A 3-D Contextual Bayesian Classifier

Rasmus Larsen

Abstract— **In this paper we will consider an extension of the Bayesian 2-D contextual classification routine developed by Owen, Hjort & Mohn to 3 spatial dimensions. It is evident that compared to classical pixelwise classification further information can be obtained by taking into account the spatial structure of image data, i.e. neighbouring pixels tend to be of the same class. The algorithm developed by Owen, Hjort & Mohn consists of basing the classification of a pixel on the simultaneous distribution of the values of a pixel and its four nearest neighbours. This includes the specification of a Gaussian distribution for the pixel values as well as a prior distribution for the configuration of class variables within the cross that is made of a pixel and its four nearest neighbours. We will extend this algorithm to 3-D, i.e. we will specify a simultaneous Gaussian distribution for a pixel and its 6 nearest 3-D neighbours, and generalise the class variable configuration distribution within the 3-D cross. The algorithm is tested on a synthetic 3-D multivariate dataset.**

Keywords—**Classification, Segmentation, Contextual methods**

I. Introduction

WHEN applying classical classification schemes in image analysis the spatial structure of the datasets is neglected. This is non-satisfying, because further information obviously can be drawn from the spatial arrangement of pixels, i.e. neighbouring pixels tend to be of the same class. We will refer to this type of information as contextual information.

Contextual information can be taken into account in a number of ways when performing classification. One important way is to include (derived) features that hold information of the neighbourhood of a given pixel, i.e. contextual features. Another way to take the spatial nature into account is in the analysis. Several algorithms have been proposed in the 2-D case. In [1] it is proposed simply to augment the feature vector with the average of the feature vector from the four neighbouring pixels. In order to find the maximum a posteriori estimate in a Markov random field model stochastic relaxation has been proposed in [2]. An approximation to the maximum a posteriori estimate using iterated conditional modes was proposed in [3]. In [4], [5], [6] a classification scheme for 2-D images that bases the actual classification of pixel on the feature vectors of the pixel itself and those of the 4 nearest neighbours is introduced. In [6] it is assumed that classes of the nearest neighbours of a pixel are conditionally independent given the class of the center pixel, whereas in [4], [5] it is assumed that the pixel size is small relative to the grains of the pattern under study, which leads to a vastly reduced set of possible class configurations among a pixel and it four nearest neighbours.

In this article we will extend the algorithm proposed in [4], [5] to 3-D images, and carry out a series of tests on a synthetic 3-D image.

II. Methods

In this Section we will develop a contextual classification rule, specify a Gaussian distribution for the observed (and derived) features, and specify a prior distribution for the class variable.

A. Construction of a Contextual Classification Rule

Suppose that a pixel is an observation from one of the classes (populations) $\pi_1, \pi_2, \ldots, \pi_k$. The classification of the observation depends on the vector of features $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^T$ of that pixel. Furthermore, let us assume knowledge of the prior distribution of the classes, i.e. the prior probabilities, $P(C = \pi_i) = p_i$, $i = 0, 1, \ldots, k$ where C is the class variable. This distribution determines the probability with which an arbitrary feature vector has been generated from a particular class.

We will denote the feature vector of the neighbouring pixels $\boldsymbol{X}_N$, $\boldsymbol{X}_E$, $\boldsymbol{X}_S$, $\boldsymbol{X}_W$, $\boldsymbol{X}_T$, and $\boldsymbol{X}_B$ for the north, east, south, west, top, and bottom pixel, respectively. The augmented feature vector consisting of the feature vector for the pixel and its neighbours will be denoted $\boldsymbol{D} = (\boldsymbol{X}^T, \boldsymbol{X}_N^T, \boldsymbol{X}_E^T, \boldsymbol{X}_S^T, \boldsymbol{X}_W^T, \boldsymbol{X}_T^T, \boldsymbol{X}_B^T)^T$.

