A Theoretical Analysis of Probabilistic Fuzzy Systems

Ludo Waltman
E-mail: lwaltman@ieee.org

March 2005

Master Thesis Informatics & Economics
Faculty of Economics
Erasmus University Rotterdam

Supervisors:
Uzay Kaymak and Jan van den Berg
Department of Computer Science

Acknowledgement

I would like to thank the supervisors of this thesis, Uzay Kaymak and Jan van den Berg, with whom I had a number of heated discussions on the subject of probabilistic fuzzy systems. These discussions were both confusing and stimulating. The supervisors of this thesis also had the difficult task of convincing me that I should be satisfied with the results of my work.

Contents

Chapter 1

Introduction

In this introductory chapter, I first give a motivation for probabilistic fuzzy systems (PFSs), which are the subject of study of this thesis. I also discuss the relevance of fuzzy systems in general and PFSs in particular to economic applications. I then formulate the research questions that are considered in the thesis. The chapter is concluded with an overview of the thesis. In this overview, the methodology that is adopted in the different parts of the thesis is also briefly discussed.

1.1 Motivation for probabilistic fuzzy systems

Many statistical models are difficult to interpret and therefore do not provide much insight into the process that is being modeled. This is especially true for the popular neural network models. In comparison with statistical modeling approaches, fuzzy systems have the important advantage that they can be interpreted much more easily. Fuzzy systems therefore result in a better understanding of the process that is being modeled. As a consequence, useful information about a process can be extracted from fuzzy systems and fuzzy systems may be used with higher confidence than black box statistical models. In addition to an improved understanding of a process, another motivation for fuzzy systems is the possibility to define a model using knowledge provided by human experts. The use of expert knowledge may result in an improved accuracy and in lower data requirements.

Due to limited availability of information, there are many problems that involve probabilistic uncertainty. As I will discuss in Subsection 1.2.2, examples of economic problems involving probabilistic uncertainty are target selection and financial markets analysis. The presence of probabilistic uncertainty makes it impossible to obtain a model that always provides correct output. To model probabilistic uncertainty in an appropriate way, one should therefore use a model that provides a probability distribution over all possible outputs. Since ordinary fuzzy systems do not have this ability, these

systems cannot appropriately model probabilistic uncertainty. PFSs, introduced in [16, 18, 34], are an extension of ordinary fuzzy systems. Because PFSs calculate a probability distribution over the output space instead of a single output, these systems allow probabilistic uncertainty to be modeled appropriately. The importance of PFSs therefore is that they are both easily interpretable and capable of modeling probabilistic uncertainty.

1.2 Economic applications of fuzzy systems and probabilistic fuzzy systems

1.2.1 Fuzzy systems

Numerous applications of fuzzy systems to economic problems exist. Some recent examples of such applications can be found in [6, 19, 20, 22, 31, 33]. I will now briefly discuss these examples.

In [6, 19, 31], fuzzy systems are used for predicting financial time series. Stock price prediction using a technical analysis approach is considered in [6]. The authors use a fuzzy system because they believe that fuzzy reasoning corresponds closely with the way humans reason in technical analysis processes. As another advantage, the authors point out that it is easy to understand the knowledge that is contained in a fuzzy system. In [19], neural networks and fuzzy systems are used for predicting currency exchange rates. The authors do not discuss the specific advantages of applying fuzzy systems to this problem. In [31], stock price prediction using fuzzy systems is considered with special emphasis on the issue of combining information provided by human experts and information contained in a data set.

In [33], an early warning system for predicting bank failures is discussed. The authors point out that traditional statistical models for studying bank failures function as black boxes. These models are not able to identify the characteristics of financial distress, which is a very important cause of bank failures. As an alternative, the authors propose to use a fuzzy system for predicting bank failures. Using a fuzzy system, the inherent characteristics of failed banks can be identified. It is interesting to note that the problem of predicting whether a bank will fail or survive involves probabilistic uncertainty. The probabilistic uncertainty is caused by the limited availability of relevant information, which makes it impossible to obtain a model that always provides a correct prediction. Because of the presence of probabilistic uncertainty, bank failures may be modeled in a more appropriate way by using a PFS.

Examples of the use of fuzzy reasoning in decision support systems and expert systems are given in [20, 22]. In [20], an intelligent system for developing a marketing strategy is described. The author argues that the development of a marketing strategy involves a high degree of uncertainty

and ambiguity. He proposes to use fuzzy reasoning for dealing with these factors. In [22], a fuzzy system that supports corporate acquisition processes is described. The authors use fuzzy reasoning because they want to enable users to adapt the system to their requirements.

Outside academic research, fuzzy systems have been applied to numerous real-world business problems. A list of examples is given in [5] (p. 36–43).

1.2.2 Probabilistic fuzzy systems

PFSs have been applied to target selection problems and to financial markets analysis. I will now briefly discuss these applications.

In [15, 17], PFSs are applied to a target selection problem. The problem is to predict whether a customer will respond to a mailing. Since only limited information is available on each customer, it is not possible to predict with certainty whether a customer will respond. The problem therefore involves probabilistic uncertainty. As a consequence of this uncertainty, instead of predicting whether a customer will respond, a more appropriate approach is to predict a customer's probability of response. In [15, 17], PFSs are used for predicting this probability. The target selection problem studied in [15, 17] will also be considered in Section 5.4 of this thesis.

The application of PFSs to financial markets analysis is considered in [18, 34]. The authors use PFSs to obtain linguistic descriptions of the characteristics of financial time series. They study both an artificial time series following a GARCH process and a real-world time series consisting of returns of the Dow Jones index. Both time series exhibit volatility variations. As discussed in Subsection 1.2.1, ordinary fuzzy systems may be used for predicting the future state of a financial market. In [18, 34], however, the authors recognize that predicting the future state of a financial market involves probabilistic uncertainty. Instead of giving a single prediction of a financial market's future state, the authors therefore choose to predict the probabilities of different future states. To accomplish this, the authors make use of PFSs.

1.3 Research questions

In this thesis, I consider two research questions. The first research question is concerned with fuzzy histograms. Fuzzy histograms, introduced in [16, 18, 34], are nonparametric density estimators that can be seen as fuzzy generalizations of ordinary crisp histograms. Because fuzzy histograms are related to PFSs, studying the properties of fuzzy histograms may be expected to result in an improved understanding of PFSs. The following research question concerning fuzzy histograms is considered in this thesis:

1. What are the statistical properties of fuzzy histograms and how do

these properties compare with the properties of other nonparametric density estimators?

This question has not been addressed before in the literature.

The second research question considered in this thesis is concerned with the estimation of the parameters in a PFS:

2. How can the parameters in a PFS be estimated in a way that is, in some sense, optimal?

To answer this question, it is necessary both to choose a criterion of optimality and to provide a method for obtaining parameter estimates that satisfy this criterion. In this thesis, the second research question is split into the following three, partly overlapping subquestions:

2.1. How can the probability parameters in a PFS be estimated in a way that is, in some sense, optimal?

2.2. How can the parameters in a PFS for classification tasks be estimated in a way that is, in some sense, optimal?

2.3. How can the parameters in a PFS for regression tasks be estimated in a way that is, in some sense, optimal?

The first subquestion has also been considered in [16, 18, 26, 34]. In this thesis, a detailed analysis of the approach taken in [16, 18, 34] is provided and an alternative approach is proposed. The second subquestion has also been considered in [1]. The third subquestion has not been addressed before in the literature.

1.4 Overview of the thesis

In addition to this introductory chapter, the thesis consists of six chapters. Chapter 2 is concerned with the first research question. The second research question is addressed in Chapter 3 to 6. Conclusions and issues for future research are discussed in Chapter 7. When reading the thesis, Chapter 2 can be skipped without loss of continuity. Chapter 5 and 6 can be read in arbitrary order. I will now briefly discuss the contents of Chapter 2 to 6. I will also pay some attention to the methodology that is adopted in each chapter.

In Chapter 2, a statistical analysis of fuzzy histograms is presented. The analysis provides insight into the statistical efficiency of different types of fuzzy histograms. The results of the analysis are compared with the results reported in the literature for other nonparametric density estimators. The analysis of fuzzy histograms in Chapter 2 is mainly asymptotic, as is usually the case in statistical analyses of nonparametric density estimators. The

question to what extent the results of the analysis hold for small samples is not considered.

In Chapter 3, PFSs are discussed. Although the discussion is based on the existing literature on this subject [16, 18, 34], there are some important differences and additions. The chapter also contains a proof of the functional equivalence between PFSs for classification tasks and a specific type of radial basis function networks for classification tasks.

Chapter 4 is concerned with the estimation of the probability parameters in a PFS. The probability parameters constitute a subset of all the parameters in a PFS. The statistical properties of an existing method [16, 18, 34] for estimating probability parameters are derived. Because the existing method turns out to have unsatisfactory statistical properties, an alternative method for estimating probability parameters is proposed. This method is based on the criterion of maximum likelihood. The mathematical properties of the optimization problem that results from the proposed method are also analyzed.

In Chapter 5, the issue of estimating the parameters in a PFS for classification tasks is considered. It is proposed to use the criterion of maximum likelihood for estimating both the probability parameters and the antecedent parameters in a PFS for classification tasks. Contrary to Chapter 2 to 4, which are completely theoretical, Chapter 5 also contains some practical experiments. In these experiments, the proposed method for parameter estimation is compared with two heuristic methods that are described in the literature [1, 17]. The heuristic methods make use of fuzzy clustering. One of the experiments in Chapter 5 is concerned with the target selection problem that was discussed in Subsection 1.2.2.

In Chapter 6, probabilistic fuzzy modeling in regression problems is discussed. First, it is argued using a simple example that in many regression problems a PFS with a limited number of rules does not have a satisfactory approximation accuracy. Then, an alternative approach to probabilistic fuzzy modeling in regression problems is proposed. Chapter 6 is quite theoretical, since applications of probabilistic fuzzy modeling to practical regression problems are not considered. The contribution of the chapter is to draw attention to important issues in the application of probabilistic fuzzy modeling to regression problems. The further elaboration of these issues remains for future research.

Chapter 2

Fuzzy Histograms

In this chapter, a statistical analysis of fuzzy histograms (FHs) is presented. FHs are considered from the point of view of nonparametric density estimation. The analysis of FHs is based on a book and some papers written by Scott [27, 28, 29, 30]. In his book, Scott analyses a number of nonparametric density estimators, namely (crisp) histograms, frequency polygons, averaged shifted histograms, and kernel density estimators (KDEs). The analysis of FHs in this chapter is an extension of Scott's analysis of KDEs. Based on the results of the analysis of FHs, FHs are compared to other nonparametric density estimators. It is also shown that in the special case in which there is no fuzziness, the analysis of FHs gives the same results as Scott's analysis of crisp histograms.

It should be noted that FHs are very similar to double-kernel estimators introduced in [36]. However, the analysis of FHs in this chapter differs considerably from the analysis of double-kernel estimators that is given in [36]. Furthermore, in [4] so-called soft histograms are studied. Like FHs, these soft histograms can be seen as an extension of crisp histograms based on ideas from fuzzy set theory. It is, however, important to note that the soft histograms in [4] are not identical to the FHs considered in this chapter. It should also be noted that what is called a FH in [25] is not the same as the FHs in this chapter.

This chapter is organized as follows. In Section 2.1, a statistical analysis of KDEs is given. The results in this section are taken from [30]. FHs are discussed in Section 2.2. In this section, it is also shown that FHs can be seen as KDEs with a variable kernel. Based on this observation, a statistical analysis of FHs is presented in Section 2.3. In Section 2.4, the results of this analysis are applied to crisp histograms, which are considered as a special case of FHs. It is shown that the same results are obtained as in the analysis of crisp histograms that is given in [30]. FHs that use triangular membership functions (mfs) turn out to have special properties. Some of these properties are analyzed in Section 2.5. In Section 2.6, FHs

are compared to other nonparametric density estimators. Finally, the use of FHs for estimating the mean and the variance of a population is discussed in Section 2.7. It should be noted that only univariate density estimation is considered in this chapter. Furthermore, the relation between the FHs discussed in this chapter and the PFSs discussed in the remainder of this thesis will become clear in Section 3.3 and 4.1.

2.1 Analysis of kernel density estimators

In this section, a statistical analysis of KDEs is given. All results in this section are taken from [30].

Let $x_1, \ldots, x_n$ denote a random sample of size n from a distribution with probability density function (pdf) $f(x)$. A KDE estimates $f(x)$ as follows

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i), \tag{2.1}$$

where K denotes the kernel function, which is usually a pdf, h denotes the smoothing parameter, and $K_h(t) = K(t/h)/h$.

For a specific value of x, the estimation of $f(x)$ can be seen as a standard point estimation problem in which the unknown parameter $f(x)$ is estimated by the point estimator $\hat{f}(x)$. The bias and the variance of the estimator $\hat{f}(x)$ are defined as

$$\text{Bias}\left(\hat{f}(x)\right) = \text{E}\hat{f}(x) - f(x) \tag{2.2}$$

and

$$\text{Var}\left(\hat{f}(x)\right) = \text{E}\left(\hat{f}(x) - \text{E}\hat{f}(x)\right)^2. \tag{2.3}$$

First consider the bias of $\hat{f}(x)$. The expectation of $\hat{f}(x)$, which is the first term in (2.2), can be written as

$$\text{E}\hat{f}(x) = \text{E}\left(\frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i)\right) = \frac{1}{n} \sum_{i=1}^{n} \text{E}K_h(x - x_i). \tag{2.4}$$

The random variables $x_1, \ldots, x_n$ are independent and identically distributed according to $f(x)$. Using X to denote a random variable that is distributed

according to $f(x)$, it therefore follows that

$$\begin{aligned}
\mathrm{E}\hat{f}(x) &= \mathrm{E}K_h(x-X) \\
&= \mathrm{E}\left(\frac{1}{h}K\left(\frac{x-X}{h}\right)\right) \\
&= \int \frac{1}{h}K\left(\frac{x-t}{h}\right) f(t)\, dt \\
&= \int K(w)\, f(x-hw)\, dw \\
&= \int K(w) \sum_{k=0}^{\infty} \frac{(-hw)^k f^{(k)}(x)}{k!} dw \\
&= f(x) \int K(w)\, dw - hf'(x) \int wK(w)\, dw \\
&\qquad + \frac{1}{2}h^2 f''(x) \int w^2 K(w)\, dw + O\left(h^3\right), \qquad (2.5)
\end{aligned}$$

where a Taylor series has been used. It should further be noted that in this chapter the symbol $\int$ should be read as $\int_{-\infty}^{\infty}$. Define μ_K and σ_K^2 as

$$\mu_K = \int wK(w)\, dw \qquad \text{and} \qquad \sigma_K^2 = \int w^2 K(w)\, dw. \qquad (2.6)$$

Assuming that K is a pdf implies that $\int K(w)\, dw = 1$. Equation (2.5) can then be written as

$$\mathrm{E}\hat{f}(x) = \mathrm{E}K_h(x-X) = f(x) - hf'(x)\,\mu_K + \frac{1}{2}h^2 f''(x)\,\sigma_K^2 + O\left(h^3\right). \quad (2.7)$$

Furthermore, by assuming that $\mu_K = 0$, it follows from (2.2) and (2.7) that

$$\mathrm{Bias}\left(\hat{f}(x)\right) = \frac{1}{2}h^2 f''(x)\,\sigma_K^2 + O\left(h^3\right). \qquad (2.8)$$

Now consider the variance of $\hat{f}(x)$. Since $x_1, \ldots, x_n$ are independent and identically distributed random variables, it follows that

$$\mathrm{Var}\left(\hat{f}(x)\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}\left(K_h(x-x_i)\right) = \frac{1}{n}\mathrm{Var}\left(K_h(x-X)\right). \qquad (2.9)$$

Notice that

$$\mathrm{Var}\left(K_h(x-X)\right) = \mathrm{E}\left(K_h(x-X)^2\right) - \left(\mathrm{E}K_h(x-X)\right)^2. \qquad (2.10)$$

The first term in (2.10) can be written as

$$\begin{aligned}
\mathrm{E}\left(K_h\left(x-X\right)^2\right) &= \mathrm{E}\left(\left(\frac{1}{h}K\left(\frac{x-X}{h}\right)\right)^2\right) \\
&= \int \frac{1}{h^2}K\left(\frac{x-t}{h}\right)^2 f\left(t\right)dt \\
&= \int \frac{1}{h}K\left(w\right)^2 f\left(x-hw\right)dw \\
&= \int \frac{1}{h}K\left(w\right)^2 \sum_{k=0}^{\infty}\frac{\left(-hw\right)^k f^{(k)}\left(x\right)}{k!}dw \\
&= \frac{f\left(x\right)R\left(K\right)}{h} + O\left(1\right), \qquad (2.11)
\end{aligned}$$

where $R\left(K\right)$ measures the roughness of the kernel function K. The roughness $R\left(\phi\right)$ of a function ϕ is defined as

$$R\left(\phi\right) = \int \phi\left(x\right)^2 dx. \qquad (2.12)$$

The second term in (2.10) equals the square of (2.7). Since $\mu_K = 0$, this results in

$$\left(\mathrm{E}K_h\left(x-X\right)\right)^2 = f\left(x\right)^2 + O\left(h^2\right). \qquad (2.13)$$

Substituting (2.11) and (2.13) in (2.10) and (2.10) in (2.9) gives

$$\mathrm{Var}\left(\hat{f}\left(x\right)\right) = \frac{f\left(x\right)R\left(K\right)}{nh} + O\left(\frac{1}{n}\right). \qquad (2.14)$$

The bias and the variance of the estimator $\hat{f}\left(x\right)$ are criteria for measuring the error of $\hat{f}\left(x\right)$ for a specific value of x. When $\hat{f}\left(x\right)$ is used for estimating an entire pdf rather than a single point on a pdf, one may want to consider a global error criterion. One such criterion is the mean integrated squared error (MISE) (also called the integrated mean squared error or IMSE) of $\hat{f}\left(x\right)$, which is defined as

$$\begin{aligned}
\mathrm{MISE}\left(\hat{f}\right) &= \mathrm{E}\left(\int\left(\hat{f}\left(x\right)-f\left(x\right)\right)^2 dx\right) \\
&= \int \mathrm{E}\left(\hat{f}\left(x\right)-f\left(x\right)\right)^2 dx \\
&= \int \left(\mathrm{E}\hat{f}\left(x\right)-f\left(x\right)\right)^2 + \mathrm{E}\left(\hat{f}\left(x\right)-\mathrm{E}\hat{f}\left(x\right)\right)^2 dx \\
&= \int \mathrm{Bias}^2\left(\hat{f}\left(x\right)\right) + \mathrm{Var}\left(\hat{f}\left(x\right)\right) dx \\
&= \mathrm{ISB}\left(\hat{f}\right) + \mathrm{IV}\left(\hat{f}\right), \qquad (2.15)
\end{aligned}$$

where the integrated squared bias (ISB) of $\hat{f}(x)$ and the integrated variance (IV) of $\hat{f}(x)$ are given by

$$\text{ISB}\left(\hat{f}\right) = \int \text{Bias}^2\left(\hat{f}(x)\right) dx \tag{2.16}$$

and

$$\text{IV}\left(\hat{f}\right) = \int \text{Var}\left(\hat{f}(x)\right) dx. \tag{2.17}$$

In the case that $\hat{f}(x)$ is a KDE, the ISB and the IV can be derived from (2.8) and (2.14), respectively. This results in

$$\text{ISB}\left(\hat{f}\right) = \frac{1}{4}h^4 R\left(f''\right)\sigma_K^4 + O\left(h^5\right) \tag{2.18}$$

and

$$\text{IV}\left(\hat{f}\right) = \frac{R(K)}{nh} + O\left(\frac{1}{n}\right). \tag{2.19}$$

In order to compare different kernel estimators, the optimal rate of convergence of the MISE as the sample size $n \to \infty$ is usually considered. For convergence of the MISE, it is necessary that $h \to 0$ and $nh \to \infty$ as $n \to \infty$. This follows from (2.18) and (2.19). In an asymptotic analysis of a kernel estimator, the asymptotic mean integrated squared error (AMISE) criterion is used, which is equal to the MISE criterion without the terms that are insignificant in the asymptotic analysis. For a KDE $\hat{f}(x)$, the AMISE equals

$$\text{AMISE}\left(\hat{f}\right) = \frac{R(K)}{nh} + \frac{1}{4}h^4 R\left(f''\right)\sigma_K^4. \tag{2.20}$$

The asymptotically optimal smoothing parameter, denoted by h^*, is the smoothing parameter h that results in the optimal rate of convergence of the AMISE (or MISE). From (2.20), it follows that

$$h^* = \left(\frac{R(K)}{R\left(f''\right)\sigma_K^4}\right)^{1/5} n^{-1/5}. \tag{2.21}$$

The optimal AMISE, denoted by AMISE*, is obtained by substituting (2.21) in (2.20). The result is given in the following theorem.

