CHAPTER 6

Estimation and Detection

R. Srinivasan (University of Twente)
G.H.L.M. Heideman (University of Twente)

Introduction

The early part of the last century saw the development of the mathematical theories of statistical estimation and detection. Since then, these theories have played an important role in many areas of engineering. They have laid down guiding principles for processing of signals in the areas of communications, radar, sonar, radio astronomy, seismic processing, meteorology, underwater and deep space exploration, and biomedical research. These principles have given rise to powerful algorithms in numerous applications, as evidenced by the highly reliable and sophisticated processing systems that are in use today. The applications are too many to list here. However, a common conceptual thread that links them all is the extraction of information from signals that are inherently stochastic in nature.

Bayesian reasoning and the *principle of maximum likelihood* (ML) are the classic paradigms of statistical estimation and decision theory. The development of optimal signal detection techniques and the associated processing algorithms has its roots firmly embedded in statistical decision theory and the testing of hypotheses. In digital communications, for example, optimum statistical signal processing is crucial in order to achieve, or at least to come close to achieving, the benefits of reliable information transfer as promised by the fundamental limit theorems of

[1] This chapter covers references [511] – [561].

information theory. Whereas some of the coding theorems of information theory are predicated on the assumption of maximum likelihood decoding, the ML principle and Bayesian approach have guided the development of optimum estimation and detection structures that achieve minimum probability of error performances in a variety of realistic environments. Another landmark that occurred more than half a century ago is the use of likelihoods (by Woodward, Kotelnikov, and others) in devising optimum methods for target detection in radar systems. At the other end of the applications spectrum these same principles, together with measures of information inspired by Shannon's work, have resulted in estimation and detection techniques for the processing of signals arising from biological phenomena. This has led to the development of powerful systems for the detection and diagnosis of medical anomalies in humans and animals.

Despite the existence of an immense literature on estimation and detection as distinct areas of research, their roles are usually hard to delineate in the operation of any real processing system. Nevertheless, in this chapter, we have attempted to categorize papers on the two topics in separate sections, notwithstanding the close interrelationships that exist in some cases. An attempt has also been made, as far as possible, to provide a commentary on these WIC contributions while keeping information theoretic considerations in mind. The papers have roughly been grouped into three categories: estimation, detection, and pattern recognition and classification. The few papers that fall outside this categorization but nevertheless fall within the general purview of the aim of this chapter have been treated separately at the end.

6.1 Information Theoretic Measures in Estimation

Several theoretical and application oriented papers on estimation are described in this section.

6.1.1 Time Delay Estimation

The use of entropy and mutual information measures have produced several results in estimation applications. An important application has been the analysis of electroencephalogram (EEG) signals in animal and human brains for understanding the mechanisms that cause epileptic seizures. Several results in this area, which are due to Moddemeijer, are described herein. Estimation of time delays between recordings of EEG signals from different channels is a principal approach for analysis of these signals.

Several methods are in use for time-delay estimation. The cross-correlation and mutual information methods search for the maximum correspondence of pairs of samples $(\underline{X}(t), \underline{Y}(t+\tau))$ as a function of the time shift τ, disregarding the dependence of subsequent sample pairs. Other well-known methods are maximum likelihood delay estimation, see Knapp and Carter [38], and those that employ autoregressive moving average (ARMA) modeling (cf. Section 6.1.2). In addition to these, there is a large number of phase measurement methods defined in the frequency domain which use the same signal model as that in [38].

The connection between time-delay estimation and mutual information and entropies (and therefore probability density functions) is relatively easy to illustrate. The time shift τ that maximizes the mutual information between the X and Y signals is considered to be a good estimate of the delay between the two signals. As is well known, mutual information can be expressed as a function of individual and joint entropies. Estimation of these information measures therefore requires knowledge (or at least estimates) of underlying density functions. Consequently, estimation of joint density functions has been the subject of many research efforts, and several methods have been developed.

In [524], a histogram method is presented for estimating a two-dimensional continuous probability distribution, from which estimates of entropy and mutual information are obtained. Using bias correction and variance estimation, results at least as good as those reported for other estimation techniques have been obtained.