We obtain the Bayes solution for the case of equal losses by setting the discriminant score equal to the maximum a posteriori probability. The posterior distribution for the class variable becomes

$$f(\pi_\nu \mid \boldsymbol{d}) = P(C = \pi_\nu \mid \boldsymbol{D} = \boldsymbol{d}) = \frac{P(C=\pi_\nu)P(\boldsymbol{D}=\boldsymbol{d}|C=\pi_\nu)}{\sum_{i=1}^{k} P(C=\pi_i)P(\boldsymbol{D}=\boldsymbol{d}|C=\pi_i)} =$$

$$\frac{\sum_{a,b,c,d,e,f} p_\nu P(\boldsymbol{D}=\boldsymbol{d}|\boldsymbol{C}=(\pi_\nu,\pi_a,\pi_b,\pi_c,\pi_d,\pi_e,\pi_f)) g(\pi_a,\pi_b,\pi_c,\pi_d,\pi_e,\pi_f|\pi_\nu)}{h(\boldsymbol{d})}$$

where $h(\boldsymbol{d})$ is the unconditional density of the augmented feature vector, (a, b, c, d, e, f) is one of the possible k^6 configurations of the class variables of the neighbouring pixels, $\boldsymbol{C}$ is the class configuration corresponding to the augmented feature vector $\boldsymbol{D}$, and $g(\pi_a, \pi_b, \pi_c, \pi_d, \pi_e, \pi_f \mid \pi_\nu)$ is the probability of the configuration of the class variables of the neighbouring pixels given that the center pixel has class π_ν.

Contextual information comes into the model in two ways, first in the spatial dependence of the feature vectors (specification of the conditional distribution of the augmented feature vector), and second in the specification of g.

B. Specification of a Gaussian distribution

We assume that each feature vector may be written as a sum of two terms, i.e.

$$\boldsymbol{X} = \boldsymbol{Y} + \boldsymbol{\epsilon} \tag{1}$$

where

$$\begin{aligned} &\boldsymbol{Y} \mid C = \pi_i \in N(\boldsymbol{\mu}_i, (1-\theta)\boldsymbol{\Sigma}) \\ &(\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_N) \text{ multinormal with} \\ &\quad E\{\boldsymbol{\epsilon}_j\} = 0, E\{\boldsymbol{\epsilon}_{j1}\boldsymbol{\epsilon}_{j2}\} = \rho^{|j1-j2|}\theta\boldsymbol{\Sigma} \end{aligned} \tag{2}$$

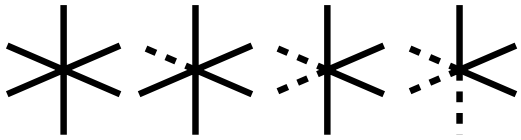
The author is with the Section for Image Analysis, Department of Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark. Phone: +45 4588 1433, FAX: +45 4588 1397

The $\boldsymbol{Y}$ terms are independent given the classes and model the class dependency of the feature vectors, whereas the $\boldsymbol{\epsilon}$'s are autocorrelated noise terms. The indices j, $j1$, and $j2$ refer to pixel numbers, and $|j1 - j2|$ is the Euclidean distance between pixels $j1$ and $j2$, N is the total number of pixels. ρ is the autocorrelation between first order neighbours, and θ is the proportion of the covariance matrix $\boldsymbol{\Sigma}$ that is due to autocorrelated noise. Note that we choose to use an isotropic autocorrelation function, the extension to an anisotropic function is straightforward.

Now it is possible to write the conditional distribution of the augmented feature vector

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{X}_N \\ \boldsymbol{X}_E \\ \boldsymbol{X}_S \\ \boldsymbol{X}_W \\ \boldsymbol{X}_T \\ \boldsymbol{X}_B \end{bmatrix} \in N_{7p} \left[\begin{bmatrix} \boldsymbol{\mu}_\nu \\ \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \\ \boldsymbol{\mu}_c \\ \boldsymbol{\mu}_d \\ \boldsymbol{\mu}_e \\ \boldsymbol{\mu}_f \end{bmatrix}, \begin{bmatrix} 1 & \alpha & \alpha & \alpha & \alpha & \alpha & \alpha \\ \alpha & 1 & \beta & \gamma & \beta & \beta & \beta \\ \alpha & \beta & 1 & \beta & \gamma & \beta & \beta \\ \alpha & \gamma & \beta & 1 & \beta & \beta & \beta \\ \alpha & \beta & \gamma & \beta & 1 & \beta & \beta \\ \alpha & \beta & \beta & \beta & \beta & 1 & \gamma \\ \alpha & \beta & \beta & \beta & \beta & \gamma & 1 \end{bmatrix} \otimes \boldsymbol{\Sigma} \right] \tag{3}$$

where $\otimes$ denotes the tensor product, and given that the classes are π_ν, π_a, π_b, π_c, π_d, π_e, and π_f, respectively. Furthermore $\alpha = \rho\theta$, $\beta = \rho^{\sqrt{2}}\theta$, and $\gamma = \rho^2\theta$ are the correlations between first-order, second-order, and third-order neighbours, respectively.