Theorem 2.1 *Let the kernel K be a pdf with $\mu_K = 0$. A KDE $\hat{f}(x)$ that uses kernel K has an AMISE given by (2.20) and an asymptotically optimal smoothing parameter given by (2.21). The optimal AMISE equals*

$$\text{AMISE}^*\left(\hat{f}\right) = \frac{5}{4}(\sigma_K R(K))^{4/5} R\left(f''\right)^{1/5} n^{-4/5}. \tag{2.22}$$

Therefore, the optimal AMISE of $\hat{f}(x)$ decreases at a rate of $O\left(n^{-4/5}\right)$.

The optimal kernel, i.e. the kernel K that minimizes $\sigma_K R(K)$, is known as the Epanechnikov kernel and is given by (see [30])

$$K(t) = \begin{cases} \frac{3}{4}\left(1-t^2\right) & \text{if } -1 \leq t \leq 1 \\ 0 & \text{otherwise.} \end{cases} \tag{2.23}$$

For this kernel, $\sigma_K = 1/\sqrt{5}$ and $R(K) = 3/5$. Therefore, a KDE $\hat{f}(x)$ that uses the Epanechnikov kernel has an optimal AMISE given by

$$\text{AMISE}^*\left(\hat{f}\right) = \frac{5}{4}\left(\frac{3}{5\sqrt{5}}\right)^{4/5} R\left(f''\right)^{1/5} n^{-4/5}. \tag{2.24}$$

2.2 Fuzzy histograms as kernel density estimators with a variable kernel

FHs have been introduced in [16, 18, 34]. In a FH, the sample space X is partitioned in a number of fuzzy sets A_j. The partitioning is done in such a way that

$$\sum_j \mu_{A_j}(x) = 1 \qquad \forall x \in X. \tag{2.25}$$

If the condition in (2.25) is satisfied, the sample space X is said to be 'well-defined' [16, 18, 34, 35] or, equivalently, the fuzzy sets A_j are said to be 'normalized disjunct' [36]. A FH estimates a pdf $f(x)$ as follows

$$\hat{f}(x) = \sum_j f(x|A_j)\, p_j, \tag{2.26}$$

where $f(x|A_j)$ denotes the conditional pdf of x given fuzzy event A_j. $f(x|A_j)$ is assumed to be given by

$$f(x|A_j) = \frac{\mu_{A_j}(x)}{\int \mu_{A_j}(x)\, dx}. \tag{2.27}$$

Furthermore, p_j in (2.26) denotes an estimate of the probability $\Pr(A_j)$. This estimate is calculated on the basis of a random sample $x_1, \ldots, x_n$ in the following way

$$p_j = \frac{1}{n}\sum_{i=1}^{n} \mu_{A_j}(x_i). \tag{2.28}$$

The KDE defined by (2.1) uses the same kernel for all values of x. A more general KDE is obtained by allowing the use of different kernels for different values of x. Such a KDE with a variable kernel is given by

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_x(x - x_i), \tag{2.29}$$

where the smoothing parameter has been omitted. (To avoid confusion, it should be noted that in the literature on KDEs [30] the term 'variable kernel' is sometimes used in a different sense than in this thesis.) Now consider the following theorem.

Theorem 2.2 *A FH given by (2.26), (2.27), and (2.28) is mathematically equivalent to a KDE with a variable kernel given by (2.29) if the kernels K_x are chosen as follows*

$$K_x(w) = \sum_j \frac{\mu_{A_j}(x-w)\,\mu_{A_j}(x)}{\int \mu_{A_j}(x)\,dx}. \tag{2.30}$$

Furthermore, if condition (2.25) is satisfied, then the kernels K_x in (2.30) are valid pdfs.

Proof: Substituting (2.27) and (2.28) in (2.26) and rearranging terms gives the following result for a FH

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\sum_j \frac{\mu_{A_j}(x_i)\,\mu_{A_j}(x)}{\int \mu_{A_j}(x)\,dx}. \tag{2.31}$$

The same result is obtained for a KDE with a variable kernel by substituting (2.30) in (2.29). This proves the mathematical equivalence of a FH and a KDE with a variable kernel given by (2.30). Also, it follows from (2.30) that

$$\begin{aligned}\int K_x(w)\,dw &= \int \sum_j \frac{\mu_{A_j}(x-w)\,\mu_{A_j}(x)}{\int \mu_{A_j}(x)\,dx}dw \\ &= \sum_j \frac{\int \mu_{A_j}(x-w)\,dw}{\int \mu_{A_j}(x)\,dx}\mu_{A_j}(x) \\ &= \sum_j \mu_{A_j}(x)\,. \end{aligned} \tag{2.32}$$

If condition (2.25) is satisfied, then (2.32) implies that $\int K_x(w)\,dw = 1$ for all $x \in X$. Moreover, since mfs are nonnegative, it follows from (2.30) that the kernels K_x are also nonnegative. Therefore, the kernels K_x are valid pdfs. This completes the proof of the theorem.

2.3 Analysis of fuzzy histograms

In this section, a statistical analysis of FHs is presented. Since FHs can be seen as KDEs with a variable kernel, the analysis in this section is actually an extension of the analysis of KDEs with a fixed kernel that was given in Section 2.1.

In the analysis in this section, the sample space is assumed to be $\mathbb{R}$. Moreover, the analysis is restricted to the special case that the sample space is uniformly partitioned. The sample space $\mathbb{R}$ is said to be uniformly partitioned if there are an infinite number of fuzzy sets A_j, with $j \in \mathbb{Z}$, and these fuzzy sets have mfs given by

$$\mu_{A_j}(x) = \mu_A\left(\frac{x}{h} - j\right) \qquad \forall j \in \mathbb{Z}, \tag{2.33}$$

where $0 < h < \infty$ is a smoothing parameter and μ_A is an arbitrary mf that satisfies

$$\int x\mu_A(x)\,dx = 0 \tag{2.34}$$

and

$$\sum_{j\in\mathbb{Z}} \mu_A(x + j) = 1 \qquad \forall x \in \mathbb{R}. \tag{2.35}$$

Equation (2.35) ensures that the partitioning of the sample space given by (2.33) satisfies condition (2.25). Also, (2.35) implies that

$$\begin{aligned}
\int \mu_A(x)\,dx &= \sum_{j\in\mathbb{Z}} \int_j^{j+1} \mu_A(x)\,dx \\
&= \sum_{j\in\mathbb{Z}} \int_0^1 \mu_A(x + j)\,dx \\
&= \int_0^1 \sum_{j\in\mathbb{Z}} \mu_A(x + j)\,dx \\
&= \int_0^1 dx \\
&= 1.
\end{aligned} \tag{2.36}$$

In a uniformly partitioned sample space, the denominator in (2.30) can be rewritten as follows by using (2.36)

$$\int \mu_{A_j}(x)\,dx = \int \mu_A\left(\frac{x}{h} - j\right) dx = h\int \mu_A(w)\,dw = h. \tag{2.37}$$

Substituting (2.33) and (2.37) in (2.30) gives

$$K_{x,h}(w) = \frac{1}{h}\sum_{j\in\mathbb{Z}} \mu_A\left(\frac{x - w}{h} - j\right)\mu_A\left(\frac{x}{h} - j\right). \tag{2.38}$$

Therefore, in a uniformly partitioned sample space, a FH is mathematically equivalent to a KDE that uses the variable kernel given by (2.38).

Consequently, a statistical analysis of a FH can be given by analyzing the equivalent KDE. This approach is followed below.

To determine the bias of a KDE $\hat{f}(x)$ that uses a variable kernel $K_{x,h}$, first consider the expectation of $\hat{f}(x)$

$$\begin{aligned}
\mathrm{E}\hat{f}(x) &= \mathrm{E}K_{x,h}(x-X) \\
&= \int K_{x,h}(x-t) f(t)\, dt \\
&= \int K_{x,h}(w) f(x-w)\, dw \\
&= \int K_{x,h}(w) \sum_{k=0}^{\infty} \frac{(-w)^k f^{(k)}(x)}{k!} dw \\
&= \sum_{k=0}^{\infty} \frac{(-1)^k f^{(k)}(x)}{k!} \int w^k K_{x,h}(w)\, dw. \qquad (2.39)
\end{aligned}$$

Since $\int K_{x,h}(w)\, dw = 1$, the bias of $\hat{f}(x)$ equals

$$\mathrm{Bias}\left(\hat{f}(x)\right) = \sum_{k=1}^{\infty} \frac{(-1)^k f^{(k)}(x)}{k!} \int w^k K_{x,h}(w)\, dw. \qquad (2.40)$$

For a FH in a uniformly partitioned sample space, the kernels $K_{x,h}$ are given by (2.38). Therefore, the integral in (2.39) and (2.40) can be written as

$$\begin{aligned}
\int w^k K_{x,h}(w)\, dw &= \int \frac{w^k}{h} \sum_{j\in\mathbb{Z}} \mu_A\left(\frac{x-w}{h} - j\right) \mu_A\left(\frac{x}{h} - j\right) dw \\
&= \sum_{j\in\mathbb{Z}} \mu_A\left(\frac{x}{h} - j\right) \int \frac{w^k}{h} \mu_A\left(\frac{x-w}{h} - j\right) dw \\
&= \sum_{j\in\mathbb{Z}} \mu_A\left(\frac{x}{h} - j\right) \int (x - hj - hv)^k \mu_A(v)\, dv.
\end{aligned} \qquad (2.41)$$

Now consider the ISB of $\hat{f}(x)$

$$\begin{aligned}
\mathrm{ISB}\left(\hat{f}\right) &= \int \mathrm{Bias}^2\left(\hat{f}(x)\right) dx \\
&= \int \left(\sum_{k=1}^{\infty} \frac{(-1)^k f^{(k)}(x)}{k!} \int w^k K_{x,h}(w)\, dw \right)^2 dx \\
&= h \int \left(\sum_{k=1}^{\infty} \frac{(-1)^k f^{(k)}(hu)}{k!} \int w^k K_{hu,h}(w)\, dw \right)^2 du, \qquad (2.42)
\end{aligned}$$

where x has been replaced by hu in the last step. Replacing x by hu in (2.41) results in

$$\int w^k K_{hu,h}(w)\,dw = h^k \sum_{j\in\mathbb{Z}} \mu_A(u-j) \int (u-j-v)^k \mu_A(v)\,dv. \tag{2.43}$$

For $k=1$, this gives

$$\int wK_{hu,h}(w)\,dw = h\sum_{j\in\mathbb{Z}} (u-j)\,\mu_A(u-j), \tag{2.44}$$

where (2.34) and (2.36) have been used. Using the results of (2.43) and (2.44), (2.42) can be written as

$$\begin{aligned}
\mathrm{ISB}\left(\hat{f}\right) &= h\int \left(-hf'(hu)\sum_{j\in\mathbb{Z}} (u-j)\,\mu_A(u-j) + \ldots\right)^2 du \\
&= h^3 \int f'(hu)^2 \left(\sum_{j\in\mathbb{Z}} (u-j)\,\mu_A(u-j)\right)^2 du + \ldots \\
&= h^3 \sum_{i\in\mathbb{Z}} \int_{i-\frac{1}{2}}^{i+\frac{1}{2}} f'(hu)^2 \left(\sum_{j\in\mathbb{Z}} (u-j)\,\mu_A(u-j)\right)^2 du + \ldots \\
&= h^3 \sum_{i\in\mathbb{Z}} f'(h\eta_i)^2 \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\sum_{j\in\mathbb{Z}} (u-j)\,\mu_A(u-j)\right)^2 du + \ldots,
\end{aligned} \tag{2.45}$$

where the last step follows from the generalized mean value theorem. This step is valid for some collection of points η_i, where $i-\frac{1}{2} < \eta_i < i+\frac{1}{2}$. The generalized mean value theorem states that

$$\int_a^b \phi(x)\,g(x)\,dx = \phi(c)\int_a^b g(x)\,dx, \tag{2.46}$$

for some value of c such that $a<c<b$. The functions ϕ and g are assumed to be continuous on the finite interval $[a,b]$, and g is also assumed to be nonnegative on this interval. The last step in (2.45) also uses the following result

$$\int_a^{a+1} \left(\sum_{j\in\mathbb{Z}} (u-j)\,\mu_A(u-j)\right)^2 du = c \qquad \forall a\in\mathbb{R}, \tag{2.47}$$

where the value of c is constant for all $a\in\mathbb{R}$. Equation (2.45) can be rewritten using numerical integration approximations. This results in

$$\mathrm{ISB}\left(\hat{f}\right) = h^2 P(\mu_A)\,R\left(f'\right) + O\left(h^3\right), \tag{2.48}$$

where $P(\phi)$ is defined as

$$P(\phi) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\sum_{j \in \mathbb{Z}} (x-j)\, \phi(x-j) \right)^2 dx. \tag{2.49}$$

The variance of a KDE $\hat{f}(x)$ with a variable kernel $K_{x,h}$ is given by

$$\mathrm{Var}\left(\hat{f}(x)\right) = \frac{1}{n}\mathrm{Var}\left(K_{x,h}(x-X)\right), \tag{2.50}$$

where

$$\mathrm{Var}\left(K_{x,h}(x-X)\right) = \mathrm{E}\left(K_{x,h}(x-X)^2\right) - \left(\mathrm{E}K_{x,h}(x-X)\right)^2. \tag{2.51}$$

The first term in (2.51) can be written as

$$\begin{aligned}
\mathrm{E}\left(K_{x,h}(x-X)^2\right) &= \int K_{x,h}(x-t)^2 f(t)\, dt \\
&= \int K_{x,h}(w)^2 f(x-w)\, dw \\
&= \int K_{x,h}(w)^2 \sum_{k=0}^{\infty} \frac{(-w)^k f^{(k)}(x)}{k!} dw \\
&= \sum_{k=0}^{\infty} \frac{(-1)^k f^{(k)}(x)}{k!} \int w^k K_{x,h}(w)^2 dw.
\end{aligned} \tag{2.52}$$

For a FH in a uniformly partitioned sample space, the kernels $K_{x,h}$ are given by (2.38). The integral in (2.52) can therefore be written as

$$\begin{aligned}
\int w^k K_{x,h}(w)^2 dw &= \int \frac{w^k}{h^2} \left(\sum_{j \in \mathbb{Z}} \mu_A\left(\frac{x-w}{h} - j\right) \mu_A\left(\frac{x}{h} - j\right) \right)^2 dw \\
&= \frac{1}{h} \int (x - hv)^k \left(\sum_{j \in \mathbb{Z}} \mu_A(v-j)\, \mu_A\left(\frac{x}{h} - j\right) \right)^2 dv.
\end{aligned} \tag{2.53}$$

Notice further that the second term in (2.51) equals the square of (2.39).

The IV of $\hat{f}(x)$ is given by

$$\mathrm{IV}\left(\hat{f}\right) = \int \mathrm{Var}\left(\hat{f}(x)\right) dx. \tag{2.54}$$

Replacing x by hu results in

$$\begin{aligned}
\mathrm{IV}\left(\hat{f}\right) &= h \int \mathrm{Var}\left(\hat{f}(hu)\right) du \\
&= \frac{h}{n} \int \mathrm{E}\left(K_{hu,h}(hu-X)^2\right) du \\
&\quad - \frac{h}{n} \int \left(\mathrm{E}K_{hu,h}(hu-X)\right)^2 du,
\end{aligned} \tag{2.55}$$

where (2.50) and (2.51) have been used. Replacing x by hu in (2.53) gives

$$\int w^k K_{hu,h}(w)^2 dw = h^{k-1} \int (u-v)^k \left(\sum_{j \in \mathbb{Z}} \mu_A(v-j)\, \mu_A(u-j) \right)^2 dv. \tag{2.56}$$

Using (2.52) and (2.56), the first term in (2.55) can be written as

$$\begin{aligned}
&\frac{h}{n} \int \mathrm{E}\left(K_{hu,h}(hu - X)^2 \right) du \\
&\quad = \frac{1}{n} \sum_{k=0}^{\infty} \frac{(-h)^k}{k!} \int f^{(k)}(hu) \int (u-v)^k \left(\sum_{j \in \mathbb{Z}} \mu_A(v-j)\, \mu_A(u-j) \right)^2 dv du \\
&\quad = \frac{1}{n} \int f(hu) \int \left(\sum_{j \in \mathbb{Z}} \mu_A(v-j)\, \mu_A(u-j) \right)^2 dv du + \ldots \\
&\quad = \frac{1}{n} \sum_{i \in \mathbb{Z}} \int_{i-\frac{1}{2}}^{i+\frac{1}{2}} f(hu) \int \left(\sum_{j \in \mathbb{Z}} \mu_A(v-j)\, \mu_A(u-j) \right)^2 dv du + \ldots \\
&\quad = \frac{1}{n} \sum_{i \in \mathbb{Z}} f(h\eta_i) \int_{-\frac{1}{2}}^{\frac{1}{2}} \int \left(\sum_{j \in \mathbb{Z}} \mu_A(v-j)\, \mu_A(u-j) \right)^2 dv du + \ldots .
\end{aligned} \tag{2.57}$$

The last step, which is valid for some collection of points η_i ($i - \frac{1}{2} < \eta_i < i + \frac{1}{2}$), follows from the generalized mean value theorem. Furthermore, the following result has been used

$$\int_a^{a+1} \int \left(\sum_{j \in \mathbb{Z}} \mu_A(v-j)\, \mu_A(u-j) \right)^2 dv du = c \qquad \forall a \in \mathbb{R}, \tag{2.58}$$

where the value of c is constant for all $a \in \mathbb{R}$. Rewriting (2.57) using numerical integration approximations gives

$$\begin{aligned}
&\frac{h}{n} \int \mathrm{E}\left(K_{hu,h}(hu - X)^2 \right) du \\
&\quad = \frac{1}{nh} \int f(x)\, dx \int_{-\frac{1}{2}}^{\frac{1}{2}} \int \left(\sum_{j \in \mathbb{Z}} \mu_A(v-j)\, \mu_A(u-j) \right)^2 dv du + O\left(\frac{1}{n} \right) \\
&\quad = \frac{Q(\mu_A)}{nh} + O\left(\frac{1}{n} \right),
\end{aligned} \tag{2.59}$$

where $Q(\phi)$ is defined as

$$Q(\phi) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \int \left(\sum_{j \in \mathbb{Z}} \phi(w-j)\, \phi(x-j) \right)^2 dw dx. \tag{2.60}$$

The last step in (2.59) can be made because $f(x)$ is a pdf and, consequently, $\int f(x)\,dx = 1$. Using (2.35), (2.36), (2.39), and (2.43), the second term in (2.55) can be written as

$$\begin{aligned}
&\frac{h}{n}\int \left(\mathrm{E}K_{hu,h}\left(hu - X\right)\right)^2 du \\
&\quad = \frac{h}{n}\int \left(\sum_{k=0}^{\infty} \frac{(-1)^k f^{(k)}(hu)}{k!} \int w^k K_{hu,h}(w)\,dw\right)^2 du \\
&\quad = \frac{h}{n}\int f(hu)^2\,du + \ldots .
\end{aligned} \tag{2.61}$$

In a similar way as the first term in (2.55), the second term can be rewritten using the generalized mean value theorem and numerical integration approximations. Equation (2.61) then becomes

$$\frac{h}{n}\int \left(\mathrm{E}K_{hu,h}\left(hu - X\right)\right)^2 du = \frac{R(f)}{n} + O\left(\frac{h}{n}\right). \tag{2.62}$$

The IV of $\hat{f}(x)$ is obtained by substituting (2.59) and (2.62) in (2.55). Since (2.59) decreases at a lower rate than (2.62) as the sample size $n \to \infty$ and the smoothing parameter $h \to 0$, the IV of $\hat{f}(x)$ can be written as

$$\mathrm{IV}\left(\hat{f}\right) = \frac{Q(\mu_A)}{nh} + O\left(\frac{1}{n}\right). \tag{2.63}$$

The AMISE of a KDE $\hat{f}(x)$ with a variable kernel $K_{x,h}$ given by (2.38) follows from (2.48) and (2.63)

$$\mathrm{AMISE}\left(\hat{f}\right) = \frac{Q(\mu_A)}{nh} + h^2 P(\mu_A) R\left(f'\right). \tag{2.64}$$

This equation is valid under the assumption that $P(\mu_A) > 0$. From (2.64), it follows that the asymptotically optimal smoothing parameter is given by

$$h^* = \left(\frac{Q(\mu_A)}{2P(\mu_A) R(f')}\right)^{1/3} n^{-1/3}. \tag{2.65}$$

The optimal AMISE is given in the following theorem.

Theorem 2.3 *Let $\hat{f}(x)$ be a FH given by (2.26), (2.27), and (2.28), let the sample space $\mathbb{R}$ be uniformly partitioned as defined by (2.33), (2.34), and (2.35), and let $P(\mu_A) > 0$. Then, $\hat{f}(x)$ has an AMISE given by (2.64) and an asymptotically optimal smoothing parameter given by (2.65). The optimal AMISE equals*

$$\mathrm{AMISE}^*\left(\hat{f}\right) = \left(\frac{27}{4} P(\mu_A) R\left(f'\right)\right)^{1/3} Q(\mu_A)^{2/3} n^{-2/3}. \tag{2.66}$$

Therefore, the optimal AMISE of $\hat{f}(x)$ decreases at a rate of $O\left(n^{-2/3}\right)$.