In [529], an attempt at developing a unifying concept underlying the different methods of time-delay estimation mentioned above is discussed. It resulted in the proposed maximum average log-likelihood (MALL) method. The concept is based on (a generalization of) defining an average log-likelihood function and using it as an estimate of the mean log-likelihood (MLL). Then a search is carried out for a parameter vector which maximizes this average. The maximum thus obtained, or MALL, is then considered to be an estimate of the negative entropy, where the latter is well approximated by the maximum of the MLL. This leads to an estimate for an unknown probability density function that can be used in time-delay estimation. The different biases of this procedure are related to the histogram-based estimators proposed in [524]. Jumping ahead to [556], Moddemeijer studies the probability distribution of the MALL statistic. He shows that, under certain conditions, the distribution of the MALL is a sum of independent contributions. In particular, in the asymptotic situation of a large number of observations, it is obtained as the sum of a normal distributed component and a χ^2 distributed component. These findings indeed provide theoretical justification for the assumptions made by the author in his earlier results ([552] and [554]) on AR order estimation based on hypotheses testing. The latter are described in the sequel.

An interesting further result due to Moddemeijer is an information theoretic time-delay estimator [531]. The proposed method is model-free and non-parametric, and sets up a measure of mutual information between processes to define time delay. Two stochastic processes are considered, where one process is a sample sequence shifted j samples in the future. Each process is partitioned into two parts: an infinite sample sequence representing the past and one representing the future. The past vectors of both processes are concatenated into one past vector, and the same is done for the future vectors. A mutual information measure is set up between the joint past and joint future by considering both original processes to be of length $2M$ and then allowing $M \to \infty$. It is shown that for stationary processes and under certain convergence conditions, this mutual information possesses a unique minimum with respect to the time shift j. This minimizing value of j is then defined as the information theoretic time delay between the two processes. The interpretation is that for this specific time shift, there exists a joint process with a minimum transport of information between the past and future. The minimum mutual information method proposed herein is discussed in comparison with other methods. It is shown for example that this method is, to an approximation, a generalization of the maximum likelihood method. For exposition of this estimator, normally distributed sequences are considered. It is demonstrated that the mutual information can be calculated by operations on the determinants of estimated covariance matrices of the processes. Numerical results are promising.

6.1.2 Autoregressive Processes

The modeling of time series data using autoregressive (AR), moving average (MA), or mixed ARMA processes has long been a powerful approach for characterizing various kinds of signals arising in practice. These are signal models which are driven, usually, by stationary uncorrelated Gaussian sequences of known or unknown variance. Such models lend themselves well to estimation activities, especially for methods based on Kalman and least- squares filtering and prediction. Multichannel ARMA processes are closely related to the state-space models arising in Kalman-Bucy filtering. This is a reason for their importance in the statistical analysis of speech, biomedical signals, weather data, and a host of other applications. We remind the reader that a scalar (single-channel) stationary ARMA process $\{x_n\}$ has a model that can be written as

$$x_n = \varepsilon_n - \sum_{i=1}^{m} a_i x_{n-i} + \sum_{i=1}^{p} b_i \varepsilon_{n-i}. \tag{6.1}$$

It is a model driven by the stationary white Gaussian noise sequence $\{\varepsilon_n\}$ with variance σ^2 and the model may include initial conditions. The parameters a_i and b_i denote the AR and MA parameters, respectively. Together with σ^2, they represent the model parameters in an application. It is usual to refer to the process as an ARMA(m, p) sequence with AR order m and MA order p.

In practice, choosing a model, determining model order, and estimating parameters within the model are real problems to be solved. The decision to model a process by ARMA, AR, or MA models usually depends on some prior information regarding the physics of the phenomenon under study. The second two estimation tasks are handled by well-known powerful methods. For example, the model order can be determined using Akaike's information criterion (AIC), final prediction error (FPE), or the minimum description length (MDL) information theoretic criterion, with parameter estimation based on ML or on least squares methods.

In [523], Liefhebber describes the *minimum information* approach for model selection and order determination. It is in fact an application of the *principle of maximum entropy*, a formalism based on statistical estimation and information theoretic considerations that arose almost 40 years ago. The minimum information approach to model identification involves the use of a normalized power spectrum (as a spectral density function) to define a spectral entropy and then maximizing this entropy subject to a set of constraints on the correlation coefficients estimated from a finite realization of a discrete random process with continuous power spectrum obtained as observed data. Such a procedure is considered to provide a process model which is least presumptive or minimally prejudiced to the observations. The result is a parametric model for the power spectrum as a representation of the observed data. By means of spectral factorization, an equivalent time-domain model is obtained. It is finally shown that an *a priori* choice for an AR, MA, or ARMA model for the observed data is violated if the minimum information principle is imposed on the data. In the first two cases, applying the principle leads to increased-order *a posteriori* representations for the data, whereas the ARMA case leads to a nonparametric representation. The author recommends further investigations into this problem.