(a) P (b) Q (c) R (d) S

Fig. 1. Patterns in the model. Within the "cross", that represents the neighbourhood of a pixel, i.e. the six nearest neighbours, it it is assumed that at most two classes are present, and that the only possible configurations are these.

C. Specification of a g

Assuming that pixels in a scene are assigned populations by a stochastic process, we regard a scene with pixels that have not been assigned populations. As the first step in the process we divide the scene by planes distributed by a stochastic process. Each pixel will now be part of a region. If the size of the regions are large compared to the pixel size, it can be assumed that on the borders between regions we only have the patterns shown in Figure 1.

By rotation we obtain six, twelve, and eight different Q, R, and S patterns, respectively. These patterns are assigned positive a priori probabilities, while all other patterns are assigned the probability zero. The probabilities for P, Q, R, and S are denoted p, q, r, and s, respectively.

As the second step we assign a population to each region independently, according to the a priori probabilities for the populations. If two neighboring regions are assigned the same population we can delete the border between these regions.

Under these assumptions we have the following expression for the probabilities, for each of the possible patterns.

$$\begin{aligned}
P:\ & g(\pi_\nu,\pi_\nu,\pi_\nu,\pi_\nu\pi_\nu,\pi_\nu \mid \pi_\nu) = p + (q+r+s)\cdot p_\nu \\
Q:\ & g(\pi_i,\pi_\nu,\pi_\nu,\pi_\nu,\pi_\nu,\pi_\nu \mid \pi_\nu) = \\
& g(\pi_\nu,\pi_i,\pi_\nu,\pi_\nu,\pi_\nu,\pi_\nu \mid \pi_\nu) = \\
& g(\pi_\nu,\pi_\nu,\pi_i,\pi_\nu,\pi_\nu,\pi_\nu \mid \pi_\nu) = \\
& g(\pi_\nu,\pi_\nu,\pi_\nu,\pi_i,\pi_\nu,\pi_\nu \mid \pi_\nu) = \\
& g(\pi_\nu,\pi_\nu,\pi_\nu,\pi_\nu,\pi_i,\pi_\nu \mid \pi_\nu) = \\
& g(\pi_\nu,\pi_\nu,\pi_\nu,\pi_\nu,\pi_\nu,\pi_i \mid \pi_\nu) = \tfrac{1}{6} q p_i \\
R:\ & g(\pi_i,\pi_i,\pi_\nu,\pi_\nu,\pi_\nu,\pi_\nu \mid \pi_\nu) = \\
& g(\pi_\nu,\pi_i,\pi_i,\pi_\nu,\pi_\nu,\pi_\nu \mid \pi_\nu) = \\
& g(\pi_\nu,\pi_\nu,\pi_i,\pi_i,\pi_\nu,\pi_\nu \mid \pi_\nu) = \\
& g(\pi_i,\pi_\nu,\pi_\nu,\pi_i,\pi_\nu,\pi_\nu \mid \pi_\nu) = \\
& g(\pi_i,\pi_\nu,\pi_\nu,\pi_\nu,\pi_i,\pi_\nu \mid \pi_\nu) = \\
& g(\pi_\nu,\pi_\nu,\pi_i,\pi_\nu,\pi_i,\pi_\nu \mid \pi_\nu) = \\
& g(\pi_i,\pi_\nu,\pi_\nu,\pi_\nu,\pi_\nu,\pi_i \mid \pi_\nu) = \\
& g(\pi_\nu,\pi_\nu,\pi_i,\pi_\nu,\pi_\nu,\pi_i \mid \pi_\nu) = \\
& g(\pi_\nu,\pi_i,\pi_\nu,\pi_\nu,\pi_i,\pi_\nu \mid \pi_\nu) = \\
& g(\pi_\nu,\pi_\nu,\pi_\nu,\pi_i,\pi_i,\pi_\nu \mid \pi_\nu) = \\
& g(\pi_\nu,\pi_i,\pi_\nu,\pi_\nu,\pi_\nu,\pi_i \mid \pi_\nu) = \\
& g(\pi_\nu,\pi_\nu,\pi_\nu,\pi_i,\pi_\nu,\pi_i \mid \pi_\nu) = \tfrac{1}{12} r p_i \\
S:\ & g(\pi_i,\pi_i,\pi_\nu,\pi_\nu,\pi_i,\pi_\nu \mid \pi_\nu) = \\
& g(\pi_i,\pi_i,\pi_\nu,\pi_\nu,\pi_\nu,\pi_i \mid \pi_\nu) = \\
& g(\pi_\nu,\pi_i,\pi_i,\pi_\nu,\pi_i,\pi_\nu \mid \pi_\nu) = \\
& g(\pi_\nu,\pi_i,\pi_i,\pi_\nu,\pi_\nu,\pi_i \mid \pi_\nu) = \\
& g(\pi_\nu,\pi_\nu,\pi_i,\pi_i,\pi_i,\pi_\nu \mid \pi_\nu) = \\
& g(\pi_\nu,\pi_\nu,\pi_i,\pi_i,\pi_\nu,\pi_i \mid \pi_\nu) = \\
& g(\pi_i,\pi_\nu,\pi_\nu,\pi_i,\pi_i,\pi_\nu \mid \pi_\nu) = \\
& g(\pi_i,\pi_\nu,\pi_\nu,\pi_i,\pi_\nu,\pi_i \mid \pi_\nu) = \tfrac{1}{8} s p_i
\end{aligned} \tag{4}$$