2.4 Crisp histograms as a special case of fuzzy histograms

If the fuzzy sets used by a FH for partitioning the sample space are actually crisp sets (i.e. the degree of membership is always either 0 or 1), then the FH reduces to a crisp histogram. Crisp histograms can therefore be considered as a special case of FHs. Consequently, the results of the analysis of FHs presented in the previous section can be applied to crisp histograms.

Consider a sample space $\mathbb{R}$ that is uniformly partitioned according to (2.33). Let μ_A, which must satisfy the conditions in (2.34) and (2.35), be defined as

$$\mu_A(x) = \begin{cases} 1 & \text{if } -\frac{1}{2} \leq x < \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{2.67}$$

Notice that (2.26), (2.27), (2.28), (2.33), and (2.67) define a crisp histogram with a fixed bin width.

For an asymptotic analysis of a crisp histogram $\hat{f}(x)$ with a fixed bin width, $P(\mu_A)$ and $Q(\mu_A)$ need to be calculated. Substituting (2.67) in (2.49) gives

$$\begin{aligned} P(\mu_A) &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\sum_{j \in \mathbb{Z}} (x-j)\, \mu_A(x-j) \right)^2 dx \\ &= \lim_{a \uparrow \frac{1}{2}} \int_{-\frac{1}{2}}^{a} \left(\sum_{j \in \mathbb{Z}} (x-j)\, \mu_A(x-j) \right)^2 dx \\ &= \lim_{a \uparrow \frac{1}{2}} \int_{-\frac{1}{2}}^{a} x^2 dx \\ &= \frac{1}{12}, \end{aligned} \tag{2.68}$$

and substituting (2.67) in (2.60) gives

$$\begin{aligned} Q(\mu_A) &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \int \left(\sum_{j \in \mathbb{Z}} \mu_A(w-j)\, \mu_A(x-j) \right)^2 dwdx \\ &= \lim_{a \uparrow \frac{1}{2}} \int_{-\frac{1}{2}}^{a} \int \left(\sum_{j \in \mathbb{Z}} \mu_A(w-j)\, \mu_A(x-j) \right)^2 dwdx \\ &= \lim_{a \uparrow \frac{1}{2}} \int_{-\frac{1}{2}}^{a} \int \mu_A(w)^2\, dwdx \\ &= \lim_{a \uparrow \frac{1}{2}} \int_{-\frac{1}{2}}^{a} \int \mu_A(w)\, dwdx \\ &= 1. \end{aligned} \tag{2.69}$$

Using (2.68) and (2.69), the following asymptotic results are obtained for $\hat{f}(x)$

$$\text{AMISE}\left(\hat{f}\right) = \frac{1}{nh} + \frac{1}{12} h^2 R\left(f'\right), \tag{2.70}$$

$$h^* = \left(\frac{6}{R\left(f'\right)} \right)^{1/3} n^{-1/3}, \tag{2.71}$$

and

$$\text{AMISE}^*\left(\hat{f}\right) = \left(\frac{3}{4}\right)^{2/3} R\left(f'\right)^{1/3} n^{-2/3}. \tag{2.72}$$

These results correspond with the results obtained by Scott in his analysis of crisp histograms (see Theorem 3.3 in [30]).

Scott [30] has also studied the properties of averaged shifted histograms. An averaged shifted histogram results from averaging a number of crisp histograms with the same bin width but with different locations for the bin edges. Like ordinary crisp histograms, averaged shifted histograms can be considered as a special case of FHs. For example, an averaged shifted histogram constructed from m histograms that are all given equal weight is mathematically equivalent to a FH with μ_A given by

$$\mu_A(x) = \begin{cases} \dfrac{1}{m} & \text{if } -\dfrac{1}{2}m \leq x < \dfrac{1}{2}m \\ 0 & \text{otherwise.} \end{cases} \tag{2.73}$$

Notice that (2.73) satisfies the conditions in (2.34) and (2.35). Also, (2.73) reduces to (2.67) for $m = 1$.

2.5 Fuzzy histograms with triangular membership functions

Consider a FH in a uniformly partitioned sample space $\mathbb{R}$ with μ_A given by

$$\mu_A(x) = \begin{cases} \frac{1}{m} + \frac{x}{m^2} & \text{if } -m \leq x < 0 \\ \frac{1}{m} - \frac{x}{m^2} & \text{if } 0 \leq x < m \\ 0 & \text{otherwise,} \end{cases} \tag{2.74}$$

where m is a positive integer. It can be shown that (2.74) satisfies the conditions in (2.34) and (2.35). (Notice that m is restricted to integer values in order to satisfy (2.35).) Equation (2.74) results in a partitioning of the sample space in fuzzy sets with triangular mfs.

It will now be shown that $P(\mu_A) = 0$ for μ_A given by (2.74). From (2.49), it can be seen that for a continuous function $\phi(x)$

$$\begin{aligned} P(\phi) = 0 &\Leftrightarrow \forall x \in \left[-\tfrac{1}{2}, \tfrac{1}{2}\right] : \sum_{j\in\mathbb{Z}} (x-j)\,\phi(x-j) = 0 \\ &\Leftrightarrow \forall x \in \mathbb{R} : \sum_{j\in\mathbb{Z}} (x-j)\,\phi(x-j) = 0. \end{aligned} \tag{2.75}$$

Assume, without loss of generality, that $0 \leq x < 1$. It can then be observed that for μ_A given by (2.74)

$$\begin{aligned} &\sum_{j\in\mathbb{Z}} (x-j)\,\mu_A(x-j) \\ &\quad = \sum_{j=1-m}^{0} (x-j)\left(\frac{1}{m} - \frac{x-j}{m^2}\right) + \sum_{j=1}^{m} (x-j)\left(\frac{1}{m} + \frac{x-j}{m^2}\right) \\ &\quad = \sum_{j=1-m}^{0} \left(-\frac{x^2}{m^2} - \frac{j^2}{m^2} + \frac{2xj}{m^2} + \frac{x}{m} - \frac{j}{m}\right) \\ &\qquad + \sum_{j=1}^{m} \left(\frac{x^2}{m^2} + \frac{j^2}{m^2} - \frac{2xj}{m^2} + \frac{x}{m} - \frac{j}{m}\right) \\ &\quad = \left(\frac{2x}{m^2} \sum_{j=1-m}^{0} j\right) - \left(\frac{2x}{m^2} \sum_{j=1}^{m} j\right) + 2x \\ &\quad = \left(-x + \frac{x}{m}\right) - \left(x + \frac{x}{m}\right) + 2x \\ &\quad = 0. \end{aligned} \tag{2.76}$$

From (2.75) and (2.76), it follows that $P(\mu_A) = 0$ for all values of m. Because $P(\mu_A) = 0$, the results in Theorem 2.3 cannot be applied to FHs

with triangular mfs. The asymptotic properties of FHs with triangular mfs will therefore be derived in the remainder of this section.

Let $\hat{f}(x)$ denote a FH in a uniformly partitioned sample space $\mathbb{R}$ and let $P(\mu_A) = 0$. The ISB of $\hat{f}(x)$ can be derived from (2.42) and (2.43). For $k = 1$, (2.43) has been rewritten in (2.44). Since $P(\mu_A) = 0$, (2.44) can be combined with (2.75), which results in

$$\int wK_{hu,h}(w)\,dw = h\sum_{j\in\mathbb{Z}}(u-j)\,\mu_A(u-j) = 0. \tag{2.77}$$

Furthermore, for $k = 2$ (2.43) becomes

$$\int w^2K_{hu,h}(w)\,dw = h^2\left(\int v^2\mu_A(v)\,dv + \sum_{j\in\mathbb{Z}}(u-j)^2\mu_A(u-j)\right), \tag{2.78}$$

where (2.34), (2.35), and (2.36) have been used. Using (2.77) and (2.78), (2.42) can be written as

$$\begin{aligned}
\mathrm{ISB}\left(\hat{f}\right) &= h\int\left(\frac{1}{2}h^2f''(hu)\left(\int v^2\mu_A(v)\,dv\right.\right.\\
&\qquad\left.\left.+\sum_{j\in\mathbb{Z}}(u-j)^2\mu_A(u-j)\right)+\ldots\right)^2du\\
&= \frac{1}{4}h^5\sum_{i\in\mathbb{Z}}f''(h\eta_i)^2\int_{-\frac{1}{2}}^{\frac{1}{2}}\left(\int v^2\mu_A(v)\,dv\right.\\
&\qquad\left.+\sum_{j\in\mathbb{Z}}(u-j)^2\mu_A(u-j)\right)^2du+\ldots,
\end{aligned} \tag{2.79}$$

where the last step, which is valid for some collection of points η_i $(i-\frac{1}{2} < \eta_i < i+\frac{1}{2})$, follows from the generalized mean value theorem. Rewriting (2.79) using numerical integration approximations gives

$$\mathrm{ISB}\left(\hat{f}\right) = \frac{1}{4}h^4S(\mu_A)\,R\left(f''\right)+O\left(h^5\right), \tag{2.80}$$

where $S(\phi)$ is defined as

$$S(\phi) = \int_{-\frac{1}{2}}^{\frac{1}{2}}\left(\int w^2\phi(w)\,dw + \sum_{j\in\mathbb{Z}}(x-j)^2\phi(x-j)\right)^2dx. \tag{2.81}$$

The equation for the IV of a FH is the same for $P(\mu_A) = 0$ and $P(\mu_A) > 0$. The IV of $\hat{f}(x)$ is therefore given by (2.63). The asymptotic properties of $\hat{f}(x)$ follow from (2.63) and (2.80) and are summarized in the following theorem.

Theorem 2.4 *Let $\hat{f}(x)$ be a FH given by (2.26), (2.27), and (2.28), let the sample space $\mathbb{R}$ be uniformly partitioned as defined by (2.33), (2.34), and (2.35), and let $P(\mu_A) = 0$. Then,*

$$\text{AMISE}\left(\hat{f}\right) = \frac{Q(\mu_A)}{nh} + \frac{1}{4}h^4 S(\mu_A) R\left(f''\right), \tag{2.82}$$

$$h^* = \left(\frac{Q(\mu_A)}{S(\mu_A) R(f'')}\right)^{1/5} n^{-1/5}, \tag{2.83}$$

and

$$\text{AMISE}^*\left(\hat{f}\right) = \frac{5}{4}\left(S(\mu_A) R\left(f''\right)\right)^{1/5} Q(\mu_A)^{4/5} n^{-4/5}. \tag{2.84}$$

Therefore, the optimal AMISE of $\hat{f}(x)$ decreases at a rate of $O\left(n^{-4/5}\right)$.

For the triangular mfs given by (2.74), it turns out that the calculation of $Q(\mu_A)$ and $S(\mu_A)$ for an arbitrary value of m is rather complicated. For $m = 1$, however, it can be shown that $Q(\mu_A) = \frac{1}{2}$ and $S(\mu_A) = \frac{7}{60}$. Substituting these values in (2.84) gives the following result.

Theorem 2.5 *Let $\hat{f}(x)$ be a FH given by (2.26), (2.27), and (2.28), let the sample space $\mathbb{R}$ be uniformly partitioned as defined by (2.33), (2.34), and (2.35), let μ_A be a triangular mf as defined by (2.74), and let $m = 1$. Then,*

$$\text{AMISE}^*\left(\hat{f}\right) = \frac{5}{8}\left(\frac{7}{30}R\left(f''\right)\right)^{1/5} n^{-4/5}. \tag{2.85}$$

The density estimator $\hat{f}(x)$ in Theorem 2.5 is equivalent with a density estimator that has been studied in the statistical literature. In [11], this estimator is called a frequency polygon based on a linearly weighted discretization of the data. From Equation (3.2) in [11], the same expression can be derived for the optimal AMISE as in (2.85).

Finally, it should be noted that the triangular mfs defined by (2.74) are not the only mfs for which $P(\mu_A) = 0$. It can be shown that some trapezoidal mfs also satisfy $P(\mu_A) = 0$.

2.6 Comparison between fuzzy histograms and other nonparametric density estimators

When choosing a nonparametric density estimator, two important criteria are the statistical efficiency and the computational efficiency of an estimator. In this thesis, the statistical efficiency of an estimator is determined by the estimator's AMISE*. Using the criteria of statistical efficiency and computational efficiency, crisp histograms and KDEs can be considered as two extremes. Crisp histograms are very efficient computationally but quite

inefficient statistically. KDEs, on the other hand, are quite efficient statistically but very inefficient computationally. The difference in statistical efficiency between crisp histograms and KDEs follows from their AMISE*. Comparing (2.22) and (2.72), the AMISE* of a crisp histogram turns out to decrease at a rate of $O\left(n^{-2/3}\right)$, whereas the AMISE* of a KDE turns out to decrease at a rate of $O\left(n^{-4/5}\right)$.

In order to combine the advantages of crisp histograms and KDEs, other nonparametric density estimators have been introduced (see for example Scott [30] and Jones [11]). Examples include the frequency polygon, which is formed by linear interpolation of adjacent mid-bin values of a crisp histogram, and the averaged shifted histogram, which was briefly discussed in Section 2.4. The frequency polygon and the averaged shifted histogram can also be combined, which results in the frequency polygon of an averaged shifted histogram. This density estimator has an AMISE* given by (derived from Theorem 5.2 in [30])

$$\text{AMISE}^*\left(\hat{f}\right) = \frac{5}{12}\left(\left(\frac{4}{3} + \frac{4}{3m^2} + \frac{3}{5m^4}\right) R\left(f''\right)\right)^{1/5} n^{-4/5}, \tag{2.86}$$

where m denotes the number of underlying histograms used in the construction of the averaged shifted histogram. For $m = 1$, (2.86) reduces to the AMISE* of an ordinary frequency polygon (cf Theorem 4.1 in [30]). As $m \to \infty$, (2.86) approaches the AMISE* of a KDE that uses a triangular kernel. Notice that for all values of m, the AMISE* in (2.86) decreases at the same rate as the AMISE* of a KDE. This means that the combination of a frequency polygon and an averaged shifted histogram results in a high level of statistical efficiency.

For a FH given by (2.26), (2.27), and (2.28) in a sample space $\mathbb{R}$ that is uniformly partitioned as defined by (2.33), (2.34), and (2.35), the rate of convergence of the AMISE* depends on the value of $P\left(\mu_A\right)$. For $P\left(\mu_A\right) > 0$, the AMISE* decreases at a rate of $O\left(n^{-2/3}\right)$, as stated in Theorem 2.3. For $P\left(\mu_A\right) = 0$, Theorem 2.4 states that the AMISE* decreases at a rate of $O\left(n^{-4/5}\right)$. In other words, if $P\left(\mu_A\right) > 0$, the AMISE* of a FH has the same rate of convergence as the AMISE* of a crisp histogram. This means that in this case, a FH is statistically quite inefficient. On the other hand, if $P\left(\mu_A\right) = 0$, the AMISE* of a FH decreases at the same rate as the AMISE* of a KDE, which is statistically quite efficient. Therefore, by choosing an appropriate mf μ_A, for example a triangular mf as defined by (2.74), a FH can be a statistically quite efficient density estimator. Since a FH is computationally much more efficient than a KDE (as the sample size $n \to \infty$), a FH can be used to obtain both a high level of statistical efficiency and a high level of computational efficiency.

In Table 2.1, a comparison is made between a number of nonparametric density estimators. All estimators in the table have a high level of statistical efficiency, since they all have an AMISE* that decreases at a rate of

Density estimator	Relative AMISE*	Equivalent sample size
KDE	1.000	1.000
FH	1.071	1.089
FP	1.210	1.269
FP-ASH ($m = 2$)	1.062	1.078
FP-ASH ($m = 3$)	1.034	1.043
FP-ASH ($m \to \infty$)	1.011	1.014

Table 2.1: Comparison between a number of nonparametric density estimators.

$O\left(n^{-4/5}\right)$. In the first column of the table, 'KDE' refers to a KDE that uses the optimal Epanechnikov kernel. 'FH' refers to a FH in a uniformly partitioned sample space that uses the triangular mf given by (2.74) for $m = 1$. (Notice that it might be possible that other mfs are statistically more efficient.) Furthermore, 'FP' refers to a frequency polygon, and 'FP-ASH' refers to a frequency polygon of an averaged shifted histogram. In the second column of Table 2.1, the relative AMISE* of each density estimator is reported. The relative AMISE* of an estimator equals the ratio between the AMISE* of that estimator and the AMISE* of a KDE that uses the Epanechnikov kernel. The equivalent sample size for an estimator, reported in the third column of the table, gives the ratio between the sample size required by that estimator and the sample size required by a KDE that uses the Epanechnikov kernel such that they both have the same AMISE*. It is interesting to observe that the FH in Table 2.1 is only slightly less statistically efficient than the KDE. The AMISE* of the FH is 7.1% higher than the AMISE* of the KDE, and the FH requires 8.9% more data to obtain the same AMISE* as the KDE. Also, the FH performs much better than the frequency polygon, which has a 21.0% higher AMISE* than the KDE and requires 26.9% more data to obtain the same AMISE*. Compared to the combination of a frequency polygon and an averaged shifted histogram, the FH is competitive when the averaged shifted histogram is constructed from only two underlying histograms ($m = 2$). When more than two histograms are used ($m > 2$), the frequency polygon of the averaged shifted histogram is statistically somewhat more efficient than the FH.

2.7 Estimation of the mean and the variance of a population

Given a random sample $x_1, \ldots, x_n$, one may be interested in estimating the population mean μ and the population variance σ^2. Statistical theory states

that unbiased estimates of μ and σ^2 are given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.87}$$

and

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\left(\left(\sum_{i=1}^{n} x_i^2\right) - n\bar{x}^2\right), \tag{2.88}$$

respectively.

When it is assumed that the sample $x_1, \ldots, x_n$ has been drawn from a distribution with an unknown pdf $f(x)$ and when an estimate $\hat{f}(x)$ of this pdf has been obtained, one may choose to use the mean and the variance of $\hat{f}(x)$ as estimates of μ and σ^2. This results in

$$\bar{x} = \int \hat{f}(x)\, x\, dx \tag{2.89}$$

and

$$s^2 = \int \hat{f}(x)\,(x - \bar{x})^2\, dx = \int \hat{f}(x)\, x^2\, dx - \bar{x}^2. \tag{2.90}$$

(Notice that this approach to estimating the mean and the variance is followed in PFSs, as will be discussed in Chapter 3.) Now suppose that $\hat{f}(x)$ is a FH given by (2.26), (2.27), and (2.28) in a uniformly partitioned sample space $\mathbb{R}$ as defined by (2.33), (2.34), and (2.35). Equation (2.89) then becomes

$$\bar{x} = \sum_j \frac{\int \mu_{A_j}(x)\, x\, dx}{\int \mu_{A_j}(x)\, dx} p_j = \sum_j \frac{h^2 j}{h} p_j = h\sum_j j p_j \tag{2.91}$$

and (2.90) becomes

$$\begin{aligned} s^2 &= \left(\sum_j \frac{\int \mu_{A_j}(x)\, x^2\, dx}{\int \mu_{A_j}(x)\, dx} p_j\right) - \bar{x}^2 \\ &= h^2\left(\sum_j \left(j^2 + \int \mu_A(x)\, x^2\, dx\right) p_j\right) - \bar{x}^2, \end{aligned} \tag{2.92}$$

where p_j is obtained from the sample $x_1, \ldots, x_n$ using (2.28).

An interesting question is whether (2.91) and (2.92) are equivalent to the unbiased estimates in (2.87) and (2.88). In general, it turns out that this is not the case. However, if mf μ_A is triangular, then (2.91) and (2.87) can be shown to be equivalent. This result is given in the following theorem.

Theorem 2.6 *Let $\hat{f}(x)$ be a FH given by (2.26), (2.27), and (2.28), let the sample space $\mathbb{R}$ be uniformly partitioned as defined by (2.33), (2.34), and (2.35), and let μ_A be a triangular mf as defined by (2.74). Then, (2.91) and (2.87) are equivalent.*

Proof: Substitution of (2.28) and (2.33) in (2.91) results in

$$\bar{x} = h\sum_j j\frac{1}{n}\sum_{i=1}^{n}\mu_A\left(\frac{x_i}{h} - j\right) = \frac{1}{n}\sum_{i=1}^{n} h\sum_j j\mu_A\left(\frac{x_i}{h} - j\right). \tag{2.93}$$

To prove the theorem, (2.93) and (2.87) must be equivalent. This is the case if

$$h\sum_j j\mu_A\left(\frac{x}{h} - j\right) = x \qquad \forall x \in \mathbb{R}. \tag{2.94}$$

Now assume, without loss of generality, that $x \in [0, h)$. Substitution of (2.74) in the left part of (2.94) then gives

$$\begin{aligned} h\sum_j j\mu_A\left(\frac{x}{h} - j\right) &= h\sum_{j=1-m}^{-1} j\left(\frac{1}{m} - \frac{1}{m^2}\left(\frac{x}{h} - j\right)\right) \\ &\qquad + h\sum_{j=1}^{m} j\left(\frac{1}{m} + \frac{1}{m^2}\left(\frac{x}{h} - j\right)\right) \\ &= \frac{x}{m^2}\sum_{j=1}^{m-1} j + \frac{x}{m^2}\sum_{j=1}^{m} j \\ &= (m-1)\frac{x}{2m} + (m+1)\frac{x}{2m} \\ &= x. \end{aligned} \tag{2.95}$$

For μ_A defined by (2.74), the condition in (2.94) is therefore satisfied. This completes the proof of the theorem.