Using the ARMA model approach, Moddemeijer presents in [527] a slightly different method for order determination than conventional ARMA estimation. EEG signal models typically involve a large number of parameters. While the Akaike criterion is used to select the optimal model, the parameter space of the ARMA model signal is split into two parts, containing active and inactive parameters. Optimization of an appropriate cost function is then carried out with respect to the active parameters. Application of this approach using numerical examples indicates somewhat better results when compared with the conventional method.

Continuing this line of research in [554], Moddemeijer uses a distinction between the correct or true AR model and an optimal model to present an algorithm for model identification. These two models differ in the following way. If in the correct AR model a parameter is small, then it is neglected or set equal to zero in the optimal model. This is carried out for all the parameters. Such a procedure sacrifices flexibility but reduces the variance by allowing some bias to enter into the estimation. In practice, neither the AR order nor the number of negligible parameters is known *a priori*. An algorithm to estimate the configuration of significant parameters is proposed based on the ARMA estimation algorithm studied in the preceding paragraph combined with an AR order estimation procedure using a *modified information criterion* suggested by the same author. An AR model order and values of the nonzero coefficients of the model are first estimated. This model has a parameter vector consisting of independently adjustable parameters. Fixing some of these parameters to zero leads to a reduced dimension for the parameter vector. Models with different configurations (or parameter vectors) are treated as multiple hypotheses. Then the optimal configuration is selected via hypotheses testing based on an *a-priori* specified value of false alarm probability of selecting an excessively high order. The hypotheses testing aspects ([552]) are dealt with in Section 6.2.4 for papers written by Moddemeijer. Using examples, the author shows that the method performs satisfactorily.

6.1.3 Miscellany

In [512], Boel addresses the question of estimating the intensity of a Poisson process. An explicit, recursive, optimal estimator is sought. Boel shows that the solution is a stochastic linear partial differential equation with the observed Poisson process as input. In an example, it is assumed that the intensity is the square of an Ornstein-Uhlenbeck process, which is related to models for optical communications and communication networks.

In [514], Kwakernaak proposes an algorithm for the fundamentally important problem of estimating arrival times and heights of pulses of known shape in the presence of additive white noise. In the realistic situation of an unknown number of pulses, maximum likelihood procedures encounter the same difficulties as for order estimation of an unknown system. He proposes a solution for this based on Rissanen's *shortest data description* criterion (equivalent to the MDL mentioned in Section 6.1.2) and establishes consistency of the estimation algorithm. An example from seismic data processing serves to illustrate the algorithm.

The mathematical paper by Berlinet, Györfi and Van der Meulen [548] concerns the ever important problem of estimating the quality of density estimators. In particular, the Kullback-Leibler number or information divergence of two densities is used. They study a histogram-based density estimator proposed by Barron in [72] and a related distribution estimator proposed by Barron, Györfi and Van der Meulen in [87]. In the latter paper, the authors established sufficient conditions for consistency, based on information divergence, of the histogram density estimator. In the present paper ([548]), a limit law is derived for the centered information divergence of the same estimator. The centered divergence is defined as the random part of the information divergence. It is shown that a suitably normalized form of the centered information divergence is asymptotically normal with asymptotic variance less than or equal to unity. They show that the centered divergence is smaller (asymptotically) than the non-random part of the information divergence, the latter representing the expected global error in estimation. The result therefore strengthens the proposed density estimation procedure.

6.2 Detection Theory and Applications

In this section we attempt to describe the work carried out in detection. The topics dealt with are diverse, ranging from abstract concepts through typical signal detection problems in communications to biomedical applications.

6.2.1 Change Detection

Jump or change detection (also called the change-point problem) has been studied by several researchers because of its importance in many applications. A rather large body of literature exists on various aspects of this problem. Applications of jump detection are in image processing, oil exploration, underwater signal processing, radar tracking of maneuvering targets, and in many more areas. The basic problem is one of detecting a sudden jump in a noisy signal. The size of the jump may be known or unknown. The so-called “quickest detection” problem can also be considered as a case of change detection. It is one of detecting the change in the shortest time possible.

Much is known about optimal methods for detecting jumps in random signals when the size of the jump is known. Relatively less is known about how to deal with the general case of unknown jump size. In the latter case, the problem naturally becomes one of simultaneous detection and estimation. This is the subject of the paper by Vellekoop [558]. A brief background on this problem is useful. The setting is one wherein the noise is additive and white Gaussian. It has been established that for a known jump size in the stochastic signal, the optimum detection rule produces an alarm whenever the conditional probability that a jump has occurred exceeds a certain threshold. This conditional probability can be determined in terms of a likelihood ratio. This is referred to as a Shirayev detector [44]. On the other hand, when the time of occurrence of the jump is known, the solution to the estimation problem is just the Kalman filter. The Kalman filter of course is optimal if the signal has a Gaussian distribution. The general case where both jump size and time of occurrence are unknown is much harder. In the present paper, Vellekoop proposes an algorithm which projects the nonlinear filtering Zakai equation on a statistical manifold using the Kullback-Leibler information criterion. This results in a structure which is a mixture of the Shirayev detector and the Kalman filter. The equations provide estimates of the conditional probability that a jump has occurred and size of the jump. The paper then establishes convergence properties of the filtering algorithm.