where $\nu \neq i$, and $\nu, i = 1, \ldots, k$.

In this way we have obtained a huge reduction in the number of terms in the contextual classification rule.

III. Results

In order to illustrate the power of this algorithm we will apply it to a two class 3-D synthetic dataset. This dataset consists of a $32 \times 32 \times 32$ data volume with one variable at every pixel. The dataset is generated by use of a (morphological) isotropic Potts model [7]. In Figure 2 four slices are shown.

The two classes are assigned mean values -1 and 1. We will consider two cases. First, the case of pure white noise, and second, the case of a mixture of white and autocorrelated noise. In both cases we will compare the contextual classifier with a classical pixelwise linear classifier (e.g. [8]).

For the sake of evaluation the mean values and variances are estimated from 2/3 of the pixels in the data volume picked at random. The classification is then evaluated on the remaining 1/3 of the pixels. We will assume equal prior probabilities for the two classes.

A. Case 1: White noise

In this case we will degrade the dataset with independent, identically distributed Gaussian noise, with standard deviations 1 and 2, respectively. In Figure 3 the degraded slices from Figure 2 are shown.

Fig. 2. Slices 4, 12, 20, 28 of the original data volume.

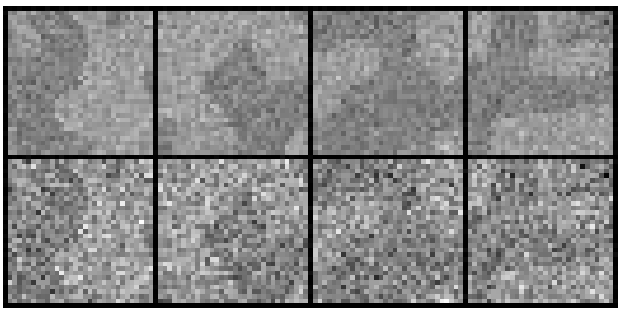


Fig. 3. Slices 4, 12, 20, 28 of the noisy data volumes. The two rows have noise with standard deviation 1 and 2, respectively.

