings of the 7th Annual ACM Workshop on Computational Learning Theory* 7:76–87. Morgan Kaufman, San Mateo, CA.

Haykin S (1999) *Neural Networks: A comprehensive foundation*. 2nd edition. Prentice-Hall, Englewood Cliffs, NJ.

Hertz J, Krogh A, Palmer RG (1991) *Introduction to the Theory of Neural Computation*. Addison Wesley, Reading, MA.

Hills M (1967) Discrimination and allocation with discrete data. *Applied Statistics* 16(3): 237–50.

Hinton GE (1989) Connectionist learning procedures. *Artificial Intelligence* 40:185–234.

Ho TK, Basu M (2000) Measuring the complexity of classification problems. *Proceedings of 15th International Conference on Pattern Recognition* 2:43–7. IEEE Computer Society Press, Los Alamitos, CA.

Hoerl AE, Kennard RW (1970) Ridge regression: Biased estimation for orthogonal problems. *Technometrics* 12:55–67.

Holmstrom L, Hamalainen A (1993) The self organizing reduced kernel density estimator. *Proceedings of the 1993 IEEE International Conference on Neural Networks* 1:417–421. San Francisco, CA.

Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. *Neural Networks* 2:359–66.

Huber PJ (1981) *Robust Statistics*. John Wiley, New York.

Hughes GF (1965) On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory* IT-14:55–63.

Ibaraki T, Muroga S (1970) Adaptive linear classifier by linear programming. *IEEE Transactions on Systems Science and Cybernetics* SSC-6:53–62.

- Jain AK and Chandrasekaran B (1982) Dimensionality and sample size considerations in pattern recognition practice. In PR Krishnaiah and LN Kanal (edotors), *Handbook of Statistics*, 2:835–85. North-Holland, Amsterdam.
- Jain AK and Dubes R (1978) Feature definition in pattern recognition with small sample size. *Pattern Recognition* 10:85–97.
- Jain AK, Duin RPW and Mao J (2000) Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-22:4– 37.
- Jain AK, Mao J and Mohiuddin M (1996) Neural networks: a tutorial. *IEEE Computer* 29:31–44.
- Jain AK and Waller WG (1978) On the optimal number of features in the classification of multivariate Gaussian data. *Pattern Recognition* 10:367–74.
- Jain AK and Zongker D (1997) Feature selection: evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-19:153–8.
- John S (1961) Errors in discrimination. *Annals of Mathematical Statistics* 32:1125–44.
- Kanal L, Chandrasekaran B (1971) On dimensionality and sample size in statistical pattern classification. *Pattern Recognition* 3:238–55.
- Karnin ED (1990) A simple procedure for pruning back-propagation trained neural networks. *IEEE Transactions on Neural Networks* 1:239–42.
- Keehn D (1965) A note on learning for Gaussian properties. *IEEE Transactions on Information Theory* IT-11:126–31.
- Kharin YS (1996) *Robustness in Statistical Pattern Recognition*. Kluwer Academic Publishers, Dordrecht.
- Kligiene N (1977) Asymptotic estimate of the probability of misclassification of autoregressive sequences. In L Telksnys (editor), *Statistical Problems of Control*, 19:81–102. Institute of Mathematics and Informatics, Vilnius (in Russian).
- Koford JS, Groner GF (1966) The use of an adaptive threshold element to design a linear optimal pattern classifier. *IEEE Transactions on Information Theory* IT-2:42–50.
- Kohonen T (1986) Learning vector quantisation for pattern recognition. *Technical Report TKK-F-A601*. Helsinki University of Technology, Helsinki.

- Kowalczyk A (1996) Model of generalisation error in learning systems with training error selection. In S Amari, L Xu, LW Chan, I King and KS Leung (editors), *Progress in Neural Information Processing: Proc ICONIP'96*, 180–7. Springer, Hong Kong.
- Kraaijeveld MA, Duin RPW (1994) The effective capacity of multilayer feedforward network classifiers. *Proceedings of 12th International Conference on Pattern Recognition 2*: 99–103. IEEE Computer Society Press, Los Alamitos, CA
- Kruskal IB Jr (1956) On the shortest spanning subtree of a graph and the travelling salesman problem. *Proceedings of American Mathematical Society* 7:48–50.
- Kubat M (1998) Decision trees can initialize radial-basis function networks. *IEEE Transactions on Neural Networks* 9:813–21.
- Lachenbruch PA, Goldstein M (1979) Discriminant analysis. *Biometrics* 5(3):9–85.
- Lachenbruch PA, Mickey RM (1968) Estimation of error rates in discriminant analysis. *Technometrics* 10(1):1–11.
- Lbov GS (1965) A selection of an efficient system of statistically dependent variables. In NG Zagoruiko (editor), *Computing Systems*, 19:21–34. Institute of Mathematics, Novosibirsk (in Russian).
- Lbov GS (1966) On representativeness of the sample size while choosing the effective measurement system. In: NG Zagoruiko (editor), *Computing Systems*, 22:39–58. Institute of Mathematics, Novosibirsk (in Russian).
- Lbov GS (1981) *Methods of Processing Experimental Data with Mixed Variables*. Nauka, Novosibirsk (in Russian).
- Le Cun Y (1986) Learning process in an asymmetric threshold network. In E Bienenstock, F Fogelman-Soulie and G Weisbuch (editors), *Disordered systems and biological organizations*, 233–240. Proceedings of NATO Workshop, Les Houches, March 1985. Springer.
- Le Cun Y (1987) *Modeles Connexionistes de l'Apprentissage*. Ph D thesis. University Paris 6.
- Le Cun Y, Denker JS, Solla S (1990) Optimal brain damage. In D Tourecky (editor), *Advances in Neural Information Processing Systems*, 598–605. Morgan Kaufmann, San Mateo, CA.