In the two papers [547] and [550], written before the one by Vellekoop discussed just above, Hupkens studies the problem of quickest detection of changes in random fields. The classical quickest detection problem, as solved by Shirayev, is defined for unidirectional stochastic processes, i.e. those that evolve in time. The solution is specified in terms of a stopping rule given by a generalized sequential probability ratio test. If the signal under study is a random field, this causality is no longer available. The change may be present at any arbitrary site of the field from which measurements are taken. Examples of such a situation arise in several spatial search applications. In his first paper [547], Hupkens develops a mathematical formulation of this problem. He demonstrates that in its full generality, the change detection problem for random fields is difficult to solve. Assuming that the prior distribution of changes is known and making some simple assumptions on a cost function, he approaches the problem from a Bayesian viewpoint in his second paper [550]. Thus a Bayes cost is set up, and a Bayes stopping strategy that minimizes the cost is the required solution. Even here it is shown that the problem cannot be solved explicitly without making further restrictions. For cases where change detection can be modeled as a simple hypotheses testing problem, the author obtains an approximate solution, and he provides numerical results which match well with the exact solutions for some simple cases.

6.2.2 Biomedical Applications

An early paper on transient detection in EEG signals is the one by Kemp [518]. A simple model describes the EEG signal as observations of a known amplitude modulated signal in additive white Gaussian noise. The author makes use of Ito's differentiation rule and a filter result of Wonham. Using a martingale representation of the amplitude modulated transient, he derives an optimal estimator-detector structure for sleep states. The relationship between the estimation and detection operations is examined.

The detection of brain state during sleep using EEG observations is the subject of the paper by Kemp and Jaspers [521]. Here, brain state is modeled as a 4-state Markov process. Using a feedback loop driven by white noise with the Markov process as a modulating signal, they adopt a generator model for the EEG signal. Then martingale theory is used to derive filtered estimates of the state. Optimal state decisions are then obtained by minimizing the average cost in the usual Bayes cost formulation employing uniform costs. It is shown that the resulting detection rule is easy to implement and that extension to a larger number of states is straightforward.

In a further attempt toward developing automated sleep stage monitoring systems, Kemp in [528] proposes a model for the occurrence of bursts of rapid eye movements (REMs). Various stages of human sleep produce different eye and body movements. REMs occur irregularly, but exclusively during waking or during a sleep stage called REM-sleep. In this paper, REM bursts are modeled as stochastic processes simulated by a Poisson counting process with a rate that depends on a binary Markov sleep state. Using this model, a stochastic differential equation driven by a martingale process results, and this describes the REM burst counting process. The likelihood ratio for the problem of testing whether or not the observations belong to a REM state is set up. The detection problem is then investigated using a Bayes optimal threshold, the latter being obtained by simplifying the Poisson rate to be one of two constant values. The rates are the reciprocal of the average sojourn times in each state (REM and non-REM), experimentally observed, and their ratio forms the test threshold. The structure of this minimum probability of error detector is derived, and the required processing is revealed. Although performance results have not been presented, the author feels that better detectors can be obtained using these methods.

More recent research on the analysis of EEG recordings is contained in the paper by Cremer and Veelenturf [549]. The problem investigated is that of spike-wave detection, an application somewhat different from the one mentioned in the preceding paragraph. Spike waves are randomly occurring waveforms sometimes present in EEG signals, and they usually mark the start of an epileptic seizure. They are difficult to characterize mathematically, as they have very different shapes and durations. Detection of such phenomena is therefore only possible by learning from examples. This is the motivation for the authors to use neural networks, in particular Kohonen's neural network. Using single-channel EEG data, they implement and compare 6 different detection methods. Three of these use a variant of the Kohonen network. The conventional detection methods used are correlation detection, parametric, and non-parametric density estimation for determining likelihood functions. The neural based methods (combined with statistical signal detection) are non-parametric and semi-parametric density estimation, and parametric signal detection. The conclusion is that parametric signal detection combined with a neural network gives the best trade-off between the number of calculations required and the occurrence of false alarms.