The misclassification rates for the four classifications are shown in Table I. For the non-contextual classifier, if the parameters (μ, $\mathbf{\Sigma}$) where known, the classification rule should be a theshhold at 0, which for the two values of the standard deviation, σ, corresponds to $1 \cdot \sigma$ and $0.5 \cdot \sigma$. Assuming normality, this should result in misclassification rates of 15.866% and 30.854%, respectively. The obtained results agree well with this. When compared with the contextual classifier, we see that the inclusion of spatial information results in misclassification rates that for $\sigma = 1$ is a factor 8 lower and for $\sigma = 2$ is a factor 2.5 lower.

B. Case 2: Autocorrelated and white noise

In this case we will degrade the dataset with independent, identically distributed Gaussian noise mixed with autocorrelated noise. The white noise and the autocorrelated noise have the same variance, and we will use autocorrelated noise with an exponentially decaying autocorrelation. The autocorrelation in lag 1 is 0.6. Again we will apply the algorithms to two cases with a pixelwise standard deviations 1 and 2, respectively. In Figure 5 the degraded slices from Figure 2 are shown.

The misclassification rates for the four classifications are shown in Table I. Again, we see good agreement between the misclassification rates for the non-contextual classifications and the theoretical rates derived in the previous Section. With respect to the contextual method we see that the improvement over the non-contextual method now is reduced to a factor 2.2 and 1.4, respectively.

IV. Conclusion

We have described an extension of a 2-D contextual classification algorithm by Owen, Hjort & Mohn to the 3-D case. The algorithm includes contextual information for each pixel by including the feature vector of that pixel as well as the feature vectors of the 6 nearest neighbouring pixels in the decision. A joint Gaussian distribution for these feature vectors given the classes of the pixels has been specified. It is assumed that the noise can be modelled as a sum of white noise and autocorrelated noise, where the autocorrelation function is exponentially decaying with Euclidean distance. Furthermore, a joint prior distribution of the class variables of a pixel and its 6 nearest neighbours has been specified. It is assumed that the pixel size is small relative to the region sizes in the image, thus vastly decreasing the number of possible configurations to in principle four types.

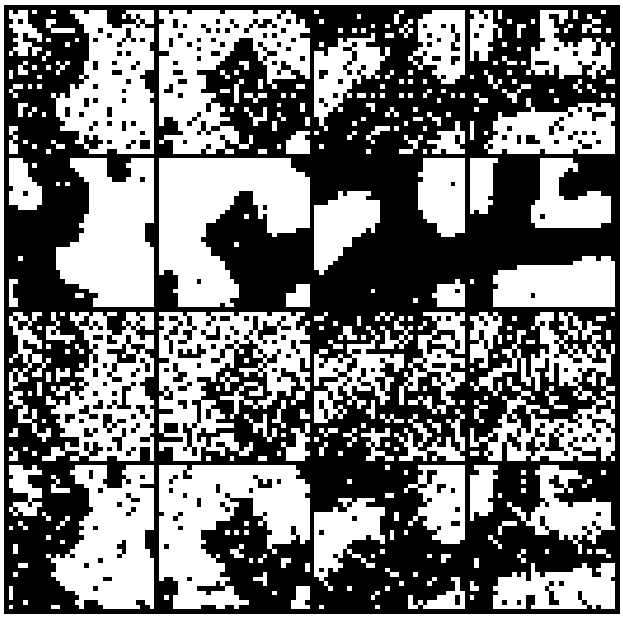


Fig. 4. Slices 4, 12, 20, 28 of the restored data volume. The two top rows are on the 1 std. dev. dataset, using pixelwise and contextual algorithms respectively. The two bottom rows are on the 2 std.dev. dataset using pixelwise and contextual algorithms, respectively.

The algorithm is tested on a synthetic two-class 3-D image. For moderate white noise levels the misclassification rate is a factor 8 lower than the rate obtained using an ordinary linear pixelwise classifier. The relative improvement in misclassification rate decreases with increasing noise level. In the case of a mixture of white and autocorrelated noise the improvement in misclassification rate over the pixelwise method is a factor 2.2 for moderate noise levels.