- Le Cun Y, Kanter I, Solla S (1991) Eigenvalues of covariance matrices: application to neural-network learning. *Physical Review Letters* 66(18):2396–9.
- Levin E, Tishby N, Solla SA (1990). A statistical approach to generalization in layered neural networks. *Proceedings of the IEEE* 78:2133–50.
- Linhart G (1959) Techniques of discriminant analysis with discrete variables. *Metrika* 2:116, 138–40.
- Lim TS, Loh WY, Shih YS (2000) A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning* 40(3):203–28.
- Lipskaya VA (1981) An optimal decision rule for a problem of identification. *Theory of Probabilities and Mathematical Statistics*, 24:91–102. Visha Shkola Publishers, Kiev (in Russian).
- Malinovskij LG (1979) *Objects Classification by Means of Discriminant Analysis*. Nauka, Moscow (in Russian).
- Mao J, Jain A (1993) Regularisation techniques in artificial neural networks. *Proceedings of the World Congress on Neural Networks*, IV:75–9. IEEE Computer Society Press, Portland.
- Mao J, Jain A (1995) Artificial neural networks for feature extraction and multivariate data projection. *IEEE Transactions on Neural Networks* 6:296–317.
- Mazurov VD (1971) Committees of systems of inequalities and recognition problem. *Cybernetics* N3:140–7 (in Russian).
- McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* 5:115–33.
- McKay RJ, Campbell NA (1982a) Variable selection techniques in discriminant analysis. I. Description. *British Journal of Mathematical and Statistical Psychology* 35: 1–29.
- McKay RJ, Campbell NA (1982b) Variable selection techniques in discriminant analysis. II. Allocation. *British Journal of Mathematical and Statistical Psychology*, 35: 30–41.
- McLachlan GJ (1992) *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley, New York.

- Meshalkin LD (1976) Assignment of numerical values to nominal variables. In S Raudys and L Meshalkin (editors), *Statistical Methods of Control*, 14: 49–56. Institute of Mathematics and Informatics, Vilnius (in Russian).
- Meshalkin LD (1977) Theory of Statistical Analysis of Chronical Progressing Diseases. USSR Dr Sc dissertation. Moscow State University, Moscow (in Russian).
- Meshalkin LD, Serdobolskij VI (1978) Errors in classifying multivariate observations. *Theory of Probabilities and Its Applications*, 23:772–81(in Russian).
- Michie D, Spiegelhalter DJ, Taylor CC (editors) (1994) *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, New York.
- Miyake A (1979) Mathematical aspects of optimal linear discriminant function. COMPSAC'79. *Proceedings of the IEEE Computer Society's 3rd International Computer Software and Applications Conference* 161–6. Chicago, Ill.
- Mozer MC, Smolensky P (1989) Skeletonization: a technique for trimming the fat from a network via relevance assessment. In DS Touretzky (editor), *Advances in Neural Information Processing Systems* 1:107–115. Morgan Kaufmann, San Mateo, CA.
- Nieman H (1980) Linear and nonlinear mapping of patterns. *Pattern Recognition* 12(2):83–8.
- Norusis A (1991) Construction of logical (decision tree) classifiers with the top-down search. In S Raudys (editor), *Statistical Problems of Control*, 93:131-158. Institute Mathematics and Informatics, Vilnius (in Russian).
- Okamoto M (1963) An asymptotic expansion for the distribution of linear discriminant function. *Annals of Mathematical Statistics* 34:1286–301, 39:1358–9.
- Opper M, Haussler D (1991) Calculation of the learning curve of Bayes optimal classification algorithm for learning perceptron with noise. *Proceedings of Fourth annual ACM Workshop on Computational Learning Theory* 4:75–87. Morgan Kaufman, San Mateo, CA.
- Parker DB (1985) Learning logic. *Technical Report 47*, Center for Computational Research in Economic and Management Science, MIT Press, Boston.
- Parzen E (1962) On estimation of probability function and mode. *Annals of Mathematical Statistics* 33:1065–76.