6.2.3 Communications

Bergmans [525] presents a clear and concise description of the principal operations of equalization, detection, and channel coding in a digital transmission system. This is done with the motivation of comparing the three operations with respect to their respective abilities to combat intersymbol interference (ISI), noise, and channel fluctuations. A comparison is made between the signal-to-noise ratio improvements, implementation complexities, and adaptivity. Equalizer types discussed are the linear, decision feedback, and ISI cancelers using feedback and feedforward filters. As an alternative for combatting ISI, Viterbi detection is considered. Finally, he considers channel coding for protection against noise and burst errors. As is well known now, the study concludes that channel coding has the highest complexity, but also is most effective in dealing with channel variations. Based on complexity, the ISI canceller is found to be preferable to the Viterbi detector.

In [557], Levendovszky, Kovács, Jeney and Van der Meulen address the well-known problem of developing low-complexity alternatives to maximum likelihood multiuser detection (MUD) for direct sequence code division multiple access signals. In this work, the authors employ a neural network to perform blind MUD, where channel characteristics are not known and no training sequences are used. The network used is a stochastic Hopfield net. A decorrelating algorithm is suggested that performs inverse channel identification and which can combat multiuser and intersymbol interference. Mean-square convergence of the algorithm is established and performance evaluation of the system by simulation demonstrates “near optimal” MUD detection performance.

6.2.4 Autoregressive Processes

Moddemeijer and Gröneveld address a composite hypotheses testing problem in [537]. Although not directly on AR processes, the problem discussed here has a close bearing on AR order estimation, as described in a following paper. It deals with estimation of parameters of the density function of an observed random vector. The problem is posed as one of hypotheses testing wherein one probability density function is to be selected from a set of hypothesized density functions. In this paper the set is restricted to two density functions, each containing a vector of parameters that are unknown. Thus it constitutes a composite hypotheses testing problem. Consequently, a generalized likelihood ratio test is proposed as a solution. As in [529] discussed in Section 6.1.1, the average log-likelihood is used as an estimate of the mean or expected log-likelihood and a maximization of the former is sought with respect to the unknown parameter vector. A test is derived and an improved test is suggested that compensates for the bias introduced by the approximation of the MLL.

In [552], Moddemeijer provides a solution to the problem of AR model order estimation based on composite hypotheses testing. The AIC is used as a test statistic, with the maximum of the MLL replaced by the MALL. Convergence properties of the MALL are analyzed. A modification of the test in the framework of the Neyman-Pearson criterion is suggested. Simulations carried out by the author indicate excellent match with theory.

6.2.5 Biometrics

There are two interesting papers on this subject in these proceedings: [560] and [561], which address problems in biometrics using concepts of optimal hypotheses testing. Briefly, biometric verification attempts to confirm the identity of a user based on a biometric signature data (or feature vector) provided by the user. The process typically uses stored templates obtained from a large number of users. Quite akin to signal detection, such problems are modeled well in the framework of hypotheses testing. In [560], Veldhuis, Bazen and Boersma formulate a certain multi-user verification problem. It is assumed that each of the (uncountable) multiple users can be characterized by a feature vector possessing a probability density function. A likelihood ratio test is set up for a user and its performance, in terms of a threshold and false-acceptance and false-rejection rates. By averaging over the distribution of the feature vector, an optimization problem is solved to determine optimal threshold settings. They show that the overall false-rejection rate is minimized if thresholds for all users are set to the same value. Using, as they say, an exotic example, the authors proceed to illustrate their formulation by obtaining performance curves. The example involves using signals resulting from tapped rhythms as biometric features.

In [561], Goseling, Akkermans and Baggen look at the verification problem using a somewhat different hypotheses testing formulation. A noisy version of the biometric feature of a user is available. A noisy version of another biometric feature is presented, and it has to be decided whether this new feature belongs to the first user or to a new one. Employing Gaussian distribution models for the underlying processes, the authors set up a likelihood ratio test solution. The structure of the test is examined in detail and compared with standard solutions available in the detection theory literature. A conclusion from the analysis is that the optimal decision rule is not equivalent to a situation where the reference feature can be assumed to be noiseless and adding an extra noise source to the new measurement.

6.2.6 Miscellany

The paper by Van Schuppen [515], addresses some problems in estimation and detection. It was published as a short abstract in the WIC proceedings. The topics covered here include Markov processes, stochastic filtering, Kalman-Bucy filters, detection algorithms, false alarm probabilities, and Chernoff bounds.