Theorem 2.6 indicates that FHs with triangular mfs have a special property. Generally, when a sample from a population is used for estimating a pdf and when the estimated pdf is subsequently used for estimating the population mean, a different estimate is obtained than the unbiased estimate that is provided by the sample mean. (As a very simple example, consider a crisp histogram with one of its bins between 0 and 1. Suppose that the sample consists of only one element, which has a value of 0.25. The pdf estimated using the crisp histogram then equals a uniform distribution between 0 and 1. It follows that the use of the estimated pdf for estimating the population mean results in an estimate of 0.5. This estimate does not equal the estimate of 0.25 that is provided by the sample mean.) For FHs with triangular mfs, this is not the case and the mean of the estimated pdf is always the same as the mean of the sample. Therefore, a pdf estimated using

a FH with triangular mfs contains the same information about the population mean as the sample that was used for estimating the pdf. It should be emphasized that this property only holds for the population mean. There is no similar property for the population variance.

Chapter 3

Probabilistic Fuzzy Systems

In this chapter, PFSs are discussed. The discussion is based on [16, 18, 34]. These are the papers in which the idea of PFSs has been introduced. However, compared to [16, 18, 34] the discussion in this chapter contains some important differences and additions:

1. A more accurate mathematical notation is used.

2. Equation (14) in [18] and (3) and (6) in [34], which are incorrect, are omitted.

3. The assumption of a well-defined input space, which is actually not necessary, is omitted.

4. The procedure for obtaining a conditional pdf from a conditional probability distribution over fuzzy sets is interpreted in a different way than in [16, 18, 34].

5. In addition to an estimate of the conditional expectation, an estimate of the conditional variance is provided.

The last two points only relate to PFSs that are applied to regression problems.

This chapter is organized as follows. In Section 3.1, PFSs for classification tasks are considered. These PFSs turn out to be functionally equivalent to the radial basis function networks for classification tasks that are described in [2, 21]. A proof of this functional equivalence is given in Section 3.2. PFSs for regression tasks, which can be seen as an extension of PFSs for classification tasks, are discussed in Section 3.3. It should be noted that the issue of estimating the parameters in a PFS is not considered in this chapter. This issue is studied in detail in Chapter 4 and 5.

3.1 Probabilistic fuzzy systems for classification tasks

Consider the task of determining the (crisp) class $y \in \{C_1, \ldots, C_c\}$ to which a data point $\mathbf{x} = (x_1, \ldots, x_d) \in X$ belongs. To perform this task, a PFS can be used with probabilistic fuzzy rules that have the following general form

$$\begin{aligned} \text{If } \mathbf{x} \text{ is } A_j \text{ then } \underline{y} &= C_1 \text{ with probability } p_{j,1} \text{ and} \\ \underline{y} &= C_2 \text{ with probability } p_{j,2} \text{ and} \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \underline{y} &= C_c \text{ with probability } p_{j,c}. \end{aligned} \tag{3.1}$$

Notice that in this thesis $\underline{y}$ denotes a random variable and y denotes a particular value that $\underline{y}$ may take. The fuzzy sets A_j $(j = 1, \ldots, a)$ in (3.1) are defined in the d-dimensional input space X. For each fuzzy set there is a corresponding probabilistic fuzzy rule. Furthermore, the probability parameters $p_{j,k}$ in (3.1) satisfy

$$p_{j,k} \geq 0 \qquad \text{for } j = 1, \ldots, a \text{ and } k = 1, \ldots, c \tag{3.2}$$

and

$$\sum_{k=1}^{c} p_{j,k} = 1 \qquad \text{for } j = 1, \ldots, a. \tag{3.3}$$

Let μ_{A_j} denote the mf of a fuzzy set A_j. In [16, 18, 34], the input space X is assumed to be well-defined, which means that $\sum_{j=1}^{a} \mu_{A_j}(\mathbf{x}) = 1$ for all $\mathbf{x} \in X$. In this thesis I do not make this assumption. Instead, I choose to normalize the mfs μ_{A_j}. The normalized mfs are given by

$$\bar{\mu}_{A_j}(\mathbf{x}) = \frac{\mu_{A_j}(\mathbf{x})}{\sum_{j'=1}^{a} \mu_{A_{j'}}(\mathbf{x})}. \tag{3.4}$$

A PFS with rules given by (3.1) provides an estimate of $\Pr(\underline{y}|\mathbf{x})$, the conditional probability distribution of $\underline{y}$ given $\mathbf{x}$. Similarly to Takagi-Sugeno fuzzy reasoning, the normalized mfs $\bar{\mu}_{A_j}$ determine the activations of the probabilistic fuzzy rules. The estimate $\hat{p}(C_k|\mathbf{x})$ of a conditional probability $\Pr(C_k|\mathbf{x})$ is therefore obtained as follows

$$\hat{p}(C_k|\mathbf{x}) = \sum_{j=1}^{a} \bar{\mu}_{A_j}(\mathbf{x})\, p_{j,k}. \tag{3.5}$$

The estimated conditional probability distribution $\hat{p}(\underline{y}|\mathbf{x})$ can be used for classifying a data point $\mathbf{x}$. The following classification rule minimizes the probability of misclassification

$$\hat{y} = C_k \qquad \text{if } \hat{p}(C_k|\mathbf{x}) > \hat{p}(C_{k'}|\mathbf{x}) \text{ for all } k' \neq k. \tag{3.6}$$

It may be interesting to note that the fuzzy classifiers discussed in [1] make use of probabilistic fuzzy rules that have a very similar form as in (3.1). The only difference is that in the approach followed in [1] a certainty factor is attached to each rule. In the above approach, certainty factors are not considered and all rules are given equal weight. In Section 5.3, the classification performance of PFSs is studied on two benchmark data sets and the results are compared with the results that are reported in [1]. Classifiers that are similar to the PFSs discussed in this section are also considered in [26].

3.2 Functional equivalence to radial basis function networks for classification tasks

For regression problems, it is well-known that under certain conditions Takagi-Sugeno fuzzy systems are functionally equivalent to radial basis function networks (RBFNs) [7, 9, 10]. In this section, it is shown that a similar equivalence exists for classification problems. More specifically, it is shown that for classification problems the PFSs described in Section 3.1 are functionally equivalent to the RBFNs described in [2, 21]. Because of this equivalence, the learning algorithms that are used in the RBFNs in [2, 21] can also be used in PFSs for classification tasks.

First, I briefly discuss how RBFNs can be applied to classification problems. Although the discussion follows [2, 21], I use a somewhat different mathematical notation in order to emphasize the similarities between RBFNs and PFSs. In an RBFN, the activation of output unit z_k given a data point $\mathbf{x}$ is usually calculated as follows

$$z_k(\mathbf{x}) = \sum_{j=1}^{h} w_{j,k}\, \phi_j(\mathbf{x}). \tag{3.7}$$

The functions ϕ_j are called radial basis functions. These functions determine the activations of the hidden units of an RBFN. A typical choice is to use Gaussian basis functions, but other choices are also possible. The parameters $w_{j,k}$ are the weights between the hidden units and the output units of an RBFN. (A bias weight $w_{0,k}$ is sometimes added to the expression in (3.7). However, in classification problems bias weights need not be used.) Notice that the output units of an RBFN use an identity activation function. Furthermore, in an RBFN it is also possible that the activations of the hidden units are normalized to sum to 1. Equation (3.7) then becomes

$$z_k(\mathbf{x}) = \frac{\sum_{j=1}^{h} w_{j,k}\, \phi_j(\mathbf{x})}{\sum_{j=1}^{h} \phi_j(\mathbf{x})}. \tag{3.8}$$

Now consider the application of RBFNs to classification problems as described in [2, 21]. Like in Section 3.1, let $\mathbf{x}$ denote a data point and let $y \in \{C_1, \ldots, C_c\}$ denote the class to which a data point belongs. When using an RBFN for a classification task, the network has for each class C_k a corresponding output unit z_k. The activations of the network's hidden units are normalized, which means that the activations of the network's output units are given by (3.8). These activations are interpreted as estimates of the conditional probabilities $\Pr(C_k|\mathbf{x})$. Under this interpretation, the network as a whole provides an estimate of the conditional probability distribution $\Pr(\underline{y}|\mathbf{x})$. This estimate can be used for classifying a data point $\mathbf{x}$ according to

$$\hat{y} = C_k \qquad \text{if } z_k(\mathbf{x}) > z_{k'}(\mathbf{x}) \text{ for all } k' \neq k. \tag{3.9}$$

This classification rule is very similar to the rule given by (3.6) that is used in PFSs.

The following theorem states that for classification problems PFSs and RBFNs are functionally equivalent.

Theorem 3.1 *A PFS described in Section 3.1 is functionally equivalent to an RBFN described in this section if the following conditions are satisfied:*

1. *The number of fuzzy sets a is equal to the number of radial basis functions h.*

2. *Each mf μ_{A_j} is equal to a radial basis function $\phi_{j'}$.*

3. *If $\mu_{A_j} = \phi_{j'}$, then the probability parameters $p_{j,k}$ are equal to the weights $w_{j',k}$.*

Proof: If the conditions of the theorem are satisfied, then it follows from (3.4), (3.5), and (3.8) that $\hat{p}(C_k|\mathbf{x}) = z_k(\mathbf{x})$ for $k = 1, \ldots, c$. This means that the PFS and the RBFN provide identical estimates of the conditional probability distribution $\Pr(\underline{y}|\mathbf{x})$. As a consequence, the classification rules in (3.6) and (3.9) give identical classifications, which implies that the PFS and the RBFN are functionally equivalent. This completes the proof of the theorem.

It should be emphasized that Theorem 3.1 only applies to RBFNs for classification tasks that follow the description given in [2, 21]. In the literature on RBFNs, networks that use unnormalized hidden unit activations are sometimes applied to classification problems. The activations of the output units of these networks, which are calculated using (3.7), cannot be interpreted as estimates of conditional probabilities. Instead, the activations of the output units are interpreted as discriminant functions. Theorem 3.1 does not apply to these networks.

3.3 Probabilistic fuzzy systems for regression tasks

Consider the regression problem of estimating a function $f : X \to Y$. Typically, $f(\mathbf{x})$ contains an error term and is therefore stochastic. In that case, one is usually interested in estimating the expectation of $f(\mathbf{x})$. However, it may also be useful to know, for each value of $\mathbf{x}$, the variance of $f(\mathbf{x})$ or even the entire pdf of $f(\mathbf{x})$. An estimate of the pdf of $f(\mathbf{x})$ provides the most general information, since estimates of the expectation and the variance (and other statistics) can be derived from it. Notice that estimating the pdf of $f(\mathbf{x})$ for each value of $\mathbf{x}$ is equivalent to estimating the conditional pdf $p(y|\mathbf{x})$, where $\mathbf{x} \in X$ and $y \in Y$. Therefore, one approach to solving a regression problem is to estimate the conditional pdf. This is the approach that is followed in this section.

In Subsection 3.3.1, the use of PFSs for estimating conditional pdfs is discussed. The application of PFSs to regression problems is discussed in Subsection 3.3.2. In this subsection, it is pointed out how PFSs that provide estimates of conditional pdfs can be used for estimating the expectation and the variance of stochastic functions.

3.3.1 Estimation of a conditional probability density function

A PFS for estimating conditional pdfs uses probabilistic fuzzy rules that have the following general form

$$\begin{aligned} \text{If } \mathbf{x} \text{ is } A_j \text{ then } & \underline{y} \text{ is } C_1 \text{ with probability } p_{j,1} \text{ and} \\ & \underline{y} \text{ is } C_2 \text{ with probability } p_{j,2} \text{ and} \\ & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ & \underline{y} \text{ is } C_c \text{ with probability } p_{j,c}, \end{aligned} \tag{3.10}$$

where $j = 1, \ldots, a$. These rules are very similar to the rules in (3.1), which are used in PFSs for classification tasks. The conditions in (3.2) and (3.3) also apply to the rules in (3.10). The difference with PFSs for classification tasks is that $C_1, \ldots, C_c$ denote fuzzy sets instead of crisp classes. These fuzzy sets are defined in the output space Y in such a way that Y is well-defined. This means that the mfs μ_{C_k} satisfy

$$\sum_{k=1}^{c} \mu_{C_k}(y) = 1 \qquad \forall y \in Y. \tag{3.11}$$

Like in PFSs for classification tasks, the activations of the probabilistic fuzzy rules are determined by the normalized mfs of the antecedent fuzzy sets A_j. Therefore, the estimates $\hat{p}(C_k|\mathbf{x})$ of the conditional probabilities $\Pr(C_k|\mathbf{x})$ of the consequent fuzzy sets C_k are given by (3.4) and (3.5). Using these estimates, the conditional pdf $p(y|\mathbf{x})$ can be estimated in a similar way

as in the fuzzy histograms that were discussed in Chapter 2. This results in (cf (2.26))

$$\hat{p}(y|\mathbf{x}) = \sum_{k=1}^{c} p(y|C_k)\hat{p}(C_k|\mathbf{x}), \tag{3.12}$$

where $p(y|C_k)$ denotes the conditional pdf of $\underline{y}$ given fuzzy event C_k. Usually, $p(y|C_k)$ will be unknown. In that case, the assumption is made that $p(y|C_k)$ is given by (cf (2.27))

$$p(y|C_k) = \frac{\mu_{C_k}(y)}{\int \mu_{C_k}(y)dy}. \tag{3.13}$$

It is important to note that in (3.12) and (3.13) (as well as in (2.26) and (2.27)) fuzzy histograms are given a different interpretation than in [16, 18, 34]. In the above mathematical notation, a PFS for estimating a conditional pdf relies on the assumption given by (3.13). Without this assumption (and the assumption that Takagi-Sugeno fuzzy reasoning is used for interpolation between rules), the estimate of a conditional pdf provided by a PFS does not follow deductively from the system's rule base. Of course, the approximation accuracy of a PFS is affected by the extent to which the assumption in (3.13) actually holds. Furthermore, remember that fuzzy histograms are very similar to double-kernel estimators, which are studied in [36]. It may be interesting to note that in Section 4.6 of [36] a similar assumption is made as in (3.13).

3.3.2 Estimation of the expectation and the variance of a stochastic function

This subsection is concerned with the application of PFSs to regression problems. The assumption is made that the function to be estimated, which is given by $f : X \to Y$, contains an error term. It follows from this assumption that $f(\mathbf{x})$ is stochastic. Also, it is assumed that the distribution of the error term of $f(\mathbf{x})$ is unknown and may depend on the value of $\mathbf{x}$. Given these assumptions, one may be interested in estimating, for each value of $\mathbf{x}$, both the expectation and the variance of $f(\mathbf{x})$. In this subsection, it is shown how estimates of the expectation and the variance of $f(\mathbf{x})$ can be derived from an estimated conditional pdf $\hat{p}(y|\mathbf{x})$ given by (3.12).

An estimate $\hat{f}(\mathbf{x})$ of the expectation of $f(\mathbf{x})$ can be derived as follows

$$\begin{aligned}
\hat{f}(\mathbf{x}) &= \hat{\mathrm{E}} f(\mathbf{x}) \\
&= \hat{\mathrm{E}}(\underline{y}|\mathbf{x}) \\
&= \int y \hat{p}(y|\mathbf{x}) dy \\
&= \int y \sum_{k=1}^{c} p(y|C_k) \hat{p}(C_k|\mathbf{x}) dy \\
&= \sum_{k=1}^{c} \hat{p}(C_k|\mathbf{x}) \int y p(y|C_k) dy \\
&= \sum_{k=1}^{c} \hat{p}(C_k|\mathbf{x}) \mathrm{E}(\underline{y}|C_k), \qquad (3.14)
\end{aligned}$$

where

$$\mathrm{E}(\underline{y}|C_k) = \int y p(y|C_k) dy = \frac{\int y \mu_{C_k}(y) dy}{\int \mu_{C_k}(y) dy}. \qquad (3.15)$$

The last step in (3.15) follows from the assumption in (3.13). Notice that under this assumption $\mathrm{E}(\underline{y}|C_k)$ is equal to the centroid of fuzzy set C_k.

The variance $\sigma^2(\mathbf{x})$ of $f(\mathbf{x})$ is given by

$$\sigma^2(\mathbf{x}) = \mathrm{E}(f(\mathbf{x}) - \mathrm{E}f(\mathbf{x}))^2 = \mathrm{E}f(\mathbf{x})^2 - (\mathrm{E}f(\mathbf{x}))^2. \qquad (3.16)$$

Consequently, an estimate $\hat{\sigma}^2(\mathbf{x})$ of the variance of $f(\mathbf{x})$ can be obtained as follows

$$\hat{\sigma}^2(\mathbf{x}) = \hat{\mathrm{E}} f(\mathbf{x})^2 - \left(\hat{\mathrm{E}} f(\mathbf{x})\right)^2 = \hat{\mathrm{E}} f(\mathbf{x})^2 - \hat{f}(\mathbf{x})^2. \qquad (3.17)$$

Using (3.12), the first term in (3.17) can be written as

$$\begin{aligned}
\hat{\mathrm{E}} f(\mathbf{x})^2 &= \hat{\mathrm{E}}\left(\underline{y}^2|\mathbf{x}\right) \\
&= \int y^2 \hat{p}(y|\mathbf{x}) dy \\
&= \int y^2 \sum_{k=1}^{c} p(y|C_k) \hat{p}(C_k|\mathbf{x}) dy \\
&= \sum_{k=1}^{c} \hat{p}(C_k|\mathbf{x}) \int y^2 p(y|C_k) dy. \qquad (3.18)
\end{aligned}$$

Finally, substitution of (3.18) in (3.17) gives

$$\hat{\sigma}^2(\mathbf{x}) = \left(\sum_{k=1}^{c} \hat{p}(C_k|\mathbf{x}) \int y^2 p(y|C_k) dy \right) - \hat{f}(\mathbf{x})^2. \qquad (3.19)$$

Chapter 4

Estimation of Probability Parameters

As discussed in Chapter 3, the rule base of a PFS consists of fuzzy rules that have multiple consequent parts. Each consequent part has an associated probability parameter. This chapter is concerned with the estimation of the probability parameters in a PFS. It is assumed in this chapter that both the antecedent and the consequent mfs have already been determined and need not be further optimized. (For classification problems, the estimation of the mfs is considered in Chapter 5.) In [16, 18, 34], probability parameters are estimated using a fuzzy generalization of the statistical formula for the estimation of conditional probabilities (see also [35]). This method for estimating probability parameters will be referred to as the conditional probability method in the remainder of this thesis. In this chapter, I show that the conditional probability method generally does not give optimal results in terms of the approximation accuracy of a PFS. As an alternative, I propose to use the maximum likelihood (ML) criterion for estimating the probability parameters in a PFS.

This chapter is organized as follows. In Section 4.1, the conditional probability method for estimating the probability parameters in a PFS is discussed. It is shown, both for classification problems and for regression problems, that probability parameters estimated using the conditional probability method are biased, asymptotically biased, and inconsistent and do not satisfy the ML criterion. In Section 4.2, a new method for estimating the probability parameters in a PFS is proposed. This method is based on the ML criterion. The properties of the optimization problem that results from the ML criterion are also considered in Section 4.2. Notice further that in the experiments described in Chapter 5 the conditional probability method and the ML method are compared empirically by applying both methods to a number of classification problems.

4.1 The conditional probability method

Let $(x_1, y_1), \ldots, (x_n, y_n)$ denote a random sample of size n. Using this sample, an estimate of $\Pr(C|A)$, the conditional probability of event C given event A, is provided by the following statistical formula

$$\hat{p}(C|A) = \frac{\sum_{i=1}^{n} \chi_A(x_i)\chi_C(y_i)}{\sum_{i=1}^{n} \chi_A(x_i)}, \tag{4.1}$$

where the characteristic functions χ_A and χ_C are given by

$$\chi_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

and

$$\chi_C(y) = \begin{cases} 1 & \text{if } y \in C \\ 0 & \text{otherwise.} \end{cases} \tag{4.3}$$

Now suppose that A and C are fuzzy events instead of ordinary crisp events. This means that A and C are defined by mfs μ_A and μ_C instead of characteristic functions χ_A and χ_C. Equation (4.1) can then be generalized by replacing the characteristic functions χ_A and χ_C by the mfs μ_A and μ_C. This results in

$$\hat{p}(C|A) = \frac{\sum_{i=1}^{n} \mu_A(x_i)\mu_C(y_i)}{\sum_{i=1}^{n} \mu_A(x_i)}. \tag{4.4}$$

This formula is based on Zadeh's definition of the probability of a fuzzy event [37]. A derivation of (4.4) can be found in [35, 36].