- Patrick EA, Shen LYL (1971) Interactive use of problem knowledge for clustering and decision making. *IEEE Transactions on Electronic Computers* EC-10(2):216–22.
- Patterson DW, Mattson RL (1966) A method of finding linear discriminant functions for a class of performance criteria. *IEEE Transactions on Information Theory* IT-12:380–7.
- Pietrantonio H, Jurs PC (1972) Iterative least squares development of discriminant functions for spectroscopic data analysis by pattern recognition. *Pattern Recognition* 4:391–400.
- Pikelis V (1974) *Analysis of Learning Speed of Three Linear Classifiers*. Ph.D. dissertation. Institute of Physics and Mathematics, Vilnius (in Russian).
- Pikelis V (1976) Comparison of methods of computing the expected classification errors. *Automatic and Remote Control* N5:59–63 (in Russian).
- Pikelis V (1991) Calculating statistical characteristics of experimental process for selecting the best version. In S Raudys (editor) *Statistical Problems of Control*, 93:46–56. Institute of Mathematics and Informatics, Vilnius (in Russian).
- Pinsker IS (1973) Estimation of learning method and learning sample. In IS Pinsker (editor) *Simulation and Automatic Analysis of Electrocardiograms*, 13–23. Nauka, Moscow (in Russian).
- Pivoriunas V (1978) The linear discriminant function for the identification of spectra. In S Raudys (editor), *Statistical Problems of Control*, 27:71–90. Institute of Mathematics and Informatics, Vilnius (in Russian).
- Pivoriunas V, Raudys S (1978) On the accuracy of “leaving-one-out” estimate. In S Raudys (editor), *Statistical Problems of Control*, 27:53–70. Institute of Mathematics and Informatics, Vilnius (in Russian).
- Plaut, DC, Nowlan SJ, Hinton GE (1986) Experiments on learning by back propagation. *Technical Report CMU-CS-86-126*. Carnegie-Mellon University, Pittsburg, PA.
- Quinlan JR (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.
- Randles RH, Brofitt JD, Ramberg IS, Hogg RV (1978) Generalized linear and quadratic discriminant functions using robust estimates. *Journal of the American Statistical Association* 73(363):564–8.

Rao CR (1949) On some problems arising of discrimination with multiple characters. *Sankya* 9:343–65.

Rastrigin LA, Erenstein RCh (1981) *Method of Collective Recognition*. Energoizdat, Moscow (in Russian).

Raudys A (2000) Interactive initialisation of the multilayer perceptron. *Pattern Recognition Letters*, 21907-16.

Raudys A, Long A (2000) MLP based linear feature extraction for nonlinearly separable data. *Pattern Analysis and Applications* (accepted).

Raudys S (1967) On determining training sample size of linear classifier. In NG Zagoruiko (editor) *Computing Systems*, 28:79–87. Institute of Mathematics, Novosibirsk (in Russian).

Raudys S (1970) On the problems of sample size in pattern recognition. In VS Pugatchiov (editor) *Detection, Pattern Recognition and Experiment Design*, 2:64–76. Proceedings of the 2nd All-Union Conference Statistical Methods in Control Theory. Nauka, Moscow (in Russian).

Raudys S (1972) On the amount of a priori information in designing the classification algorithm. *Engineering Cybernetics* N4:168–74 (in Russian).

Raudys S (1973) Estimation of probability of misclassification In L Telksnys (editor), *Statistical Problems of Control*, 5:9–44. Institute of Mathematics and Informatics, Vilnius (in Russian).

Raudys S (1976) Limitation of sample size in clasification problems. *Statistical Problems of Control*, 18:1-186. Institute of Mathematics and Informatics, Vilnius (in Russian).

Raudys S (1979a) Classification errors when features are selected. In S Raudys (editor), *Statistical Problems of Control*, 38:9–26. Institute of Mathematics and Informatics, Vilnius (in Russian).

Raudys S (1979b) Determination of optimal dimensionality in statistical pattern classification. *Pattern Recognition* 11: 263–71.

Raudys, S. (1981) Influence of sample size on the accuracy of model selection in pattern recognition. In S Raudys (editor), *Statistical Problems of Control*, 50:9–30. Institute of Mathematics and Informatics, Vilnius (in Russian).