Gröneveld and Kleima examine m-fold detection in a general setting in [519]. They show that each optimal detector uses a partition of the $(m-1)$-dimensional simplex of the likelihood ratios in convex regions. The proof is based on optimality criteria that do not use prior distributions and loss functions. A converse is also shown wherein every partition represents an optimal detector. It turns out that selecting an optimum detector implies always selecting a Bayes detector which in turn implies certain priors and loss function.

In [540], Vanroose addresses the well-known *NP*-complete problem of constructing optimal binary decision trees and test algorithms for the identification of objects. With a simple example, he points out the deficiencies of various heuristically proposed cost functions that have been used for designing test algorithms. The author then introduces the aspect of reliability by assigning probability distributions to the important features of the objects to be identified. This is incorporated into the cost function and an unreliability measure is set up and interpreted as a conditional entropy. A test procedure based on evaluation of such a measure is then proposed as a more reliable method.

6.3 Pattern Recognition

In this section we describe papers that deal with the subjects of classification and pattern recognition, including the use of neural networks in applications.

6.3.1 Neural Networks

The brain is the most advanced information processing machine, and therefore it should be of much interest to information theorists to know how neural networks can mimic some properties of the brain. At least there is some hope that neural networks do so. It is somewhat surprising that neural networks received so little attention in the WIC community. From the ten papers that are devoted to neural networks in the past 25 years, half of them appeared in the proceedings of 1989. The other half is distributed over the next ten years.

In 1989, a lot was known about different types of neural networks: multi-layer networks, Kohonen networks, Hopfield networks, and so on. Therefore, most of the papers are concerned with learning algorithms, i.e., Hebbian rules, stability and convergence problems, and applications of neural networks in different classification and estimation applications.

A popular learning algorithm is the back-propagation learning algorithm for multilayer feedforward networks. In order to effect learning, one has to determine the weights of the connections between neurons of different layers. To do so, we need an error function. This may be a nonlinear function of the state of the output layers. Usually the gradient descent method is used.

One problem with the back-propagation algorithm is the slow convergence in some cases. De Wilde suggests in [532] to use the Marquardt algorithm. This method is a hybrid between the gradient descent and the Gauss-Newton methods. He shows that the Marquardt algorithm can be used for online learning in a similar way as gradient descent.

The article of Piret [534] is devoted to the analysis of a class of Hopfield associative memories. It analyzes a modification of the common Hebbian rule. An application of a neural network with Hebbian learning and with transmission delays can be found in the paper of Coolen and Kuijk [533]. They show that such a system will automatically perform variant pattern recognition for a one-parameter transformation group. Such a network needs a learning phase in which static objects are presented as well as objects that continuously undergo small transformations. The system does not need any *a-priori* knowledge of the transformation group itself. It learns from the information contained in the “moving” input and creates its internal representation of the transformation.

In [536] Vandenberghe and Vandewalle also mention the central problem in the use of neural networks for pattern recognition and image and signal processing, i.e., the development of training and learning algorithms. The authors discuss a number of dynamic properties of neural networks and indicate how these considerations can lead to improvements. They realize that specifications on the behavior of neural networks can generally be written as linear equations with unknown coefficients. They suggest that a systematic approach to derive adaptive training algorithms should consist of applying classical relaxation methods of solving sets of linear inequalities. They demonstrate their ideas with a design of a neural network that should recognize characters $(0, 1, ..., 9)$ as images of 15×20 pixel size and for edge detection and noise removal.

An important property of neural network design and analysis is the robustness of the construction in the presence of possible weight errors. The paper of Levendovszky, Mommaerts and Van der Meulen [544] determines some basic properties of neural networks, i.e., the convergence speed and tolerated level of inaccuracy in the implementation of the weight matrix. This kind of network qualification is suitable for engineering design in terms of computing these properties in advance. Tolerance analysis is of particular interest for both feedforward and Hopfield neural networks. The authors compute the basic properties of the nets from the weight matrix and assess the minimum tolerated weight error. In carrying out the tolerance analysis on Hopfield nets, a statistical evaluation of the network can be performed, providing statistical bounds for the convergence speed and the tolerated level of inaccuracy.

It is often said that neural networks, specifically multilayer feedforward networks, can outperform other statistical techniques because they do not estimate parameters of the classes to be distinguished, but directly “learn” the class-separating hyperplanes. Multilayer feedforward networks can approximate any class-separating function arbitrarily well, provided that enough neurons are available. In [543], De Bruin raises the question: how do multilayer feedforward networks perform the mapping? Therefore he carries out an experiment with a feedforward neural net with one hidden layer, containing 5 neurons. He concludes that the neural classifier does not simply make decisions on features in the first layer which are then combined in the second layer. His conclusion is that the idea that neural network class-separating hyperplanes are built up from parts of hyperplanes defined by hidden-layer neurons may not be correct.