V. Acknowledgements

Some of the software used in this work was programmed by M. Sc. Anders Rosholm and M. Sc. Jørgen Folm Hansen, De-

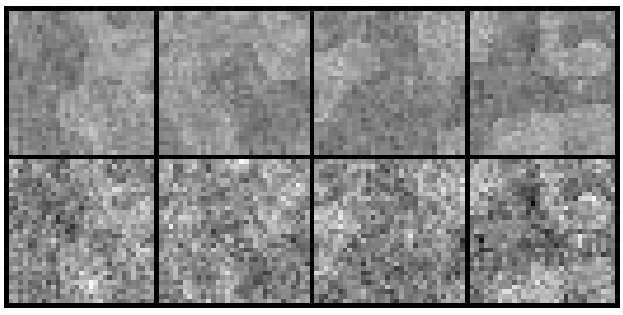


Fig. 5. Slices 4, 12, 20, 28 of the noisy data volumes. The two rows have noise with standard deviation 1 and 2, respectively.

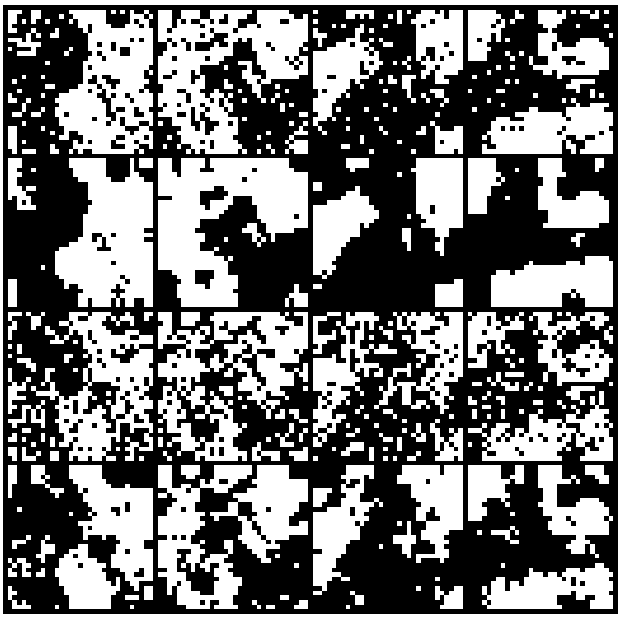


Fig. 6. Slices 4, 12, 20, 28 of the restored data volumes. The two top rows are on the 1 std. dev. dataset, using pixelwise and contextual algorithms respectively. The two bottom rows are on the 2 std.dev. dataset using pixelwise and contextual algorithms, respectively.

TABLE I

MISCLASSIFICATION RATES FOR EACH OF THE COMBINATIONS BETWEEN CLASSIFIER AND NOISE LEVEL.

	White noise		Autocorrelated noise	
	$\sigma = 1$	$\sigma = 2$	$\sigma = 1$	$\sigma = 2$
Non-contextual	16.0	30.8	15.7	30.9
Contextual	2.0	12.3	7.3	22.2

partment of Mathematical Modelling, Technical University of Denmark. This and their valuable comments on the work reported here is hereby greatfully acknowledged.

REFERENCES

[1] Paul Switzer, "Extension of discriminant analysis for statistical classification of remotely sensed satellite imagery," *Journal of the International Association for Mathematical Geology*, vol. 12, pp. 367–376, 1980.

[2] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721–741, 1984.

[3] Julian Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society, Series B*, vol. 48, no. 3, pp. 259–302, 1986.

[4] A. Owen, "A neighbourhood-based classifier for LANDSAT data," *The Canadian Journal of Statistics*, vol. 12, pp. 191–200, 1984.

[5] Niels Lid Hjort, "Estimating parameters in neighbourhood based classifiers for remotely sensed data, using unclassified vectors," in *Contextual classification of remorely sensed data: Statistical methods and development of a system*, H. V. Sæbø, K. Bråten, Niels Lid Hjort, B. Llewellyn, and Erik Mohn, Eds. Norwegian Computing Center, 1985, Technical report No. 768.

[6] John Haslett, "Maximum likelihood discriminant analysis on the plane using a markovian model of spatial context," *Pattern Recognition*, vol. 18, no. 3, pp. 287–296, 1985.

[7] J. M. Carstensen, "Morphological Markov random fields," *Statistics and probability letters*, vol. 20, no. 4, pp. 321–326, 1994.

[8] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, John Wiley, New York, second edition, 1984, 675 pp.