Raudys S (1987) On the accuracy of model selection in data analysis. *Proceedings of the III-rd International Conference on Data Analysis and Informatics* 1:91–100. INRIA Press, Paris.

- Raudys S (1988) On the accuracy of a bootstrap estimate of the classification error. *Proceedings of 9th International Joint Conference on Pattern Recognition*, 1230–2. IEEE press, Los Alamitos, CA.
- Raudys, S (1991) On the effectiveness of Parzen window classifier. *Informatica* 2:434–54. Institute of Mathematics and Informatics, Vilnius.
- Raudys S (1993) On shape of pattern error function, initializations and intrinsic dimensionality in ANN classifier design. *Informatica* 4:360–83. Institute of Mathematics Informatics, Vilnius.
- Raudys S (1994) Why do ANN classifiers have favourable small sample properties? In: ES Gelsema and LS Kanal (editors), *Pattern Recognition in Practice IV*: 287–298. Elsevier Science, Amsterdam.
- Raudys S. (1995a) Generalization of linear and non-linear adaptive classifiers. *Proceedings ICANN'95* 1:183–90. EC2 & Cie Press, Paris.
- Raudys S (1995b) A negative weight decay or antiregularisation, *Proceedings ICANN'95* 2:449–54. EC2 & Cie Press, Paris.
- Raudys, S (1996) Linear classifiers in perceptron design. *Proceedings of 13th International Conference on Pattern Recognition* 4(D):763–7. IEEE Computer Society Press, Los Alamitos, CA.
- Raudys S (1997) On dimensionality, sample size and classification error of nonparametric linear classification algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-19:669–71.
- Raudys S (1998a) Use of statistical hypothesis in neural network design. In: SA Aivazian, YS Kharin (editors), *Computer Data Analysis and Modelling*, 2:55–63. Minsk University Press, Minsk.
- Raudys S (1998b) Evolution and generalization of a single neurone. I. SLP as seven statistical classifiers. *Neural Networks* 11: 283–96.
- Raudys S (1998c) Evolution and generalization of a single neurone. II. Complexity of statistical classifiers and sample size considerations. *Neural Networks* 11:297–313.
- Raudys S (2000a) How good are support vector machines? *Neural Networks* 13:9–11.
- Raudys S (2000b) Evolution and generalization of a single neurone. III. Primitive, regularized, standard, robust and minimax regressions. *Neural Networks* 13(4-5):507–23.

- Raudys S (2000c) Scaled rotation regularization. *Pattern Recognition* 33:1989–98.
- Raudys S (2000d) Classifier's complexity control while training multilayer perceptrons. In: F Ferri, JM Inest, A. Amin and P Pudil (editors), *Advances in Pattern Recognition*, 32–44. Springer Lecture Notes in Computer Science, 1876.
- Raudys S, Amari S (1998) Effect of initial values in simple perception. In *Proceedings 1998 IEEE World Congress on Computational Intelligence IJCNN'98*:1530–5. IEEE Press.
- Raudys S, Duin RPW (1998). On expected classification error of the Fisher classifier with pseudoinverse covariance matrix. *Pattern Recognition Letters* 19:385–92.
- Raudys S, Jain AK (1991) Small sample size effects in statistical pattern recognition: recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-13:252–64.
- Raudys S, Pikelis V, Juskevicius K (1975) Experimental comparison of thirteen classification algorithms. In S Raudys (editor), *Statistical Problems of Control*, 11:53-80. Institute of Mathematics and Informatics, Vilnius (in Russian).
- Raudys S, Pikelis V (1980) On dimensionality, sample size, classification error and complexity of classification algorithm in pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-2:242–52.
- Raudys S, Pikelis V (1982) Collective selection of the best version of a pattern recognition system. *Pattern Recognition Letters* 1:7–13.
- Raudys S, Saudargiene A (1998) Structures of the covariance matrices in the classifier design. In: A Amin, D Dori, P Pudil, H Freeman (editors), *Advances in Pattern Recognition*, 583–92. Springer Lecture Notes in Computer Science, 1451.
- Raudys S, Saudargiene A (2001) Tree type dependency model and sample size - dimensionality properties. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23(1).
- Raudys S, Saudargiene A, Povilonis E (2000) The bias evaluation in the model selection. *Pattern Analysis and Applications* (submitted)
- Raudys S, Skurichina M (1994) Small sample properties of ridge estimate of the covariance matrix in statistical and neural net classification. In EM Tiit, T Kollo, H Niemi (editors), *New Trends in Probability and Statistics: Multivariate statistics and matrices in statistics*, 3:237–45. TEV, Vilnius and VSP, Utrecht.