Most of the results on neural networks are obtained by simulations on conventional computers. However, some advantages of neural networks are lost during simulation: speed, parallelism, fault tolerance. Dedicated VLSI processors can make networks more interesting than conventional computers. In [535], Verleysen, Martin and Jespers present a VLSI architecture for a Hopfield-like fully interconnected network with capacitors as synaptic interconnections instead of resistors or current sources. However, the connection weights are restricted to some discrete values. This type of architecture offers several advantages: the accuracy that can be reached with capacitors is increased, and the number of synapses that can be connected to the same neuron is greater. Also only the relative values of the capacitors are important; their size can be reduced to very small values. An 8-neuron network with discrete components has been realized.

A specific application of a two-layer network is proposed in [546] by Levendovszky, Van der Meulen and Poszyai for estimating the tail of aggregate traffic emitted by users of ATM networks for Call Admission Control (CAC). The authors interpret CAC as a set-separation problem. A traffic configuration is admitted or not. Learning can be regarded as a search in the parameter space to find the best point which minimizes the number of lost calls. They also compare the results. The neural network yields best approximation of the original admitted region (the number of lost calls is much lower than obtained by the Chernoff bound and also much lower than that obtained by the Hoeffding inequality).

In further work on the same application, Levendovsky, Meszaros and Van der Meulen [553] propose and evaluate various neural based learning algorithms for classification. This is done with the aim of implementing fast CAC in multi-access systems. Using non-uniform costs for the two kinds of errors, the authors study directed gradient and penalty function methods for performing classification. Based on comparisons made via numerical simulation, it is concluded that penalty function classifiers have a higher learning speed at the cost of a slight decrease in performance.

6.3.2 Classification and Expert Systems

The papers that appear here are diverse. They treat classification with and without teachers, data analysis, expert systems, and so on. We have dealt with them in a chronological order.

The first paper [511], by Backer, written in Dutch, is about minimal distortion relations in classification without a teacher. In this paper, special attention is given to the treatment of the minimal distortion criterion. The special attention to this important consideration provides insights into fuzzy relations that can lead to more sophisticated models. The author shows that decomposition of fuzzy relations can lead to new essentials in hierarchical classification.

In [513], Duin provides a discussion of the need for using *a-priori* knowledge in developing a pattern recognition system. Various possibilities and difficulties are treated. Special attention is given to a comparison of statistical and structural approaches. Also, the use of fuzzy concepts is discussed in various ways; fuzzy labeling, fuzzy relations, fuzzy classification, etc. It appears that the use of a fuzzy labeled learning set puts higher demands on the teacher and the features used than a hard labeled set does. The use of a fuzzy intermediate classifier improves the possibilities of a multistage classifier.

From the same author there is the paper [517] about small sample size considerations in discriminant analysis. This paper discusses a practical rule for avoiding the peaking phenomenon in discriminant analysis. This phenomenon is: the classification error made by a discriminant function based on a finite set of learning objects increases if the number of features used for representing the objects has been increased far enough. The conclusion is that the addition of new features should be stopped before the number of learning objects per point are in the order of one.

After a period of silence, the paper [526] by Backer and Eijlers was published. It describes an attempt to develop a knowledge base (CLUSAN1) for the expert system DELFI2. It should help the user to obtain validated results of an explorative data analysis. The resulting system appears to be particularly suitable for potential users which are non-experts but familiar with the subject matter. The art of knowledge engineering and the resulting structure of the knowledge base are reviewed.

Backer, Van der Lubbe and Krijgsman treat the modeling of uncertainty and inexactness in expert systems in [530]. The problem is that it is very difficult to represent uncertainty, inexactness, and belief that may be attached to expert opinions, judgments and solutions in a rigorous mathematical way. A proposition may be uncertain or inexact or may have a degree of belief, the degree of which can be represented by probabilities, possibilities, fuzzy sets and belief functions which when used in a particular calculus will yield an inexact reasoning. This paper attempts to put the major calculi into perspective as far as their functioning and performance related to mathematical assumptions are concerned.

The article [538] by Kleihorst and Hoeks is concerned with optical pattern recognition. The subject is identification of machine-printed characters in the electronic representation of an image, acquired by a camera or a scanner. The idea is that parts of characters can be detected with template matching. Detection of a part may be indicated by a connected cluster of pixels, called blobs. An automatic learning system constructs a list of “best” blobs, which were detected when the templates were applied to the example character images. The quality measure for blobs is based on techniques from fuzzy set theory. It involves reliability, support, and fuzziness (fuzzy entropy) of the detection blobs and the discriminative power of the template. For a limited set of input characters, the proposed system can recognize characters at high speed with a false recognition rate of 3.5%. An improvement may be reached with a larger description, though such modifications may cause some missed characters.