The result in (4.4) can be used for estimating the probability parameters in a PFS. This approach is followed in [16, 18, 34] and will be referred to as the conditional probability method in this thesis. Suppose that a data set containing n examples $(\mathbf{x}_i, y_i)$ $(i = 1, \ldots, n)$ is available for estimating the parameters in a PFS. Suppose further that both the antecedent and the consequent mfs in the system have already been determined and need not be further optimized. This means that only the probability parameters remain to be estimated. It seems reasonable to set a probability parameter $p_{j,k}$ equal to the estimated conditional probability of fuzzy event C_k given fuzzy event A_j. However, because the input space X need not be well-defined, instead of the unnormalized mf μ_{A_j} the normalized mf $\bar{\mu}_{A_j}$ must be used in the calculation of $p_{j,k}$. This gives

$$p_{j,k} = \frac{\sum_{i=1}^{n} \bar{\mu}_{A_j}(\mathbf{x}_i)\mu_{C_k}(y_i)}{\sum_{i=1}^{n} \bar{\mu}_{A_j}(\mathbf{x}_i)}. \tag{4.5}$$

Therefore, $p_{j,k}$ is actually set equal to the estimated conditional probability of fuzzy event C_k given the normalization of fuzzy event A_j. Notice further that PFSs for regression tasks that have a rule base with only one rule and

that use probability parameters $p_{j,k}$ given by (4.5) are equivalent to the fuzzy histograms discussed in Chapter 2.

In a PFS for classification tasks, in which C_k denotes a crisp output class, (4.5) can also be written as

$$p_{j,k} = \frac{\sum_{i=1}^{n} \bar{\mu}_{A_j}(\mathbf{x}_i)\chi_{C_k}(y_i)}{\sum_{i=1}^{n} \bar{\mu}_{A_j}(\mathbf{x}_i)}, \tag{4.6}$$

where the characteristic function χ_{C_k} is given by

$$\chi_{C_k}(y) = \begin{cases} 1 & \text{if } y = C_k \\ 0 & \text{otherwise.} \end{cases} \tag{4.7}$$

In the remainder of this section, the statistical properties of the parameter estimates in (4.5) and (4.6) are analyzed. Subsection 4.1.1 is concerned with PFSs for classification tasks. PFSs for regression tasks are considered in Subsection 4.1.2.

4.1.1 Statistical properties in classification problems

In this subsection, I prove that probability parameters estimated using (4.6) are biased, asymptotically biased, and inconsistent and do not satisfy the ML criterion. To prove this, it is sufficient to give a single example in which (4.6) provides estimates that are biased, asymptotically biased, and inconsistent and that do not maximize the likelihood of the available data set.

Consider a PFS that is applied to a classification problem in which there are two classes, denoted by C_1 and C_2. The PFS has an input space $X = [0, 1]$ and has a rule base that contains two probabilistic fuzzy rules. The mfs of the antecedent fuzzy sets A_1 and A_2 are given by

$$\mu_{A_1}(x) = 1 - x \qquad \text{and} \qquad \mu_{A_2}(x) = x. \tag{4.8}$$

It follows from (3.4) that $\bar{\mu}_{A_j} = \mu_{A_j}$ for $j = 1, 2$. Assume that the conditional probabilities of C_1 and C_2 equal

$$\Pr(C_1|x) = 1 - x \qquad \text{and} \qquad \Pr(C_2|x) = x. \tag{4.9}$$

These conditional probabilities are unknown and need to be estimated by the PFS. Using (3.5), it can be seen that in a PFS that correctly estimates the conditional probabilities in (4.9), the probability parameters are given by $p^*_{1,1} = p^*_{2,2} = 1$ and $p^*_{1,2} = p^*_{2,1} = 0$. (Notice that in this example the antecedent mfs in (4.8) have been chosen in such a way that it is possible to obtain a PFS that correctly estimates the conditional probabilities in (4.9). If it had not been possible to obtain a PFS that correctly estimates the conditional probabilities, then there would be no correct probability

parameters $p^*_{j,k}$ and, as a consequence, it would not be possible to analyze the bias, the asymptotic bias, and the consistency of estimates of the probability parameters.)

The following two theorems are concerned with the statistical properties of (4.6). To prove the theorems, I make use of the above example.

Theorem 4.1 *In a PFS for classification tasks, (4.6) provides estimates $p_{j,k}$ of the probability parameters $p^*_{j,k}$ that are biased, asymptotically biased, and inconsistent.*

Proof: Consider the example given above. Suppose that a data set containing n classification examples (x_i, y_i) $(i = 1, \ldots, n)$ is available for estimating the probability parameters in the PFS. For simplicity, assume that $x_1, \ldots, x_n$ have fixed values. This means that only $y_1, \ldots, y_n$ have to be treated as random variables. As an example, consider the estimate $p_{2,2}$ of the probability parameter $p^*_{2,2}$. From (4.6), (4.7), (4.8), and (4.9), it follows that

$$\begin{aligned} \mathrm{E}p_{2,2} &= \mathrm{E}\left(\frac{\sum_{i=1}^n \bar{\mu}_{A_2}(x_i)\chi_{C_2}(y_i)}{\sum_{i=1}^n \bar{\mu}_{A_2}(x_i)}\right) \\ &= \frac{\sum_{i=1}^n \bar{\mu}_{A_2}(x_i)\mathrm{E}(\chi_{C_2}(y_i))}{\sum_{i=1}^n \bar{\mu}_{A_2}(x_i)} \\ &= \frac{\sum_{i=1}^n x_i(0(1-x_i)+1x_i)}{\sum_{i=1}^n x_i} \\ &= \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i}. \end{aligned} \tag{4.10}$$

Now assume that $x_i \in (0, 1)$ for $i = 1, \ldots, n$. It then follows from (4.10) that $\mathrm{E}p_{2,2} \in (0, 1)$. Since $p^*_{2,2} = 1$, the estimate $p_{2,2}$ is biased. This argument holds independent of the number of classification examples n. Therefore, it also holds for $n \to \infty$, from which it follows that the estimate $p_{2,2}$ is also asymptotically biased.

Equation (4.6) provides consistent estimates if and only if for any positive number ϵ

$$\lim_{n\to\infty} \Pr(|p_{j,k} - p^*_{j,k}| \leq \epsilon) = 1, \tag{4.11}$$

where the estimate $p_{j,k}$ is obtained from a data set containing n classification examples. This condition can also be written as $\operatorname{plim} p_{j,k} = p^*_{j,k}$. A necessary condition for $\operatorname{plim} p_{j,k} = p^*_{j,k}$ is $\lim_{n\to\infty} \mathrm{E}p_{j,k} = p^*_{j,k}$ (see Theorem 2.9.39 in [12]), i.e. the estimate $p_{j,k}$ of $p^*_{j,k}$ must be asymptotically unbiased. However, it has already been proven that $p_{j,k}$ is an asymptotically biased estimate of $p^*_{j,k}$. It therefore follows that $p_{j,k}$ is also an inconsistent estimate of $p^*_{j,k}$. This completes the proof of the theorem.

Theorem 4.2 *Consider a PFS for classification tasks. Given a data set, the probability parameters $p_{j,k}$ estimated using (4.6) need not maximize the likelihood of the data set.*

x	0.0	0.5	0.5	1.0
y	C_1	C_1	C_2	C_2

Table 4.1: The data set that is used in the proof of Theorem 4.2.

Proof: Consider the example given above. Suppose that a data set containing four classification examples (x_i, y_i) $(i = 1, 2, 3, 4)$ is available for estimating the probability parameters in the PFS. The data set is displayed in Table 4.1. Substitution of the classification examples in (4.6) results in $p_{1,1} = p_{2,2} = 0.75$ and $p_{1,2} = p_{2,1} = 0.25$. It then follows from (3.5) that

$$\hat{p}(C_1|x) = 0.75 - 0.5x \qquad \text{and} \qquad \hat{p}(C_2|x) = 0.25 + 0.5x. \tag{4.12}$$

The likelihood of a data set is given by

$$L = \prod_{i=1}^{n} \hat{p}(y_i|x_i), \tag{4.13}$$

where it is assumed that the examples in the data set are independent of each other. For the probability parameters $p_{j,k}$ estimated using (4.6), it follows from (4.12) and (4.13) that the likelihood of the data set in Table 4.1 equals $9/64 \approx 0.14$. Now consider the alternative probability parameters $p'_{1,1} = p'_{2,2} = 1$ and $p'_{1,2} = p'_{2,1} = 0$. Using (3.5), these probability parameters result in

$$\hat{p}'(C_1|x) = 1 - x \qquad \text{and} \qquad \hat{p}'(C_2|x) = x. \tag{4.14}$$

For the alternative probability parameters $p'_{j,k}$, it follows from (4.13) and (4.14) that the likelihood of the data set in Table 4.1 equals 0.25. The alternative probability parameters therefore result in a higher value of the likelihood than the probability parameters $p_{j,k}$ estimated using (4.6). This example demonstrates that probability parameters estimated using (4.6) need not maximize the likelihood of a data set. (Actually, in the example it can be shown that the alternative probability parameters $p'_{j,k}$ maximize the likelihood of the data set. Of course, that the ML estimates of the probability parameters equal the correct probability parameters $p^*_{j,k}$ is merely a coincidence resulting from the specific data set in Table 4.1.) This completes the proof of the theorem.

It may be interesting to note that in a system in which the input space X is partitioned in a crisp way (i.e. $\bar{\mu}_{A_j}(\mathbf{x})$ equals, for $j = 1, \ldots, a$ and for all $\mathbf{x} \in X$, either 0 or 1), probability parameters estimated using (4.6) can be shown to be unbiased and consistent and to satisfy the ML criterion. (I do not prove this here.) Therefore, in such a crisp system it is possible to obtain parameter estimates with desirable statistical properties by estimating each parameter separately, as is done in (4.6). In a fuzzy system, however, it follows from Theorem 4.1 and 4.2 that parameter estimates with desirable statistical properties cannot be obtained by estimating each parameter

separately using (4.6). Instead, the parameters in a fuzzy system must be estimated simultaneously. Such an approach is proposed in Section 4.2.

4.1.2 Statistical properties in regression problems

In this subsection, I prove that probability parameters estimated using (4.5) are biased, asymptotically biased, and inconsistent and do not satisfy the ML criterion. To prove this, it is sufficient to give a single example in which (4.5) provides estimates that are biased, asymptotically biased, and inconsistent and that do not maximize the likelihood of the available data set. It should be noted that this subsection is very similar to the previous subsection. The only difference is that this subsection is concerned with PFSs for regression tasks instead of PFSs for classification tasks.

Consider a PFS that is applied to a regression problem. The PFS has an input space $X = [0, 1]$ and an output space $Y = [0, 1]$. The system's rule base contains two probabilistic fuzzy rules. The mfs of the antecedent fuzzy sets A_1 and A_2 are given by

$$\mu_{A_1}(x) = 1 - x \qquad \text{and} \qquad \mu_{A_2}(x) = x. \tag{4.15}$$

It follows from (3.4) that $\bar{\mu}_{A_j} = \mu_{A_j}$ for $j = 1, 2$. The output space Y is partitioned using two fuzzy sets, C_1 and C_2. The mfs of these fuzzy sets are given by

$$\mu_{C_1}(y) = 1 - y \qquad \text{and} \qquad \mu_{C_2}(y) = y. \tag{4.16}$$

Notice that the condition in (3.11) is satisfied, which means that Y is well-defined. Assume that the conditional pdf of $\underline{y}$ given x equals

$$p(y|x) = 4xy - 2x - 2y + 2. \tag{4.17}$$

This conditional pdf is unknown and needs to be estimated by the PFS. Using (3.5), (3.12), and (3.13), it can be seen that in a PFS that correctly estimates the conditional pdf in (4.17), the probability parameters are given by $p^*_{1,1} = p^*_{2,2} = 1$ and $p^*_{1,2} = p^*_{2,1} = 0$. (Notice that in this example the antecedent mfs in (4.15) and the consequent mfs in (4.16) have been chosen in such a way that it is possible to obtain a PFS that correctly estimates the conditional pdf in (4.17). If it had not been possible to obtain a PFS that correctly estimates the conditional pdf, then there would be no correct probability parameters $p^*_{j,k}$ and, as a consequence, it would not be possible to analyze the bias, the asymptotic bias, and the consistency of estimates of the probability parameters.)

The following two theorems are concerned with the statistical properties of (4.5). To prove the theorems, I make use of the above example.

Theorem 4.3 *In a PFS for regression tasks, (4.5) provides estimates $p_{j,k}$ of the probability parameters $p^*_{j,k}$ that are biased, asymptotically biased, and inconsistent.*

x	0.0	0.5	1.0
y	0.0	0.5	1.0

Table 4.2: The data set that is used in the proof of Theorem 4.4.

Proof: Consider the example given above. Suppose that a data set containing n examples (x_i, y_i) $(i = 1, \ldots, n)$ is available for estimating the probability parameters in the PFS. For simplicity, assume that $x_1, \ldots, x_n$ have fixed values. This means that only $y_1, \ldots, y_n$ have to be treated as random variables. As an example, consider the estimate $p_{2,2}$ of the probability parameter $p^*_{2,2}$. From (4.5), (4.15), (4.16), and (4.17), it follows that

$$\begin{aligned} \mathrm{E}p_{2,2} &= \mathrm{E}\left(\frac{\sum_{i=1}^n \bar{\mu}_{A_2}(x_i)\mu_{C_2}(y_i)}{\sum_{i=1}^n \bar{\mu}_{A_2}(x_i)}\right) \\ &= \frac{\sum_{i=1}^n \bar{\mu}_{A_2}(x_i)\mathrm{E}(\mu_{C_2}(y_i))}{\sum_{i=1}^n \bar{\mu}_{A_2}(x_i)} \\ &= \frac{\sum_{i=1}^n x_i \int_0^1 y p(y|x_i)dy}{\sum_{i=1}^n x_i} \\ &= \frac{\sum_{i=1}^n x_i \left(\frac{1}{3}x_i + \frac{1}{3}\right)}{\sum_{i=1}^n x_i}. \end{aligned} \tag{4.18}$$

Because $x_i \in [0, 1]$ for $i = 1, \ldots, n$, it follows from (4.18) that $\mathrm{E}p_{2,2} \in (1/3, 2/3]$. Since $p^*_{2,2} = 1$, the estimate $p_{2,2}$ is biased. This argument holds independent of the number of examples n. Therefore, it also holds for $n \to \infty$, from which it follows that the estimate $p_{2,2}$ is also asymptotically biased.

Equation (4.5) provides consistent estimates if and only if $\operatorname{plim} p_{j,k} = p^*_{j,k}$. A necessary condition for $\operatorname{plim} p_{j,k} = p^*_{j,k}$ is $\lim_{n\to\infty} \mathrm{E}p_{j,k} = p^*_{j,k}$ (see Theorem 2.9.39 in [12]), i.e. the estimate $p_{j,k}$ of $p^*_{j,k}$ must be asymptotically unbiased. However, it has already been proven that $p_{j,k}$ is an asymptotically biased estimate of $p^*_{j,k}$. It therefore follows that $p_{j,k}$ is also an inconsistent estimate of $p^*_{j,k}$. This completes the proof of the theorem.

Theorem 4.4 *Consider a PFS for regression tasks. Given a data set, the probability parameters $p_{j,k}$ estimated using (4.5) need not maximize the likelihood of the data set.*

Proof: Consider the example given above. Suppose that a data set containing three examples (x_i, y_i) $(i = 1, 2, 3)$ is available for estimating the probability parameters in the PFS. The data set is displayed in Table 4.2. Substitution of the examples in (4.5) results in $p_{1,1} = p_{2,2} = 5/6$ and $p_{1,2} = p_{2,1} = 1/6$. It then follows from (3.5), (3.12), and (3.13) that

$$\hat{p}(y|x) = \frac{1}{3}(8xy - 4x - 4y + 5). \tag{4.19}$$

The likelihood of a data set is given by (4.13). For the probability parameters $p_{j,k}$ estimated using (4.5), it follows from (4.13) and (4.19) that the likelihood of the data set in Table 4.2 equals $25/9 \approx 2.78$. Now consider the alternative probability parameters $p'_{1,1} = p'_{2,2} = 1$ and $p'_{1,2} = p'_{2,1} = 0$. Using (3.5), (3.12), and (3.13), these probability parameters result in

$$\hat{p}'(y|x) = 4xy - 2x - 2y + 2. \tag{4.20}$$

For the alternative probability parameters $p'_{j,k}$, it follows from (4.13) and (4.20) that the likelihood of the data set in Table 4.2 equals 4. The alternative probability parameters therefore result in a higher value of the likelihood than the probability parameters $p_{j,k}$ estimated using (4.5). This example demonstrates that probability parameters estimated using (4.5) need not maximize the likelihood of a data set. (Actually, in the example it can be shown that the alternative probability parameters $p'_{j,k}$ maximize the likelihood of the data set. Of course, that the ML estimates of the probability parameters equal the correct probability parameters $p^*_{j,k}$ is merely a coincidence resulting from the specific data set in Table 4.2.) This completes the proof of the theorem.

It may be interesting to note that in a system in which the input space X and the output space Y are partitioned in a crisp way (i.e. $\bar{\mu}_{A_j}(\mathbf{x})$ equals, for $j = 1, \ldots, a$ and for all $\mathbf{x} \in X$, either 0 or 1, and $\mu_{C_k}(y)$ equals, for $k = 1, \ldots, c$ and for all $y \in Y$, either 0 or 1), probability parameters estimated using (4.5) can be shown to be unbiased and consistent and to satisfy the ML criterion. (I do not prove this here.) Therefore, in such a crisp system it is possible to obtain parameter estimates with desirable statistical properties by estimating each parameter separately, as is done in (4.5). In a fuzzy system, however, it follows from Theorem 4.3 and 4.4 that parameter estimates with desirable statistical properties cannot be obtained by estimating each parameter separately using (4.5). Instead, the parameters in a fuzzy system must be estimated simultaneously. Such an approach is proposed in the next section.

4.2 The maximum likelihood method

In this section, I propose to use the ML criterion for estimating the probability parameters in a PFS. Contrary to the conditional probability method discussed in the previous section, all the probability parameters in a PFS are estimated simultaneously in this approach. It is assumed in this section that the antecedent and consequent mfs in a PFS have already been determined and need not be further optimized. (For classification problems, ML estimation of both the mfs and the probability parameters is considered in Section 5.2.) It should be noted that the focus of this section is on PFSs for estimating conditional pdfs. However, the results in this section also apply

to PFSs for classification tasks, since a PFS for classification tasks can be seen as a special case of a PFS for estimating conditional pdfs. Notice further that ML estimation of probability parameters has also been discussed in [26].

First notice that substitution of (3.5) in (3.12) results in

$$\hat{p}(y|\mathbf{x}) = \sum_{k=1}^{c} p(y|C_k) \sum_{j=1}^{a} \bar{\mu}_{A_j}(\mathbf{x})\, p_{j,k}. \tag{4.21}$$

The probability parameters $p_{j,k}$ in (4.21) must satisfy the conditions in (3.2) and (3.3). It follows from (3.3) that $p_{j,c} = 1 - \sum_{k=1}^{c-1} p_{j,k}$. Therefore, (4.21) can also be written as

$$\hat{p}(y|\mathbf{x}) = \left(\sum_{k=1}^{c-1} p(y|C_k) \sum_{j=1}^{a} \bar{\mu}_{A_j}(\mathbf{x})\, p_{j,k} \right) + p(y|C_c) \sum_{j=1}^{a} \bar{\mu}_{A_j}(\mathbf{x}) \left(1 - \sum_{k=1}^{c-1} p_{j,k} \right). \tag{4.22}$$

The probability parameters $p_{j,k}$ in (4.22) must satisfy

$$p_{j,k} \geq 0 \qquad \text{for } j = 1, \ldots, a \text{ and } k = 1, \ldots, c-1 \tag{4.23}$$

and

$$\sum_{k=1}^{c-1} p_{j,k} \leq 1 \qquad \text{for } j = 1, \ldots, a. \tag{4.24}$$

Now suppose that a data set containing n examples $(\mathbf{x}_i, y_i)$ $(i = 1, \ldots, n)$ is available for estimating the probability parameters in a PFS. Given a parameter matrix $\mathbf{P} = [p_{j,k}]$ $(j = 1, \ldots, a$ and $k = 1, \ldots, c-1)$, the likelihood of a data set equals

$$L(\mathbf{P}) = \prod_{i=1}^{n} \hat{p}(y_i|\mathbf{x}_i), \tag{4.25}$$

where it is assumed that the examples in the data set are independent of each other. The probability parameters in $\mathbf{P}$ must satisfy the conditions in (4.23) and (4.24). Using (4.22) and (4.25), the log-likelihood can be written as

$$\begin{aligned} l(\mathbf{P}) &= \ln \left(\prod_{i=1}^{n} \hat{p}(y_i|\mathbf{x}_i) \right) \\ &= \sum_{i=1}^{n} \ln \hat{p}(y_i|\mathbf{x}_i) \\ &= \sum_{i=1}^{n} \ln \left(\left(\sum_{k=1}^{c-1} p(y_i|C_k) \sum_{j=1}^{a} \bar{\mu}_{A_j}(\mathbf{x}_i)\, p_{j,k} \right) \right. \\ &\qquad \left. + p(y_i|C_c) \sum_{j=1}^{a} \bar{\mu}_{A_j}(\mathbf{x}_i) \left(1 - \sum_{k=1}^{c-1} p_{j,k} \right) \right). \end{aligned} \tag{4.26}$$

ML estimates of the probability parameters in a PFS are obtained by maximizing the log-likelihood function $l(\mathbf{P})$ in (4.26) with respect to the parameter matrix $\mathbf{P}$. Maximization of $l(\mathbf{P})$ is constrained by the conditions in (4.23) and (4.24). Since $l(\mathbf{P})$ is a nonlinear function, finding ML estimates of probability parameters is a nonlinear programming problem.