- Raudys S, Skurichina M (1992) The role of the number of training samples on weight initialisation of artificial neural net classifier. In *Neuroinformatics and Neurocomputers. Proceedings of RNNS/IEEE Symposium*, 343-353. IEEE Press.
- Raudys S, Skurichina M, Cibas T, Gallinari P (1995) Ridge estimates of the covariance matrix and regularization of artificial neural network classifier. *Pattern Recognition and Image Processing* N4:633-50. International Journal of Russian Academy of Sciences, Moscow.
- Raudys S, Vaitukaitis V (1984) Methods to estimate the probability of misclassification. In: S Raudys (editor), *Statistical Problems of Control*, 66:43-65. Institute of Mathematics and Informatics, Vilnius (in Russian).
- Reed R (1993) Pruning algorithms: a survey. *IEEE Transactions on Neural Networks*, 4:740-7.
- Reed R, Marks RJ, Oh S (1995) Similarities of error regularization, sigmoid gain scaling, target smoothing, and training with jitter. *IEEE Transactions on Neural Networks* 6:529-38.
- Ripley B, (1994) Neural networks and related methods for classification. *Journal of Royal Statistical Society Series B56*: 409-56.
- Rosenblatt F (1958) The Perceptron: A probabilistic model for information storage in the brain. *Psychological Review* 65:386-408.
- Rosenblatt M (1956) Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics* 27:832-37.
- Rumelhart DE, Hinton GE, Williams RJ (1986) Learning internal representations by error propagation. In: DE Rumelhart, JL McClelland (editors), *Parallel Distributed Processing: Explorations in the microstructure of cognition* I:318-62. Bradford Books, Cambridge, MA.
- Sammon JW (1969) A non-linear mapping for data structure analysis. *IEEE Transactions on Computers* C-18:401-9.
- Saudargiene A (1999) Structurization of the covariance matrix by process type and block diagonal models in the classifier design. *Informatica* 10(2): 245-69.
- Schalakoff RJ (1992) *Pattern Recognition: Statistical, structural and neural approaches*. John Wiley, New York.
- Schuermann J (1977) *Polynomklassifikatoren für Zeichenerkennung*. Oldenbourg, Verlag, Munchen-Wien.

- Schuermann J (1996) *Pattern Classification: A unified view of statistical and neural approaches*, John Wiley, New York.
- Scott D (1992) *Multivariate Density Estimation: Theory, practice, visualization*. John Wiley, New York.
- Sebestyen GS (1962) *Decision-Making Process in Pattern Recognition*. Macmillan, New York.
- Serdobolskij VI (1979) The moments of discriminant function and classification for a large number of variables. In S Raudys (editor), *Statistical Problems of Control*, 38:27–51. Institute of Mathematics and Informatics, Vilnius (in Russian).
- Serdobolskij VI (1983a) On minimal error probability in discriminant analysis. *Reports of Academy of Sciences of the USSR* 270:1066–70 (in Russian).
- Serdobolskij VI (1983b) The influence of selecting components of random variable on classification. *Izvestija VUZOV USSR, series Mathematics* N9:46–55. Moscow (in Russian).
- Sethi I and Jain AK (editors) (1991) *Neural Networks and Statistical Pattern Recognition*, North-Holland, Amsterdam.
- Seung HS, Sompolinsky H, Tishby N (1992) Statistical mechanics from examples. *Physical Review A* 45(8): 6056–91.
- Shlesinger M (1968) Interrelation between learning and self-learning in pattern recognition. *Cybernetics* (translated from Russian Consult Bureau, NY) 4(2):81–8.
- Siadliecki, W, Siadlecka K, Sklansky J (1988) An overview of mapping techniques for exploratory pattern analysis. *Pattern Recognition* 21:411–29.
- Silverman BW (1986) *Density Estimation in Statistics*. Chapman and Hall, New York.
- Sitgreaves R (1961) Some results on the distribution of the W-classification statistics. In H Solomon (editor), *Studies in Item Selection and Prediction*, 241–61. Stanford University Press, Stanford, CA.
- Sjoberg J, Ljung L (1992) Overtraining, regularization, and searching for minimum in neural networks. *Technical Report S-581 83*. Department of Electrical Engineering, Linköping University, Sweden.
- Skurichina M (1991) Effect of the kernel functional form on the quality of nonparametric Parzen window classifier. In S Raudys (editor), *Statistical Problems of Control*, 93:167–81. Institute Mathematics and Informatics, Vilnius (in Russian).