Design principles and some features of EDAPLUS (Exploratory Data Analysis) are presented by Backer in [539]. Exploratory data analysis is characterized by multiple statistical testing, validations, and complex reasoning. Quite a number of statistical procedures have to be applied in order to understand the peculiarities of the data at hand. Such a reasoning process is associated with knowledge-based systems. There is a need for more intelligence in statistical software packages. As such, EDAPLUS is designed as a knowledge-based software package for cluster analysis. The author describes the decision network in terms of clustering tendency based upon low-level, intermediate-level, and high-level rules. An application in the domain of signal analysis is included.

Hierarchical cluster analysis is a widely used method to represent a finite number of objects in the form of a tree or dendrogram. The paper by Lankhorst and Moddemeijer [542], presents a novel approach to the automatic categorization of words from raw data. The authors count occurrences of word pairs in text and use a hierarchical clustering technique on the frequency data to obtain a classification of words into linguistic categories. The loss of mutual information, caused by combining two clusters in a single new cluster, is used as a criterion in the clustering process. Using this method, words are not only classified on the basis of their syntactic categories, but also with respect to aspects that are related to their meaning. They suggest that this method can form the basis of a system that uses a much finer categorization of words than is feasible using traditional grammar-based approaches.

Another contribution to pattern classification is treated in [545] by Vanroose, Van Gool and Oosterlinck. In this paper, the authors propose BUCA (a bottom up classification algorithm) as a general-purpose supervised learning algorithm based on the average splitting entropy concept. The classification tree is built starting from the leaves, as opposed to other classical methods. BUCA can be applied to any training set which includes class information. BUCA differs from top-down classification systems in two aspects. It recursively joins two training data subsets into a new set in a way similar to the well-known Huffman source coding algorithm, maximizing the joint dissimilarity of the two subsets with respect to the rest of the training set. Dissimilarity of the two classes is defined to be the average splitting entropy, i.e., the average log-probability of a feature value belonging to one subclass, which will be classified into another subclass erroneously. It sometimes outperforms classical classifiers, both in terms of correct classification rate and in execution time.

6.4 Miscellaneous Topics

The paper [516] of Veelenturf belongs to the subject of automata theory. He considers the adaptive identification of sequential machines. It is known that an n-state discrete-time sequential machine can be identified if the set of all input-output sequences of length $2n - 2$ is given. Algorithms that do this are complex. Performing identification using a smaller set is difficult. The author suggests an adaptive procedure which constructs a sequential machine stage by stage. The steps are described in detail and the algorithm is shown to be of reduced complexity.

In [520], written in Dutch, Schripsema and Veelenturf study Petri-networks as a representation of learning behavior. They conclude that Petri networks can be used to simulate learning behavior, but are inefficient for specific applications of learning behavior.

In [551], Slump describes applications in optics from an information theoretic viewpoint, mainly using Gabor's interpretation of information as degrees of freedom of phenomena. With optical image formation as a starting point, it is shown how the wave function characterizing an object can be expanded in terms of the Whittaker-Shannon interpolation (sampling) equation. This is used to determine the number of degrees of freedom. Then radiological imaging is described. For the case where light levels are low, the author shows that noise analysis and detection theory are required. The covariance function of the stochastic image is computed for the example of an X-ray imaging detector. The author states that a spatial information capacity can be defined and computed for such applications.

In [555], Van Someren, Wessels and Reinders tackle the important problem of information extraction from genetic data consisting of high-dimensional signal sets measured at relatively few time points. This task, of inferring gene interactions, is approached by modeling them with a linear genetic network. Advantages of the simplified model adopted include the use of a few network parameters that are easily interpretable, and the possibility of applying constraints without introducing errors in fitting the measured data. Their approach is based on empirical observations that show that genetic networks tend to be sparsely connected. The authors provide a description of the general linear model followed by a procedure to optimize it from the point of view of alleviating the dimensionality problem. In experiments conducted on real data sets, they find computational complexity to be a major obstacle. A clustering procedure is suggested to partially address this issue. In related work, Reinders [559] is concerned with the analysis of genetic data that comprise DNA microarrays. By studying gene expressions (in the enormous amounts of data produced by numerous genome projects worldwide), one can gain a better understanding of gene function, regulation, and interaction in fundamental biological phenomena. The article describes various computational tools used in microarray analysis.