Now consider the following theorem.

Theorem 4.5 *The log-likelihood function $l(\mathbf{P})$ in (4.26) is concave.*

Proof: Since a sum of concave functions is concave, it is sufficient to prove that

$$\begin{aligned}\phi(\mathbf{P}) &= \ln \hat{p}(y|\mathbf{x}) \\ &= \ln \left(\left(\sum_{k=1}^{c-1} p(y|C_k) \sum_{j=1}^{a} \bar{\mu}_{A_j}(\mathbf{x})\, p_{j,k} \right) \right. \\ &\qquad \left. + p(y|C_c) \sum_{j=1}^{a} \bar{\mu}_{A_j}(\mathbf{x}) \left(1 - \sum_{k=1}^{c-1} p_{j,k} \right) \right) \end{aligned} \tag{4.27}$$

is a concave function for all $\mathbf{x} \in X$ and all $y \in Y$. Notice that $\phi(\mathbf{P})$ is defined for any parameter matrix $\mathbf{P}$ for which $\hat{p}(y|\mathbf{x}) > 0$. Furthermore, $\phi(\mathbf{P})$ is twice continuously differentiable. The first-order and second-order partial derivatives of $\phi(\mathbf{P})$ are given by

$$\frac{\partial \phi(\mathbf{P})}{\partial p_{\alpha,\gamma}} = \frac{\bar{\mu}_{A_\alpha}(\mathbf{x})(p(y|C_\gamma) - p(y|C_c))}{\hat{p}(y|\mathbf{x})} \tag{4.28}$$

and

$$\frac{\partial^2 \phi(\mathbf{P})}{\partial p_{\alpha,\gamma}\, \partial p_{\beta,\delta}} = -\frac{\bar{\mu}_{A_\alpha}(\mathbf{x})(p(y|C_\gamma) - p(y|C_c))\bar{\mu}_{A_\beta}(\mathbf{x})(p(y|C_\delta) - p(y|C_c))}{\hat{p}(y|\mathbf{x})^2}, \tag{4.29}$$

where $\alpha, \beta = 1, \ldots, a$ and $\gamma, \delta = 1, \ldots, c-1$. From (4.29), it follows that for any parameter matrix $\mathbf{P}$ for which $\phi(\mathbf{P})$ is defined

$$\frac{\partial^2 \phi(\mathbf{P})}{\partial p_{\alpha,\gamma}^2} = -\left(\frac{\bar{\mu}_{A_\alpha}(\mathbf{x})(p(y|C_\gamma) - p(y|C_c))}{\hat{p}(y|\mathbf{x})} \right)^2 \leq 0. \tag{4.30}$$

Let D_m denote an $m \times m$ determinant given by

$$D_m = \begin{vmatrix} \dfrac{\partial^2 \phi(\mathbf{P})}{\partial p_{\alpha_1,\gamma_1}\, \partial p_{\beta_1,\delta_1}} & \cdots & \dfrac{\partial^2 \phi(\mathbf{P})}{\partial p_{\alpha_1,\gamma_1}\, \partial p_{\beta_m,\delta_m}} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial^2 \phi(\mathbf{P})}{\partial p_{\alpha_m,\gamma_m}\, \partial p_{\beta_1,\delta_1}} & \cdots & \dfrac{\partial^2 \phi(\mathbf{P})}{\partial p_{\alpha_m,\gamma_m}\, \partial p_{\beta_m,\delta_m}} \end{vmatrix}, \tag{4.31}$$

where, for $q \neq r$, $(\alpha_q, \gamma_q) \neq (\alpha_r, \gamma_r)$ and $(\beta_q, \delta_q) \neq (\beta_r, \delta_r)$. Using (4.29), it can be seen that for any determinant D_m with $m = 2$ and for any parameter matrix $\mathbf{P}$ for which $\phi(\mathbf{P})$ is defined

$$D_2 = \frac{\partial^2 \phi(\mathbf{P})}{\partial p_{\alpha_1,\gamma_1} \partial p_{\beta_1,\delta_1}} \frac{\partial^2 \phi(\mathbf{P})}{\partial p_{\alpha_2,\gamma_2} \partial p_{\beta_2,\delta_2}} - \frac{\partial^2 \phi(\mathbf{P})}{\partial p_{\alpha_2,\gamma_2} \partial p_{\beta_1,\delta_1}} \frac{\partial^2 \phi(\mathbf{P})}{\partial p_{\alpha_1,\gamma_1} \partial p_{\beta_2,\delta_2}} = 0. \tag{4.32}$$

By applying Laplace determinant expansion, it follows from (4.32) that D_m also equals 0 for any determinant with $m > 2$. Combined with (4.30), this result implies that the Hessian matrix of $\phi(\mathbf{P})$ is negative semidefinite everywhere (see Theorem 1.E.11 in [32]) and, consequently, that $\phi(\mathbf{P})$ is a concave function. This completes the proof of the theorem. Notice that the theorem is quite general, in the sense that it makes no assumptions about the mfs μ_{A_j} in a PFS.

In the nonlinear programming problem of finding ML estimates of the probability parameters in a PFS, the functions in the constraints, given by (4.23) and (4.24), are all linear, from which it follows that these functions are all convex. Since, according to Theorem 4.5, the objective function is concave, the nonlinear programming problem is actually a convex programming problem. Convex programming problems have the convenient property that each local optimum is also a global optimum. Therefore, finding a globally optimal solution should be relatively easy.

Chapter 5

Applications to Classification Problems

This chapter is concerned with PFSs that are applied to classification problems. A new method for estimating the parameters in these systems is proposed. This method can be seen as an extension of the ML method for estimating the probability parameters in a PFS, which was introduced in Section 4.2. The method proposed in this chapter uses the ML criterion for estimating both the probability parameters and the antecedent parameters in a PFS (instead of only the probability parameters, as in Section 4.2). Maximization of the likelihood function is performed using a gradient-based optimization algorithm.

In the experiments described in this chapter, the performance of the proposed method is compared with the performance of two heuristic methods for parameter estimation. The first heuristic method is the conditional probability method, which is used for estimating the probability parameters in a PFS. This method was discussed in Section 4.1. The second heuristic method, which makes use of fuzzy c-means (FCM) clustering, estimates the antecedent parameters in a PFS. Notice that the heuristic methods are complementary to each other, since the first method only estimates probability parameters and the second method only estimates antecedent parameters. Two sets of experiments are described in this chapter. In the first set of experiments, PFSs are applied to the Wisconsin breast cancer data set and the wine data set, which are both taken from the UCI machine learning repository [3]. The results of these experiments are compared with the results that are reported in [1], where a supervised clustering algorithm is used for estimating the parameters in a probabilistic fuzzy classifier. In the second set of experiments, PFSs are applied to a target selection problem. The results of these experiments are compared with the results that are reported in [17]. It should be noted that the issue of finding the optimal number of rules in a PFS is not considered in the experiments in this chapter.

This chapter is organized as follows. In Section 5.1, the use of FCM clustering for estimating the antecedent parameters in a PFS is considered. In Section 5.2, a ML method is proposed for estimating both the probability parameters and the antecedent parameters in a PFS. Experiments with the Wisconsin breast cancer data set and the wine data set are described in Section 5.3, and experiments with a target selection problem are described in Section 5.4.

5.1 Antecedent parameter estimation using the fuzzy c-means method

In this chapter, it is assumed that each antecedent mf μ_{A_j} in a PFS is the product of d univariate Gaussian mfs $\phi(x) = \exp\left(-(x-c)^2/\sigma^2\right)$, one for each dimension of the input space X. This results in

$$\mu_{A_j}(\mathbf{x}) = \exp\left(-\sum_{l=1}^{d} \frac{(x_l - c_{j,l})^2}{\sigma_{j,l}^2}\right). \tag{5.1}$$

For each mf, the parameters that need to be estimated are given by a vector $\mathbf{c}_j = \{c_{j,1}, \ldots, c_{j,d}\}$ and a vector $\mathbf{\Sigma}_j = \{\sigma_{j,1}, \ldots, \sigma_{j,d}\}$. These vectors indicate, respectively, the center and the width of the mf in each dimension of the input space. Furthermore, it is assumed in this chapter that a data set containing n classification examples $(\mathbf{x}_i, y_i)$ $(i = 1, \ldots, n)$ is available for estimating the parameters in a PFS.

In this section, the use of the FCM algorithm for estimating the antecedent parameters $\mathbf{c}_j$ and $\mathbf{\Sigma}_j$ in a PFS is discussed. In this approach, the data set available for parameter estimation is first normalized. For the lth feature $(l = 1, \ldots, d)$ of a data point $\mathbf{x}_i$ $(i = 1, \ldots, n)$, this is done according to

$$\bar{x}_{i,l} = \frac{x_{i,l} - \mu_l}{\sigma_l}, \tag{5.2}$$

where μ_l and σ_l denote, respectively, the mean and the standard deviation of the lth feature over the entire data set. The FCM algorithm is then applied to the normalized data points $\bar{\mathbf{x}}_i$ in order to identify a predefined number of cluster centers. The FCM algorithm uses the standard Euclidean distance measure. The cluster centers obtained using FCM serve as the centers $\mathbf{c}_j$ of the Gaussian mfs μ_{A_j}. The vectors $\mathbf{\Sigma}_j$, which contain the widths of the mfs, then remain to be estimated. In the experiments in this chapter, the *nearest neighbor heuristic* [13] is used for estimating these vectors. This results in

$$\sigma_{j,l} = \min_{j' \neq j} \|\mathbf{c}_j - \mathbf{c}_{j'}\| \qquad \text{for } l = 1, \ldots, d, \tag{5.3}$$

where $\|\mathbf{c}_j - \mathbf{c}_{j'}\|$ denotes the Euclidean distance between $\mathbf{c}_j$ and $\mathbf{c}_{j'}$. Notice that an mf is given the same width in each dimension according to this heuristic.

It should be noted that the FCM algorithm and other unsupervised clustering algorithms generally do not provide optimal estimates of the antecedent parameters in a fuzzy system. In the PFSs discussed in this chapter, this is because FCM does not take the class labels y_i into account. An alternative approach is proposed in [1], where a *supervised* clustering algorithm is used for estimating the parameters in a probabilistic fuzzy classifier. Contrary to unsupervised algorithms, this algorithm takes class labels into account during clustering.

5.2 Parameter estimation using the maximum likelihood method

In this section, I propose a ML method for estimating the parameters in a PFS for classification tasks. Compared with the conditional probability method described in Section 4.1 and the FCM method described in Section 5.1, the proposed method has the advantage that it maximizes the likelihood of the data set available for parameter estimation and that it simultaneously estimates both the antecedent parameters $\mathbf{c}_j$ and $\mathbf{\Sigma}_j$ and the probability parameters $p_{j,k}$. Notice that the ML method discussed in Section 4.2 only estimates the probability parameters $p_{j,k}$.

The likelihood of a data set is given by

$$L = \prod_{i=1}^{n} \hat{p}(y_i|\mathbf{x}_i), \tag{5.4}$$

where it is assumed that the classification examples in the data set are independent of each other. Maximization of the likelihood is equivalent to minimization of the negative log-likelihood. I therefore choose to minimize the following error function

$$E = -\sum_{i=1}^{n} \ln \hat{p}(y_i|\mathbf{x}_i). \tag{5.5}$$

Notice that minimizing (5.5) is similar to minimizing the cross entropy error function that is used for training neural network classifiers [2].

Finding the parameters $\mathbf{c}_j$, $\mathbf{\Sigma}_j$, and $p_{j,k}$ that minimize the error function in (5.5) is a constrained optimization problem, since the probability parameters $p_{j,k}$ must satisfy the conditions in (3.2) and (3.3). This constrained optimization problem can be converted into an unconstrained optimization problem by using the auxiliary variables $u_{j,k}$ ($j = 1, \ldots, a$ and $k = 1, \ldots, c$). The relation between these variables and the probability parameters $p_{j,k}$ is described by the softmax function, i.e.

$$p_{j,k} = \frac{e^{u_{j,k}}}{\sum_{k'=1}^{c} e^{u_{j,k'}}}. \tag{5.6}$$

ML estimates of the parameters $\mathbf{c}_j$, $\mathbf{\Sigma}_j$, and $p_{j,k}$ can be obtained by unconstrained minimization of (5.5) with respect to $\mathbf{c}_j$, $\mathbf{\Sigma}_j$, and $u_{j,k}$. I choose not to optimize the variables $u_{j,c}$ ($j = 1, \ldots, a$). These variables are given a fixed value of 0. In this way, the size of the optimization problem is reduced, while the variables $u_{j,k}$ can still represent all possible solutions for the parameters $p_{j,k}$.

For minimizing the error function in (5.5), a gradient descent optimization algorithm is used. The stochastic variant of gradient descent is applied, which means that the available classification examples are processed one by one and that updates are performed after each example. The initial values of the antecedent parameters $\mathbf{c}_j$ and $\mathbf{\Sigma}_j$ are obtained using the FCM method described in Section 5.1, and the initial values of the probability parameters $p_{j,k}$ are obtained using the conditional probability method described in Section 4.1.

5.3 Application to the breast cancer data set and the wine data set

In order to compare the performance of the ML method (Section 5.2) with the performance of the FCM method (Section 5.1) and the conditional probability method (Section 4.1), experiments were performed with the Wisconsin breast cancer data set and the wine data set. These data sets were taken from the UCI machine learning repository [3]. In the breast cancer data set, the problem is to classify cancers as benign or malignant. Classifications are based on nine features that have integer values between one and ten. The breast cancer data set contains 683 classification examples (excluding 16 examples that have missing values). The problem in the wine data set is to distinguish three types of wine. A wine is characterized by 13 continuous features. The wine data set contains 178 examples. In the experiments, the breast cancer data set and the wine data set were normalized using (5.2).

5.3.1 Setup of the experiments

Four different types of experiments were performed. The differences between these types lie in the way in which the parameters in a PFS are estimated. In the first type of experiment, the antecedent parameters are estimated using the FCM method and the probability parameters are estimated using the conditional probability method. In the second type of experiment, the FCM method is again used for estimating the antecedent parameters. However, the probability parameters are estimated using the ML method. In the third type of experiment, the ML method is used for estimating the antecedent parameters. After each iteration of the gradient descent algorithm, the probability parameters are re-estimated using the conditional probabil-

ity method. Finally, in the fourth type of experiment, all parameters are estimated using the ML method.

In the experiments, the weighting exponent of the FCM algorithm, which determines the degree of fuzziness of the clustering, was given a value of 2. This is a typical value for this parameter. In each experiment, the FCM algorithm was run several times using different initial values. Eventually, the clustering with the lowest value of the objective function was taken. In the gradient descent algorithm that was used in the ML method, the number of iterations was set to 300 and a fixed value of 0.03 was used for the learning rate. Using these parameter settings, it was found that the gradient descent algorithm had always converged at the end of an experiment. The PFS obtained after 300 iterations was used for performance evaluation. Furthermore, in all experiments the constraint $\sigma_{j,l} \geq 0.25$ ($j = 1, \ldots, a$ and $l = 1, \ldots, d$) was imposed. This was done to avoid numerical stability problems.

The performance of a PFS was evaluated using two different error functions, denoted by E_1 and E_2. The first error function is given by

$$E_1 = \frac{n_{\text{errors}}}{n}, \tag{5.7}$$

where n_{errors} is equal to the number of examples that are misclassified. This is the usual error function for classification problems. The second error function is a normalized version of (5.5), i.e.

$$E_2 = -\frac{1}{n} \sum_{i=1}^{n} \ln \hat{p}(y_i|\mathbf{x}_i). \tag{5.8}$$

In the same way as (5.5), this error function is derived from the ML criterion.

Two different motivations can be given for error function E_2 in (5.8). Consider, for example, a system for diagnosing the presence or absence of a disease based on certain indicators of a patient's condition. Some indicators that are relevant to the diagnosis of the disease may not be available to the system. In that case, the relation between the indicators that are available and the correct diagnosis will be stochastic. As a consequence of the stochasticity, it will be impossible to obtain a system that always gives the correct diagnosis. Instead, one may want to obtain a system that provides estimates of the conditional probabilities of the different diagnoses. Since an estimate of a (conditional) probability distribution is typically evaluated using the ML criterion, the performance of such a system can be assessed in an appropriate way using error function E_2, which is derived from this criterion. In the literature [8], error function E_2 is also applied to the similar task of evaluating estimates of conditional probability density functions.

The second motivation for error function E_2 applies even when all indicators that are relevant to the diagnosis of the disease are available to

Estimation method		2 rules		4 rules	
$\mathbf{c}_j$ and $\mathbf{\Sigma}_j$	$p_{j,k}$	E_1	E_2	E_1	E_2
FCM	Eq. (4.6)	0.261 (0.036)	0.497 (0.014)	0.291 (0.028)	0.578 (0.034)
FCM	ML	0.050 (0.039)	0.339 (0.024)	0.296 (0.030)	0.563 (0.045)
ML	Eq. (4.6)	0.034 (0.025)	0.102 (0.047)	0.032 (0.023)	0.109 (0.050)
ML	ML	0.029 (0.021)	0.100 (0.060)	0.037 (0.024)	0.101 (0.060)

Table 5.1: Results for the Wisconsin breast cancer data set. The results are averages from a ten-fold cross-validation. Standard deviations are reported within parentheses.

Estimation method		3 rules		6 rules	
$\mathbf{c}_j$ and $\mathbf{\Sigma}_j$	$p_{j,k}$	E_1	E_2	E_1	E_2
FCM	Eq. (4.6)	0.034 (0.048)	0.683 (0.031)	0.068 (0.065)	0.168 (0.152)
FCM	ML	0.057 (0.075)	0.451 (0.053)	0.079 (0.081)	0.167 (0.152)
ML	Eq. (4.6)	0.028 (0.041)	0.168 (0.318)	0.028 (0.047)	0.149 (0.275)
ML	ML	0.023 (0.041)	0.121 (0.220)	0.034 (0.039)	0.079 (0.091)

Table 5.2: Results for the wine data set. The results are averages from a ten-fold cross-validation. Standard deviations are reported within parentheses.

the system. This motivation follows from the fact that only a finite set of example diagnoses will be available for training the system, implying that the correctness of a diagnosis given by the system generally cannot be guaranteed. One may therefore be interested to know the degree of confidence the system attaches to a diagnosis that it provides [23]. When the degree of confidence is expressed in terms of the probability that the diagnosis is correct, error function E_2 can be used for evaluating the accuracy of the confidence measure. The underlying idea is that an incorrect diagnosis with a probability of 0.6 attached to it is less problematic than an incorrect diagnosis with a probability of 0.9 attached to it. Conversely, a correct diagnosis is more valuable when a probability of 0.9 was attached to it than when a probability of 0.6 was attached to it.

5.3.2 Results of the experiments

The results of the experiments are reported in Table 5.1 for the Wisconsin breast cancer data set and in Table 5.2 for the wine data set. For each data set, two different values were used for the number of rules in a PFS. The results in Table 5.1 and 5.2 were obtained using ten-fold cross-validation. The splitting of a data set into ten subsets was done in such a way that the distribution of the classes was approximately the same in each subset.

First consider the estimation of the parameters $\mathbf{c}_j$ and $\mathbf{\Sigma}_j$ of the antecedent mfs in a PFS. In all experiments, the ML method performed better than the FCM method. This result holds for both error functions. Of course,

Experiment	Ref. [1]	ML
Breast cancer, 2 rules	0.074	0.029
Breast cancer, 4 rules	0.044	0.037
Wine, 3 rules	0.022	0.023

Table 5.3: Comparison between E_1 errors reported in [1] and E_1 errors of the ML method. The errors are averages from a ten-fold cross-validation.

it is not surprising that the ML method performed better, since the FCM method is unsupervised. However, the results clearly indicate the substantial performance improvement that can be realized by using a supervised method for estimating the antecedent parameters. On the other hand, it should also be noted that the performance of the FCM method may depend strongly on the heuristic that is used for estimating the parameters $\mathbf{\Sigma}_j$. It may be possible to improve the performance of the FCM method by using a different heuristic than the nearest neighbor heuristic defined by (5.3).