- Skurichina M, Raudys S, Duin RPW (2000) K-nearest neighbors directed noise injection in multilayer perceptron training, *IEEE Transactions on Neural Networks*, 11 504–11.
- Smith FW (1971) Design of minimum-error optimal classifiers for patterns form distributions with Gaussian tails. *IEEE Transactions on Information Theory* IT-17:701–7.
- Smith FW (1972) Small-sample optimality of design techniques for linear classifiers of Gaussian patterns. *IEEE Transactions on Information Theory* IT-18:118–26.
- Sohn SY (1999) Meta analysis of classification algorithms for pattern recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-21:1137–44.
- Solomon H (1956) Probability and statistics in psychometric research. In IJ Neyman (editor), *Proceedings of 3rd Berkley Symposium on Mathematical Statistics and Probability*, 169–84. University California Press, Berkley, CA.
- Somol P, Pudil P, Novovočova J and Paclik P (1999) Adaptive floating search methods in feature selection. *Pattern Recognition Letters* 20:1157-63.
- Takeshita T, Toriwaki J (1995) Experimental study of performance of pattern classifiers and the size of design samples. *Pattern Recognition Letters* 16:307–12.
- Tax DMJ, Breukelen M, Duin RPW, Kittler J (2000) Combining multiple classifiers by averaging or by multiplying? *Pattern Recognition* 33:1475–85.
- Tou JT, Gonzales RC (1974) *Pattern Recognition Principles*. Addison-Wesley, Reading, MA.
- Toussaint GT (1972) Polynomial representation of classifiers with dependent discrete-valued features. *IEEE Transactions on Computers* C-21(2):205–8.
- Toussaint GT (1974) Bibliography on estimation of misclassification. *IEEE Transactions on Information Theory* IT-20(4):472–9.
- Tsympkin Ya Z (1966) Adaptation, learning and self-learning in automatic systems. *Automatic and Remote Control* N1:23–61 (in Russian).
- Tsympkin Ya Z (1973) *Foundations of the Theory of Learning Systems*. Academic Press, New York.
- Tubbs JD, Coberley WA, Young DM (1982) Linear dimension reduction and Bayes classification with unkown parameters. *Pattern Recognition* 14(3):167–172.

- Vajda I (1970) Note on discrimination information and variation. *IEEE Transactions on Information theory* IT-16:771-773.
- Valiant LG (1984) A theory of the learnable. *Communications of ACM*, 27:1134-42.
- Van der Smagt PP (1994) Minimization methods for training feedforward neural networks. *Neural Networks* 7: 1-11.
- Vapnik VN (1976) A method of minimization of sum risk in problem of pattern recognition. In S Raudys and L Meshalkin (editors), *Statistical Problems of Control*, 14:174-7. Institute of Mathematics and Informatics, Vilnius (in Russian).
- Vapnik VN (1982) *Estimation of Dependencies Based on Empirical Data*. Springer, New York
- Vapnik VN (1995) *The Nature of Statistical Learning Theory*. Springer, Berlin.
- Vapnik VN, Chervonenkis DYa (1968) On uniform convergence of relative frequencies of events to their probabilities. *Soviet Math Doklady* 9:915-8.
- Vapnik VN, Chervonenkis DYa (1974) *Theory of Pattern Recognition: Statistical learning problems*. Nauka, Moscow (in Russian).
- Vidyasagar M (1997) *A Theory of Learning and Generalization*. Springer, London.
- Wald A (1944) On a statistical problem in the classification of an individual into one of two groups. *Annals of Mathematical Statistics* 15:145-62.
- Wald A (1950) *Statistical Decision Functions*. John Wiley, New York.
- Wand M, Jones M (1995) *Kernel Smoothing*. Chapman & Hall, New York.
- Warmack RE, Gonzales RC (1973) An algorithm for optimal solution of linear inequalities and its application to pattern recognition. *IEEE Transactions on Computers* C-22:1065-75.
- Welch BL (1939) Note on discriminant functions. *Biometrika* 31:218-20.
- Werbos PJ (1974) *Beyond Regression: New tools for prediction in the behavioral sciences*. Ph.D. Thesis, Harvard University Cambridge, MA.
- Widrow B, Hoff ME (1960) Adaptive switching circuits. *WESCON Convention Record*, 4:96-104.

