SUMMARY

This thesis is concerned with the application of mathematical-statistical tools for pattern recognition, especially multivariate classification methods (MCMs), to classification problems in analytical chemistry.

Chapter 1.1 is a general introduction and serves two purposes. First, it describes (in section 1.1.1) the background of the research in this thesis. Two practical problems, one from food chemistry and the other from clinical chemistry, are shown to have a similar structure, to be tackled by techniques for multivariate classification. A second purpose of this chapter is to give the other chapters a proper context by reviewing some related work in the literature in connection with our own results. In section 1.1.2 methods for multivariate classification are considered, including a proposal for a new MCM (ELEQDA), related to methods applied amply in later chapters. In section 1.1.3 methods and criteria for the evaluation of the success of classification procedures are reviewed.

Chapter 1.2 describes a preliminary study of the applicability of pattern recognition techniques to the problem of discriminating between wines of different geographic origin. A data set consisting of diverse chemical measurements on samples of red wines from the well-known Bordeaux and Bourgogne regions is described and used to evaluate several pattern recognition techniques available in the software packages ARTHUR, ALLOC and SPSS. It is concluded that a chemical differentiation between the two wine types is feasible. It was useful to limit the dimensionality of the measurements by variable selection procedures.

In the remaining three chapters of Part 1 the attention is focussed on applications in the field of clinical chemistry. In *Chapter 1.3* the usual practice of diagnosing patients on the basis of separate univariate reference intervals for a multitude of blood constituents is respected, and a recent method for obtaining such 'normal' intervals from a population of hospitalized (and therefore often 'non-normal') persons is critically discussed. Reference limits are determined using the database of the Groningen University Hospital.

In *Chapter 1.4* the importance of a multivariate approach to medical diagnosis is stressed. It is shown that multivariate techniques for displaying the data may disclose data characteristics that remain unnoticed in a univariate analysis.

The subject of *Chapter 1.5* is somewhat outside the main line of this thesis. The sets of tests requested from the clinical chemical labaratory by the physicians are analyzed for department-specific patterns. It is shown that missing values in a data matrix with only the results of requested tests are not occurring in a random fashion.

In *Chapters 2.1 and 2.2* two modifications of the in chemometrics well-known classification method SIMCA are proposed. It is shown how posterior probabilities of class membership can be attached to the classification results of SIMCA. The other proposed modification is more fundamental. SIMCA fits principal-component models to the data of each training class, and a different type of modelling is applied in the subspace spanned by the most important eigenvectors (termed by us the inside-model space) and the subspace of the remaining eigenvectors (the outside-model space). A new multivariate classification method is proposed, that treats the outside-model space as in SIMCA, but that uses kernel densities (well-known from the ALLOC method) for modelling in the inside-model space. The resulting method is termed CLASSY and has been implemented in the CLAS program (see Chapter 3.1). In Chapter 2.2 it is compared to other methods using some practical data sets.

Chapter 2.3 introduces stepwise deletion, a new proposal for the preprocessing of data matrices containing randomly occurring missing values. The method aims at maximizing the total number of data values in the final, missing-value-free data matrix by iteratively deleting either rows or columns in the original matrix. It is shown to retain often more of the original data than the deletion of only rows (cases) or only columns (variables).

Chapter 3.1 describes the CLAS computer program for classification and evaluation. In this program several multivariate classification methods and preprocessing methods are implemented, including the proposals from Part 2 of this thesis. The central aim of the program is comparative evaluation of the classification methods using

practical data. To this end the leave-one-out (cross-validation) method of evaluation is made available for all implemented methods, if possible by using computational shortcuts. The program is also able to compute criteria for the evaluation of the posterior probabilities given by probabilistic classification methods. A PC version of the program is described, too.

In *Chapter 3.2* it is shown by simulation studies using the CLAS program, that leave-one-out evaluation is often a good alternative for evaluation using independent test data, not only for error-rate estimation, but also if probabilistic evaluation criteria are considered. The influence of the number of objects, the number of variables, and the inter-class distance on the results of ALLOC classification are investigated in the case of correlation-free multivariate normal distributions. Overconfidence in the assigned probabilities is shown to be a problem in probabilistic classification.

Chapter 3.3 extends the results of the previous chapter by considering more classification methods (LDA, ALLOC, SIMCA, CLASSY) and more data types (simulations with uncorrelated variables, correlated variables, unequal covariance matrices, further also practical data). In the simulations CLASSY is shown to be a robust method, always performing about equally well as the best of the other three investigated methods. From the evaluations with practical data no consistent pattern emerged, suggesting that much of the data structure remains hidden for the human evaluator. A simulation method for choosing the number of principal-component axes to be included in the inside-model space of SIMCA and CLASSY class models is shown to be superior to the simple average-eigenvalue criterion.

In *Chapter 3.4* the attention is on applications of multivariate classification in the clinical chemical laboratory. The possibility of constructing a diagnostic support system using already available data is explored. Practical problems in the process are identified (and partly solved), as well as more fundamental points, such as the common occurrence of multiple diagnoses for one patient. A pilot study on heart, liver, and kidney patients is performed using the CLAS program.

Chapter 3.5 considers the question how much of a practical data set can be deleted without deterioration of the performance of the classification system. A large

clinical chemical data set is use for the discrimination between heart, liver, and kidney patients. It turns out that in this case variable selection is very fruitful even though the classification technique used (CLASSY) applies principal-component models. The number of training patients could be reduced without much loss in classificatory performance. Overconfidence in the assigned posterior probabilities was always a problem, however.

The last two chapters of this thesis, *Chapters 3.6 and 3.7*, are primarily concerned with an evaluation of the POSCON method, which is another recently developed procedure for multivariate probabilistic classification. In Chapter 3.6 this method is applied to wine data, whereas in Chapter 3.7 its performance is evaluated in a clinical chemical context. In both studies POSCON is compared to ALLOC, SIMCA and CLASSY. It is shown that the confidence intervals for the posterior probabilities, which are an unique feature of the POSCON program, are of little value in practical situations where the number of objects is not large compared to the number of measured variables. No large differences in classificatory ability were found between various POSCON models and the other investigated techniques. However, in one case a POSCON model very much related to classical LDA produced reliable (instead of overconfident) posterior probabilities.