Concerning the estimation of the probability parameters $p_{j,k}$ in a PFS, the error functions E_1 and E_2 sometimes give contradictory results. Error function E_2 indicates that in all experiments the ML method performed better than the conditional probability method, i.e. the method that uses (4.6) for estimating probability parameters. However, in some experiments the difference between the two methods is negligible. Error function E_1 does not give conclusive results. According to this error function, the conditional probability method performed better in some experiments whereas the ML method performed better in other experiments. Notice that in the breast cancer experiments with two rules and with FCM estimation of the antecedent parameters, the ML method realized a large performance improvement compared with the conditional probability method. This result suggests that at least in some cases the ML method may be preferable to the conditional probability method.

The results of the experiments can be compared with the results that are reported in [1], where a probabilistic fuzzy classifier is studied that is very similar to a PFS for classification tasks. Also, a supervised clustering algorithm for estimating the parameters in the probabilistic fuzzy classifier is proposed in [1]. In the experiments described in [1], the proposed algorithm is tested on the Wisconsin breast cancer data set and the wine data set. For performance evaluation, ten-fold cross-validation is used in combination with error function E_1. The results are shown in Table 5.3. For comparison, the results of ML estimation of all the parameters in a PFS are also shown in the table. In two experiments, the supervised clustering algorithm and the ML method have comparable performance. However, in the breast cancer experiment with two rules, the ML method has a substantially lower error than the supervised clustering algorithm. In this rather limited comparison, the ML method therefore performs somewhat better than the supervised

clustering algorithm.

5.4 Application to a target selection problem

In this section, experiments are discussed in which PFSs are applied to a target selection problem. The problem is to predict whether a customer, who is described by a number of features, will respond to a mailing concerning an offer for a product. This is a classification problem with two classes, the class of responders and the class of non-responders. By estimating each customer's probability of response, a company can choose to send an offer to its high-prospect customers only instead of to all its customers. In that way, the size of a mailing is reduced, which should result in an increase of the company's profit.

One approach to obtain a target selection model is to use a PFS for classification tasks. This approach results in a model that can be interpreted linguistically and that provides an estimate of the probability that a customer will be a responder. The approach of modeling target selection problems using PFSs is taken in [14, 15, 17]. In these papers, the parameters of the antecedent mfs in a PFS are estimated using an unsupervised clustering algorithm. The probability parameters in a PFS are estimated using the conditional probability method (Section 4.1). In this section, experiments with parameter estimation using the ML method (Section 5.2) are discussed. The results of these experiments are compared with the results that are reported in [17].

In the experiments in this section, a data set is used that has been obtained from the mailing campaigns of a charity organization. Charity organizations use target selection for selecting people that are more likely to donate money. In that way, organizations try to maximize their fund raising results. The data set that is used in the experiments consists of a training set of 4057 examples and an independent validation set of 4080 examples. Each example is described by the following three features:

1. Number of weeks since last response.
2. Number of months as a supporter.
3. Fraction of mailings responded.

For more details about the data set, the reader is referred to [24]. In the experiments in this section, the data set was normalized using (5.2).

5.4.1 Setup of the experiments

The same four types of experiments were performed as in Section 5.3. Therefore, the antecedent parameters in a PFS were estimated using either the

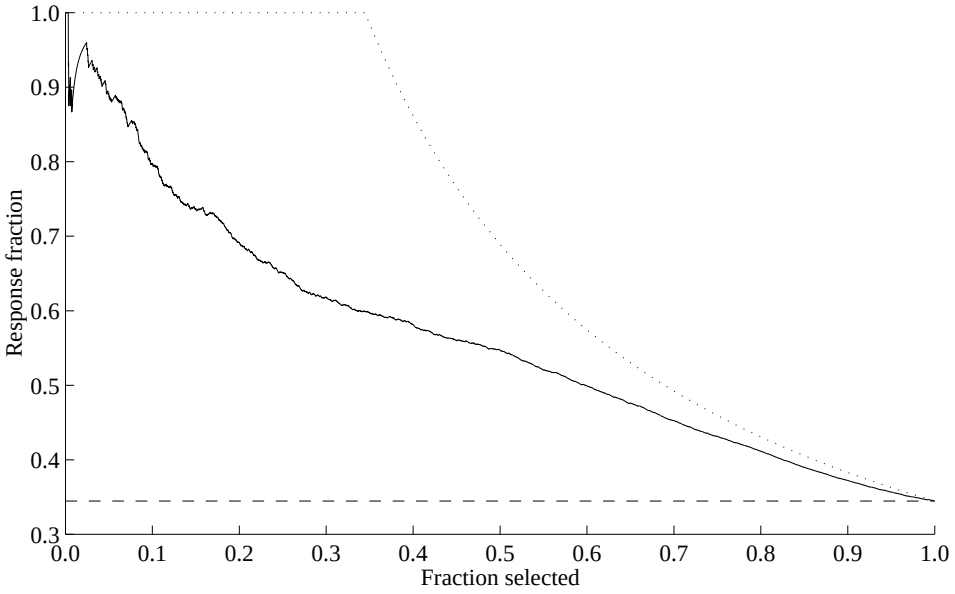

Figure 5.1: An example of a hit probability chart. The chart shows the performance of a typical target selection model (solid line), an optimal target selection model (dotted line), and a random target selection model (dashed line).

FCM method or the ML method, and the probability parameters in a PFS were estimated using either the conditional probability method or the ML method.

Like in the experiments in Section 5.3, the weighting exponent of the FCM algorithm was given a value of 2 and a fixed value of 0.03 was used for the learning rate of the gradient descent algorithm. The number of iterations of the gradient descent algorithm was set to 50, which turned out to be sufficient for convergence. Furthermore, like in Section 5.3, the constraint $\sigma_{j,l} \geq 0.25$ ($j = 1, \ldots, a$ and $l = 1, \ldots, d$) was imposed.

The performance of a PFS was evaluated using a hit probability chart. A hit probability chart shows the percentage of the mailed customers that is a responder as a function of the percentage of the customers that is mailed. An example of a hit probability chart is displayed in Figure 5.1. (Notice that in Figure 3 in [17] the hit probability chart of an optimal target selection model is drawn incorrectly.) In general, a larger area under a hit probability chart indicates a better performance of a target selection model.

To quantify the performance of a PFS, a performance index was used in the experiments. This performance index is defined as

$$I = \frac{p^* - A}{p^* \ln p^*}, \tag{5.9}$$

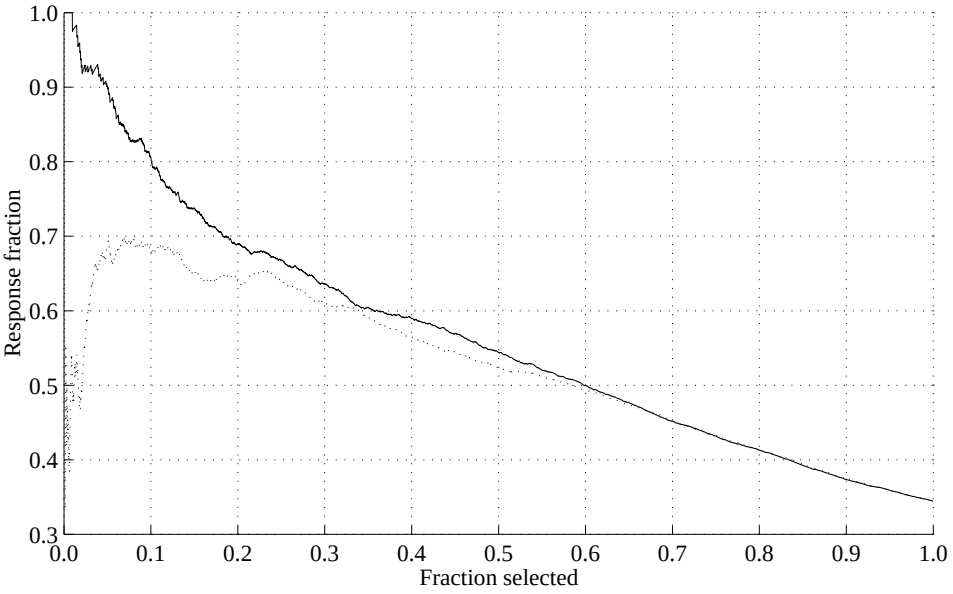

Figure 5.2: Hit probability chart for a PFS with 5 rules. The chart shows the performance of the ML method (solid line) and of the FCM method combined with the conditional probability method (dotted line).

Estimation method		Number of rules		
$\mathbf{c}_j$ and $\mathbf{\Sigma}_j$	$p_{j,k}$	5	10	15
FCM	Eq. (4.6)	0.485	0.520	0.578
FCM	ML	0.487	0.576	0.584
ML	Eq. (4.6)	0.584	0.591	0.590
ML	ML	0.600	0.602	0.590

Table 5.4: Results of the target selection experiments calculated using the performance index in (5.9).

where p^* denotes the fraction of responders in the entire data set and A denotes the area under the hit probability chart. The performance index in (5.9) equals 1 for an optimal target selection model and equals 0 for a random target selection model. (Notice that the performance index in Equation (12) in [17] is defined incorrectly, since it does not equal 1 for an optimal target selection model.)

5.4.2 Results of the experiments

Experiments were performed with 5, 10, and 15 rules in a PFS. Hit probability charts for these experiments are displayed in Figure 5.2, 5.3, and 5.4. The charts show the results of two types of experiments: experiments in which the antecedent parameters were estimated using the FCM method

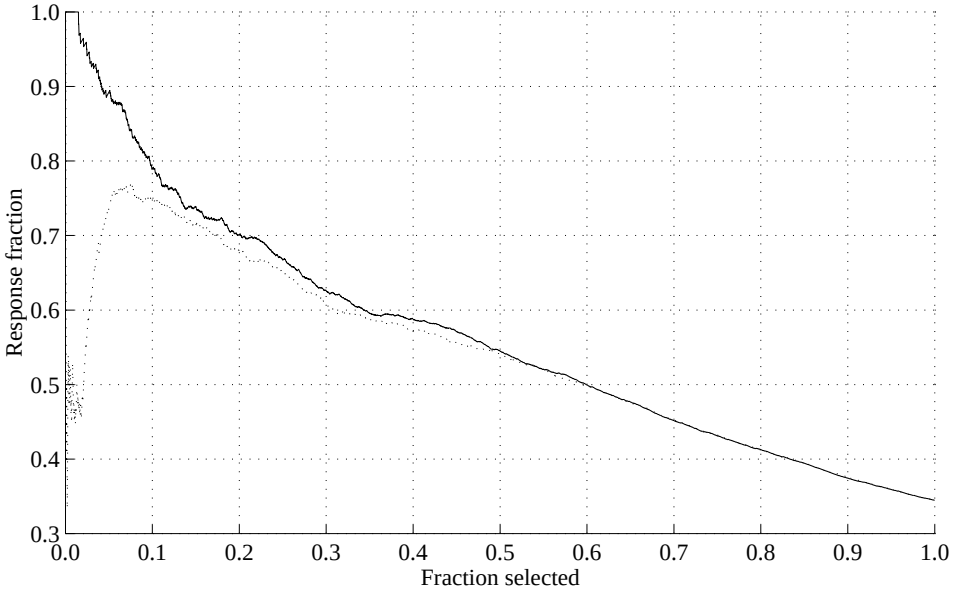

Figure 5.3: Hit probability chart for a PFS with 10 rules. The chart shows the performance of the ML method (solid line) and of the FCM method combined with the conditional probability method (dotted line).

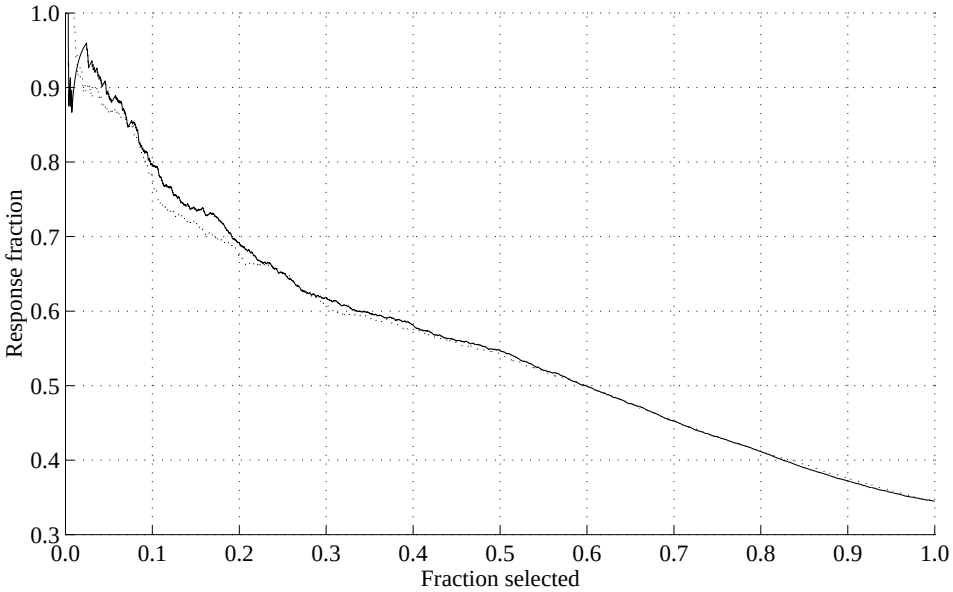

Figure 5.4: Hit probability chart for a PFS with 15 rules. The chart shows the performance of the ML method (solid line) and of the FCM method combined with the conditional probability method (dotted line).

and the probability parameters were estimated using the conditional probability method, and experiments in which all parameters were estimated using the ML method. For all four types of experiments that were performed, the value of the performance index defined by (5.9) is reported in Table 5.4. All results were obtained using the independent validation set.

The hit probability charts show that the ML method performed much better than the combination of the FCM method and the conditional probability method. This was especially the case when the number of rules in a PFS was small. The results in Table 5.4 indicate that the largest performance improvement was obtained by estimating the parameters $\mathbf{c}_j$ and $\mathbf{\Sigma}_j$ of the antecedent mfs using the ML method instead of the FCM method. In most cases, the performance improvement obtained by estimating the probability parameters $p_{j,k}$ using the ML method instead of the conditional probability method was relatively small. Notice that these outcomes are similar to the outcomes of the experiments discussed in Section 5.3. Furthermore, it is interesting to observe that when the ML method was used, a simple target selection model consisting of only 5 probabilistic fuzzy rules was sufficient to obtain a good performance. Such a model performed considerably better than a much more complex model consisting of 15 rules in which the parameters were estimated using the FCM method and the conditional probability method.

The results of the experiments can be compared with the results that are reported in [17]. In the experiments in [17], the same target selection data set is considered as in this thesis and a PFS with 15 rules is used as target selection model. The system's antecedent parameters are estimated using a weighted fuzzy clustering algorithm, and the system's probability parameters are estimated using the conditional probability method. A number of target selection models are obtained in the experiments in [17]. I only consider the model that gives the best performance. The hit probability chart of this model is shown in Figure 6 in [17]. The value of the performance index defined by (5.9) equals 0.573 for this model. (The value reported in Table I in [17] has been calculated using a different performance index and must therefore be multiplied by 1.2.) By comparing this value with the values in the rightmost column of Table 5.4, it can be seen that the best model in [17] gives approximately the same performance as a PFS in which the antecedent parameters are estimated using the FCM method and the probability parameters are estimated using the conditional probability method. However, compared with a PFS in which all the parameters are estimated using the ML method, the best model in [17] gives worse performance. This can also be observed by comparing the hit probability charts in Figure 6 in [17] and in Figure 5.4 in this thesis. Especially when less than 10 percent of the customers is mailed, a PFS with ML parameter estimates performs much better than the best model in [17].

Chapter 6

Probabilistic Fuzzy Modeling in Regression Problems

This chapter is concerned with probabilistic fuzzy modeling in regression problems. As an example, a simple regression problem is considered in which a linear function with a normally distributed error term has to be estimated. In Section 6.1, the application of PFSs to this problem is discussed. It is demonstrated that a PFS with a limited number of rules generally cannot provide a satisfactory estimate of a linear function with a normally distributed error term. In Section 6.2, an alternative approach to probabilistic fuzzy modeling is proposed. It is shown that the use of PFSs can be seen as a special case of this approach. The proposed approach is successfully applied to the problem of estimating a linear function with a normally distributed error term.

6.1 Estimation of a linear function

Consider the regression problem of estimating an unknown linear function $f : X \to Y$. $f(x)$ is given by

$$f(x) = x + N(0, 1), \tag{6.1}$$

where $N(0, 1)$ denotes an error term that is drawn from a normal distribution with mean 0 and standard deviation 1. Since $f(x)$ has an error term, the function is stochastic. The domain of $f(x)$ is $X = [0, 10]$. For estimating $f(x)$, a large data set containing $n = 10,000$ examples (x_i, y_i) $(i = 1, \ldots, n)$ was generated. To generate an example (x_i, y_i), x_i was first drawn from a uniform distribution on X and y_i was then obtained using (6.1).

A PFS for regression tasks was used for estimating $f(x)$. The system's antecedent and consequent mfs were determined first and were not optimized using the data set. The mfs of the system's antecedent fuzzy sets A_1 and

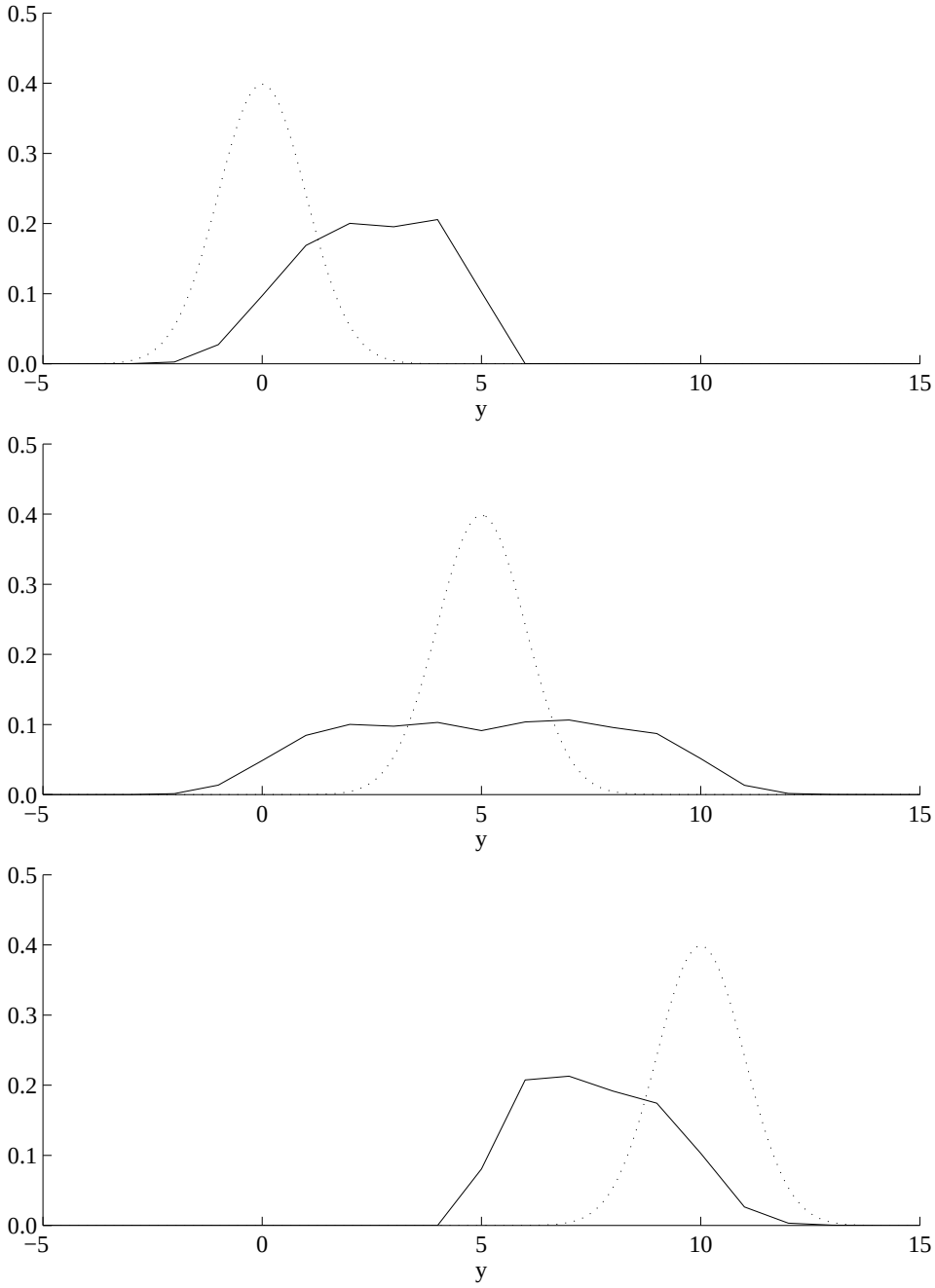

Figure 6.1: The conditional pdf $p(y|x)$ (dotted line) and the estimated conditional pdf $\hat{p}(y|x)$ (solid line) for $x = 0$ (upper panel), $x = 5$ (middle panel), and $x = 10$ (lower panel).