- Wolf AC (1966) The estimation of the optimum linear decision function with a sequential random method. *IEEE Transactions on Information Theory* IT-12:312–5.
- Wolpert DH (editor) (1995) *The Mathematics of Generalization*. Proceedings of the SFI/CNLS workshop on formal approach to supervised learning (Santa Fe (Studies in the Sciences of Complexity)). Santa Fe Institute, Santa Fe, NM.
- Wolverton CT, Wagner TJ (1969) Asymptotically optimal discriminant functions for pattern classification. *IEEE Transactions on Information Theory* IT-15:258–65.
- Wyman F, Young D, Turner D (1990) A comparison of asymptotic error rate expansions for the sample linear discriminant function. *Pattern Recognition* 23:775–83.
- Yacoub M, Bennani Y (1997) A heuristic for variable selection in multilayer artificial neural network classifier. In C Dagi, M Akay, O Ersoy, B Fernandez and A Smith (editors) *Intelligent Engineering Systems Through Artificial Neural Networks* 7:527–32. ASME, New York.
- Yau SS, Schumpert JM (1968) Design of pattern classifiers with the updating property using stochastic approximation techniques. *IEEE Transactions on Computers* C-17:908–1003.
- Young TY, Calvert TW (1974) *Classification Estimation and Pattern Recognition*. Elsevier Science, New York.
- Zarudskij VI (1979) The use of models of simple dependence problems of classification. In S Raudys (editor), *Statistical Problems of Control* 38:33–75. Institute of Mathematics and Informatics, Vilnius (in Russian).
- Zarudskij VI (1980) Determination of some graph connections for normal vectors in large dimensional case. In SA Aivazian (editor) *Algorithmic and Programic Supply of Applied Multivariate Statistical Analysis*, 189–208. Nauka, Moscow (in Russian).
- Zhezel YN (1975) A minimal essentially complete class of discriminant rules when the higher moments are unknown. *Theory of Probability and Applications* 19:832–4

# Index

- Activation function 7, 10, 136
  - hard limiting 8, 12
  - sigmoid 7, 143, 187
  - tanh 8, 10, 136, 147, 179
- Antiregularisation 142, 146, 175
- Anderson–Bahadur 34, 45, 72
- Apparent error 82, 109, 212
  - in selection 211, 242
- Architectural approach 63
- Asymptotics
  - conventional 6, 81, 133, 135
  - double 84, 92, 98, 103, 105, 133
  - thermodynamic limit 82, 133
- Autoregression 38, 258
  
- Back propagation 10, 26
- Bayes predictive approach 31, 39, 72,
- Bias
  - correction in classification 88, 106
  - correction in model selection 247
- Block diagonal matrices 36
- Bootstrap 213, 232, 251
  
- Capacity 81, 109
  - effective 168, 192
- Categorical data 67, 187
- Chaos 183
- Classification error
  - apparent 83, 109, 212
  - asymptotic 14, 21, 79, 84, 87, 92, 100, 128, 230
  - Bayes 14, 78, 210
  - conditional 16, 21, 78, 93, 227
  - expected 16, 79, 158, 210, 229
  - empirical 22, 83, 109, 212
  - leaving-one-out 213, 214, 230
- Classifier
  - Anderson–Bahadur 34, 45, 72
  - decision tree 63, 69, 73, 188
  - EDC 3, 22, 33, 115, 118, 136, 165
  - Fisher 5, 12, 22, 32, 119, 139, 159, 163, 229, 237
  - generalised Fisher 48
  - $k$ - $NV$  55, 73, 127
  - maximal margin 12, 58, 117, 142, 202
  - minimum empirical error 6, 57, 141
  - multinomial 66, 69, 73, 128, 187
  - Parzen window 51, 73, 120
  - piecewise-linear 50, 60
  - polynomial 56, 71, 146
  - potential functions 56
  - Pseudo-Fisher 44, 98, 139, 159
  - quadratic 30, 32, 36, 42, 146
  - regularised 45, 100, 138, 174, 182, 198
  - robust 47, 73, 141
  - selection 209
  - support vector 59, 142, 146
- Cluster analysis 50, 148
- Cost function 7, 136
  - dynamics 140, 152
  - surface 150, 154
- Complexity control 18, 175
- Common parameters 98, 103, 105
- Co-operation 66, 185
- Correlations 4, 32
- Covariance matrix 4, 34, 43, 93, 195
- Cross-validation
  - accuracy 226, 230
  - leaving-one-out method 213, 214, 230
  - hold-out method 211
  - $k$ -fold cross-validation 213
- Curse of dimensionality 18, 134, 237

- Data transformation
  - linear 198
  - mapping 219
  - non-linear 146, 199, 220
  - whitening 160, 196
- Decision
  - boundary 3, 6, 10, 32, 53, 121, 200
  - functions approach 4, 14, 27
  - tree 63, 69, 73, 188
- Dependence
  - tree type 37, 134, 255
  - temporal 38, 258
- Design set 2, 211, 226, 209
- Dimensionality
  - effective 17, 90
  - intrinsic 91, 116, 123, 164
  - optimal 21, 237, 252
  - reduction 21, 218, 241
- Distance
  - effective 14, 16, 89
  - Euclidean 3, 13, 53, 63
  - Mahalanobis 13, 92, 216, 222, 231
- Distribution
  - binomial 227, 245
  - conditional 27, 113, 244
  - favourable 88, 94, 108, 108, 130
  - multivariate Gaussian 30, 32
  - mixture 49, 61
  - non-spherical 4, 89
  - posterior 39, 75, 112
  - prior 39, 68, 74, 112
  - spherical 4, 86
  - unfavourable 88, 94, 111, 130
- Divergence 222
- Double asymptotics 84, 92, 98, 103, 105, 133
- Early stopping 18, 172, 174, 185, 206
- Eigenvalues 44, 91, 98, 123, 155, 198
- Eigenvectors 44, 91, 98, 155, 198
- Effective
  - dimensionality 17, 90
  - distance 14, 89
- Empirical error 6, 58, 83, 109, 141, 212
- Entropic loss 223, 233,
- Error bounds 109, 117
- Error counting 212, 215, 226
- Feature
  - definition 21
  - extraction 201, 219
  - ranking 237
  - selection 222
- Generalisation error 15, 77, 156, 229
  - and complexity 17, 19, 87, 159
- Gradient 8, 10, 136, 182
- Hold-out method 211, 226
- Huber M-estimates 47
- Ideal error in selection 242
- Initialisation 51, 118, 148, 169, 194, 204
- Intrinsic dimensionality 91, 116, 123, 164
- Kernel function 53
- k*-NN
  - classification rule 55, 73, 127
  - directed noise injection 178
- Learning curves 23, 81, 118
- Learning dynamics 11, 135, 142, 174, 197
- Learning quantity 87, 96, 115
- Learning step 8, 42, 136, 159
  - exponential increase of 23, 117, 144, 180
- Learning vector quantisation 51, 63, 149
- Leaving-one-out method 57, 213, 214, 230
- Local experts 185
- Local minima 151, 179
- Logistic regression 47
- Mixture of experts 186
- Multinomial classifier 66, 69, 129, 189
- Multilayer perceptrons
  - architecture 8
  - complexity control 173, 184
  - decision boundary 9, 146, 182
  - error surface 154
  - generalisation 161
- Newton method 182
- Noise injection
  - coloured 178, 188
  - inputs 176, 182, 183, 186
  - outputs 178, 188
  - weights 178

- Optimal values
  - of smoothing parameter 124
  - of training parameters 101, 121, 181
- Outliers 47, 141, 175, 197
- Overtraining 22, 167
  - and initialisation 23, 172
  - and early stopping 24, 174, 180
- Piecewise-linear 10, 50, 60, 149
- Peaking phenomenon 20, 158, 237, 240
- Plug-in approach 31, 36, 68, 105
- Polynomial 56, 71, 146, 235
- Principal components 4, 91, 98, 123
- Probability of misclassification – *see*  
   *Classification error*
- Process
  - chaos 183
  - noise 183
  - random 38, 258
- Pruning 21, 147, 281
- Pseudoinverse 44, 98, 139, 159
- Quadratic classifier 30, 32, 36, 42, 146
- Quadratic cost 7, 136
- Radial basis functions 50, 146, 148
- Random search 6, 112
- Regularisation 174, 181, 188
- Resubstitution 212, 228
- Robust discriminant analysis 47, 73, 141
- Saliency 223, 234
- Sample size
  - complexity relations and 18, 87, 93
  - equal in both populations 10, 16, 41,  
   86, 92, 97
  - unequal in both populations 41, 88,  
   105
- Scaled rotation 46
- Scissors effect 18, 99, 159
- Sigmoid scaling 184
- Single layer perceptron
  - as statistical classifiers 136
  - evolution 136
  - generalisation 142, 156
- Sigmoid function 7
- Selection
  - accuracy 240
  - bias 246
  - feature 222, 225
  - loss 247
  - model 193, 209
- Separability 32, 141, 222
- Singular data 91, 116, 156, 160, 196
- Smoothed error rate estimator 215
- Smoothing
  - parameter 53, 71
  - optimal 124
- Statistical–mechanics approach 81, 132
- Structured CM 34, 104, 134, 195, 203
- Support vector machines 59, 142, 146
- Targets
  - generalisation and 24, 158
  - boundary values 8, 164, 179
  - proximal 8, 164, 179
  - value control 159, 179
- Theory of the extremes 244
- Thermodynamic limit 82, 84
- Training algorithm 10, 26, 155, 173
- Training dynamics 10, 136, 144, 197
- Training error 83, 109, 212
- Transformed feature space 145, 198
- True error in selection 242
- Voting 186
- Vapnik–Chervonenkis
  - dimension 109
  - bounds 109, 117
  - effective 1668
- Weights
  - dynamics 140, 152
  - growth 141, 175
  - initial 8, 23, 64, 66, 118, 148, 169
  - space 151, 154
  - common for all classes 162
  - different for all classes 162
- Weight decay 174
- Whitening transformation 146, 181, 196
- Zero empirical error classifier 111, 115