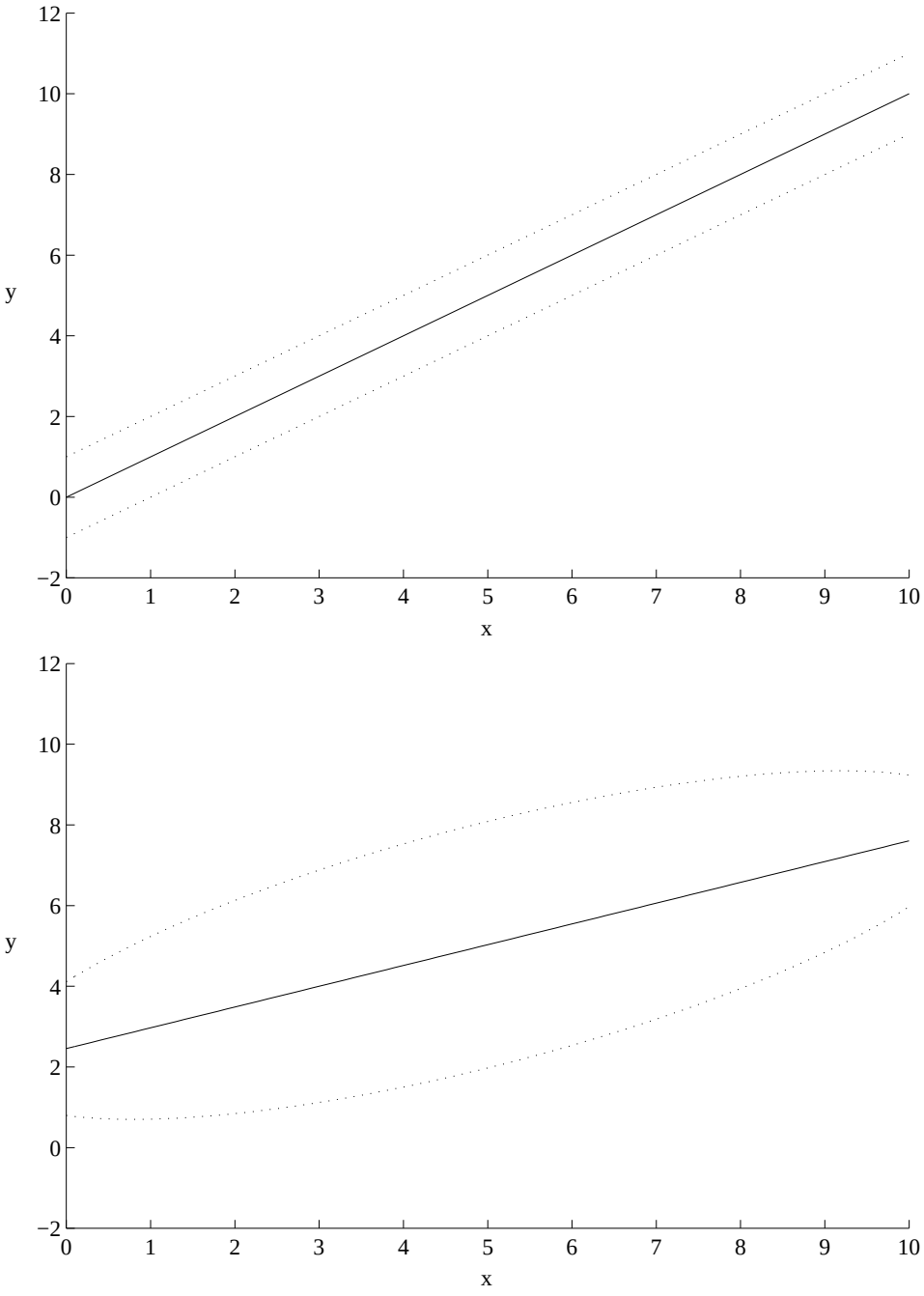

Figure 6.2: The upper panel shows the expectation of $f(x)$ (solid line) and the expectation of $f(x)$ plus or minus the standard deviation of $f(x)$ (dotted lines). The lower panel shows the estimate $\hat{f}(x)$ of the expectation of $f(x)$ (solid line) and the estimate $\hat{f}(x)$ of the expectation of $f(x)$ plus or minus the estimate $\hat{\sigma}(x)$ of the standard deviation of $f(x)$ (dotted lines).

	$j=1$	$j=2$
$k=1$	0.0000	0.0000
$k=2$	0.0000	0.0000
$k=3$	0.0003	0.0000
$k=4$	0.0027	0.0000
$k=5$	0.0273	0.0000
$k=6$	0.0972	0.0000
$k=7$	0.1689	0.0000
$k=8$	0.2002	0.0000
$k=9$	0.1954	0.0000
$k=10$	0.2057	0.0000
$k=11$	0.1022	0.0805
$k=12$	0.0000	0.2074
$k=13$	0.0000	0.2130
$k=14$	0.0000	0.1916
$k=15$	0.0000	0.1744
$k=16$	0.0000	0.1029
$k=17$	0.0000	0.0265
$k=18$	0.0000	0.0033
$k=19$	0.0000	0.0004
$k=20$	0.0000	0.0000
$k=21$	0.0000	0.0000

Table 6.1: The values of the probability parameters $p_{j,k}$.

A_2 were given by

$$\mu_{A_1}(x) = 1 - 0.1x \qquad \text{and} \qquad \mu_{A_2}(x) = 0.1x. \tag{6.2}$$

It follows from (3.4) that $\bar{\mu}_{A_j} = \mu_{A_j}$ for $j = 1, 2$. The system's output space Y was partitioned using 21 consequent fuzzy sets C_k $(k = 1, \ldots, 21)$ with triangular mfs given by

$$\mu_{C_k}(y) = \begin{cases} y - k + 7 & \text{if } k-7 \le y < k-6 \\ k - y - 5 & \text{if } k-6 \le y < k-5 \\ 0 & \text{otherwise.} \end{cases} \tag{6.3}$$

It should be noted that only 21 consequent fuzzy sets needed to be considered because it turned out that $y_i \in [-5, 15]$ for all examples in the data set.

Given the antecedent mfs μ_{A_j} and the consequent mfs μ_{C_k}, the probability parameters $p_{j,k}$ in the PFS were estimated using the ML method, as discussed in Section 4.2. The convex programming problem of finding ML estimates of the system's probability parameters was solved using the function *fmincon* in MATLAB's optimization toolbox. The standard parameter settings of this function were used, except that the medium-scale optimization algorithm was chosen instead of the large-scale optimization algorithm. The values that were obtained for the probability parameters $p_{j,k}$ are shown in Table 6.1.

The resulting PFS provided an estimate $\hat{p}(y|x)$ of the conditional pdf $p(y|x)$. For $x = 0$, $x = 5$, and $x = 10$, $p(y|x)$ and $\hat{p}(y|x)$ are shown in Figure 6.1. Following the discussion in Subsection 3.3.2, estimates of the expectation and the variance (or standard deviation) of the stochastic function $f(x)$ in (6.1) were derived from the estimated conditional pdf $\hat{p}(y|x)$. The estimate $\hat{f}(x)$ of the expectation of $f(x)$ and the estimate $\hat{\sigma}(x)$ of the standard deviation of $f(x)$ are shown in the lower panel of Figure 6.2. For comparison, the true expectation and the true standard deviation of $f(x)$ are shown in the upper panel of the figure. From Figure 6.1 and 6.2, it must be concluded that the PFS did not provide a satisfactory estimate of $f(x)$.

6.2 An alternative approach to modeling probabilistic uncertainty using fuzzy systems

To gain further insight into the reasoning mechanism that is used in PFSs for regression tasks, it is important to observe that the reasoning mechanism can be interpreted in a different way than in Chapter 3 (see also [16]). First, an estimate $\hat{p}(y|A_j)$ of the conditional pdf of $\underline{y}$ given fuzzy event A_j can be obtained as follows

$$\hat{p}(y|A_j) = \sum_{k=1}^{c} p(y|C_k) p_{j,k}, \tag{6.4}$$

where $p(y|C_k)$ is defined by (3.13). Then, an estimate $\hat{p}(y|\mathbf{x})$ of the conditional pdf $p(y|\mathbf{x})$ is given by

$$\hat{p}(y|\mathbf{x}) = \sum_{j=1}^{a} \bar{\mu}_{A_j}(\mathbf{x})\, \hat{p}(y|A_j), \tag{6.5}$$

where the normalized mf $\bar{\mu}_{A_j}$ is defined by (3.4). This probabilistic fuzzy reasoning mechanism is mathematically equivalent to the reasoning mechanism described in Chapter 3. This can be proven by showing that substitution of (6.4) in (6.5) results in the same expression as substitution of (3.5) in (3.12).

The above probabilistic fuzzy reasoning mechanism can be used for explaining the results in Figure 6.1. From (6.2) and (6.5), it follows that $\hat{p}(y|x = 0) = \hat{p}(y|A_1)$ and $\hat{p}(y|x = 10) = \hat{p}(y|A_2)$. The solid lines in the upper and the lower panel of Figure 6.1 therefore show the conditional pdfs $\hat{p}(y|A_1)$ and $\hat{p}(y|A_2)$. Since $\bar{\mu}_{A_1}(5) = \bar{\mu}_{A_2}(5) = 0.5$, it follows from (6.5) that $\hat{p}(y|x = 5)$ is obtained by taking the average of $\hat{p}(y|A_1)$ and $\hat{p}(y|A_2)$. Therefore, the solid line in the middle panel of Figure 6.1 should be the average of the solid lines in the upper and the lower panel of the figure. It can be seen that this is indeed the case.

I now propose an alternative approach to modeling probabilistic uncertainty using fuzzy systems. Consider a fuzzy system with rules that have

the following general form

$$\text{If } \mathbf{x} \text{ is } A_j \text{ then } p(y) = \phi(y; \alpha_{j,1}, \ldots, \alpha_{j,m}), \tag{6.6}$$

where $\phi(y; \alpha_{j,1}, \ldots, \alpha_{j,m})$ denotes the pdf of $\underline{y}$ that results from the jth rule ($j = 1, \ldots, a$). The function $\phi(y; \alpha_{j,1}, \ldots, \alpha_{j,m})$ is characterized by the parameters $\alpha_{j,k}$ ($k = 1, \ldots, m$). To obtain a valid pdf, it is necessary that for all values of these parameters $\phi(y; \alpha_{j,1}, \ldots, \alpha_{j,m})$ satisfies

$$\int \phi(y; \alpha_{j,1}, \ldots, \alpha_{j,m})\, dy = 1 \tag{6.7}$$

and

$$\phi(y; \alpha_{j,1}, \ldots, \alpha_{j,m}) \geq 0 \qquad \forall y \in Y. \tag{6.8}$$

In the fuzzy system, it is assumed that all rules use the same function $\phi(y; \alpha_{j,1}, \ldots, \alpha_{j,m})$. Only the values of the parameters $\alpha_{j,k}$ are different in each rule. Any function that satisfies the conditions in (6.7) and (6.8) can be chosen for $\phi(y; \alpha_{j,1}, \ldots, \alpha_{j,m})$. For example, if the normal distribution function is chosen, then $\phi(y; \alpha_{j,1}, \ldots, \alpha_{j,m})$ becomes

$$\phi(y; \alpha_{j,1}, \alpha_{j,2}) = \frac{1}{\alpha_{j,2}\sqrt{2\pi}} \exp\left(-\frac{(y - \alpha_{j,1})^2}{2\alpha_{j,2}^2}\right), \tag{6.9}$$

where $\alpha_{j,1}$ and $\alpha_{j,2} > 0$ denote, respectively, the mean and the standard deviation of a normal distribution. Other functions that may be chosen for $\phi(y; \alpha_{j,1}, \ldots, \alpha_{j,m})$ include, for example, functions based on a mixture model and functions based on a (fuzzy) histogram.

It is important to note that in a fuzzy system with rules given by (6.6), no fuzzy sets need to be defined in the output space, as is the case in a PFS for regression tasks. Instead, the assumption is made that the conditional pdf $p(y|\mathbf{x})$ can be described by a specific model, for example a normal distribution or a Gaussian mixture model. Consequently, for each rule in a system's rule base, instead of the probability parameters $p_{j,k}$ the parameters $\alpha_{j,k}$ of the model that is used for density estimation need to be determined. Notice that probability parameters in a PFS have a simple interpretation, i.e. a probability parameter $p_{j,k}$ can be interpreted as the conditional probability of fuzzy event C_k given fuzzy event A_j. Because of this property, the rule base of a PFS can be easily interpreted. Since the fuzzy sets A_j and C_k can be given linguistic values, the user of a PFS only needs to be familiar with the concept of probability to be able to understand the rules in the system. In a fuzzy system with rules given by (6.6), on the other hand, it may be more difficult to interpret the system's rule base. The user of such a system needs to be familiar with the concept of a pdf and, probably more problematic, if the system's consequent parameters have to be determined using expert knowledge, then the user also needs to understand the model that is

used for density estimation. These requirements are a potential drawback of fuzzy systems with rules given by (6.6). However, as will become clear below, this drawback may be compensated by an improved approximation accuracy.

As I discussed above, two different, but mathematically equivalent reasoning mechanisms can be used in PFSs for regression tasks:

1. The probability parameters $p_{j,k}$ can first be interpolated, and the interpolated probability parameters can then be used for estimating the conditional pdf $p(y|\mathbf{x})$. This reasoning mechanism is applied in (3.5) and (3.12).

2. The conditional pdfs $p(y|A_j)$ can first be estimated, and an estimate of the conditional pdf $p(y|\mathbf{x})$ can then be obtained by interpolating the estimates $\hat{p}(y|A_j)$. This reasoning mechanism is applied in (6.4) and (6.5).

Similar reasoning mechanisms as in PFSs for regression tasks can be used in fuzzy systems with rules given by (6.6). If the parameters $\alpha_{j,k}$ are first interpolated and the interpolated parameters are then used for estimating the conditional pdf $p(y|\mathbf{x})$, then the estimate $\hat{p}(y|\mathbf{x})$ is given by

$$\hat{p}(y|\mathbf{x}) = \phi(y; \alpha_1(\mathbf{x}), \ldots, \alpha_m(\mathbf{x})), \tag{6.10}$$

where for $k = 1, \ldots, m$

$$\alpha_k(\mathbf{x}) = \sum_{j=1}^{a} \bar{\mu}_{A_j}(\mathbf{x})\, \alpha_{j,k}. \tag{6.11}$$

On the other hand, if instead of the parameters $\alpha_{j,k}$ the pdfs $\phi(y; \alpha_{j,1}, \ldots, \alpha_{j,m})$ are interpolated, then $\hat{p}(y|\mathbf{x})$ is given by

$$\hat{p}(y|\mathbf{x}) = \sum_{j=1}^{a} \bar{\mu}_{A_j}(\mathbf{x})\, \phi(y; \alpha_{j,1}, \ldots, \alpha_{j,m}). \tag{6.12}$$

In general, the reasoning mechanism in (6.10) and (6.11) and the reasoning mechanism in (6.12) do not provide the same estimates of $p(y|\mathbf{x})$. This is an important difference with PFSs for regression tasks.

Obviously, which of the above reasoning mechanisms gives better results depends on the problem to which a fuzzy system is applied. However, it seems reasonable to assume that in many regression problems the reasoning mechanism in (6.10) and (6.11) is more appropriate than the reasoning mechanism in (6.12). As an example, consider the regression problem that was studied in Section 6.1. In this problem, the stochastic function $f(x)$ in (6.1) has to be estimated. This is a linear function with a normally

distributed error term. The standard deviation of the error term does not depend on x. For estimating $f(x)$, the input space X has been partitioned using two fuzzy sets, A_1 and A_2. The mfs of these fuzzy sets are defined by (6.2). Suppose that a fuzzy system with rules given by (6.6) is used for estimating $f(x)$. Suppose further that the normal distribution function is chosen for $\phi(y; \alpha_{j,1}, \ldots, \alpha_{j,m})$, i.e. $\phi(y; \alpha_{j,1}, \ldots, \alpha_{j,m})$ is given by (6.9). If the reasoning mechanism in (6.10) and (6.11) is used, then it is possible to obtain a perfect estimate of $f(x)$. This is accomplished by setting $\alpha_{1,1} = 0$, $\alpha_{2,1} = 10$, and $\alpha_{1,2} = \alpha_{2,2} = 1$. If, on the other hand, the reasoning mechanism in (6.12) is used, then a perfect estimate of $f(x)$ cannot be obtained. To see this, notice that according to the reasoning mechanism in (6.12), the estimated conditional pdf $\hat{p}(y|x)$ results from averaging two normal distribution functions. As a consequence, $\hat{p}(y|x)$ itself will not, in general, be a normal distribution function (in many cases, $\hat{p}(y|x)$ will be a bimodal pdf). From (6.1), it follows that $p(y|x)$ is a normal distribution function. This implies that $\hat{p}(y|x)$ cannot be a perfect estimate of $p(y|x)$. Based on these observations, it must be concluded that in the regression problem studied in Section 6.1, the reasoning mechanism in (6.10) and (6.11) is more appropriate than the reasoning mechanism in (6.12). It seems reasonable to assume that this conclusion applies to many regression problems, since in many regression problems the probabilistic uncertainty will, to some extent, have similar characteristics as in the regression problem studied in Section 6.1.

6.2.1 Generalized probabilistic fuzzy systems

A PFS for regression tasks can be seen as a special case of a fuzzy system with rules given by (6.6). Or, in other words, a fuzzy system with rules given by (6.6) can be seen as a generalization of a PFS for regression tasks. This result is stated more precisely in the following theorem.

Theorem 6.1 *A fuzzy system with rules given by (6.6) is equivalent to a PFS for regression tasks if the following conditions are satisfied:*

1. *Both systems use the same number of rules a.*
2. *Both systems use the same antecedent fuzzy sets A_j $(j = 1, \ldots, a)$.*
3. *Both systems use the same number of parameters in a rule, i.e. $m = c$.*
4. *In both systems, corresponding parameters have the same value, i.e. $\alpha_{j,k} = p_{j,k}$ $(j = 1, \ldots, a$ and $k = 1, \ldots, m)$.*
5. *In the fuzzy system with rules given by (6.6), $\phi(y; \alpha_{j,1}, \ldots, \alpha_{j,m})$ is defined as*

$$\phi(y; \alpha_{j,1}, \ldots, \alpha_{j,m}) = \sum_{k=1}^{m} p(y|C_k)\alpha_{j,k}, \tag{6.13}$$

where the functions $p(y|C_k)$ $(k = 1, \ldots, m)$ are defined in the same way as in the PFS for regression tasks.

Proof: Notice that the theorem holds both for the reasoning mechanism in (6.10) and (6.11) and for the reasoning mechanism in (6.12). First consider the reasoning mechanism in (6.10) and (6.11). If the conditions in the theorem are satisfied, then $\alpha_k(\mathbf{x})$ given by (6.11) equals $\hat{p}(C_k|\mathbf{x})$ given by (3.5). Using (6.13), it can then be seen that (6.10) and (3.12) provide the same estimate $\hat{p}(y|\mathbf{x})$ of the conditional pdf $p(y|\mathbf{x})$. Now consider the reasoning mechanism in (6.12). If the conditions in the theorem are satisfied, then $\phi(y; \alpha_{j,1}, \ldots, \alpha_{j,m})$ is given by (6.13) and equals $\hat{p}(y|A_j)$ given by (6.4). It can then be seen that (6.12) and (6.5) provide the same estimate $\hat{p}(y|\mathbf{x})$ of the conditional pdf $p(y|\mathbf{x})$. This completes the proof of the theorem.

Because a fuzzy system with rules given by (6.6) can be seen as a generalization of a PFS for regression tasks, I will refer to a fuzzy system with this type of rules using the term *generalized PFS*.

6.2.2 A simple experiment

In this subsection, a simple experiment is described in which a generalized PFS was applied to the regression problem discussed in Section 6.1. In the experiment, the normal distribution function was chosen for the function $\phi(y; \alpha_{j,1}, \ldots, \alpha_{j,m})$ in the rules of the generalized PFS. This means that $\phi(y; \alpha_{j,1}, \ldots, \alpha_{j,m})$ was given by (6.9). Furthermore, the reasoning mechanism in (6.10) and (6.11) was applied in the system. For partitioning the input space X, the same antecedent fuzzy sets A_1 and A_2 were used as in Section 6.1. The mfs of A_1 and A_2 were therefore given by (6.2). Also, the same data set was used as in Section 6.1. The parameters $\alpha_{1,1}$, $\alpha_{1,2}$, $\alpha_{2,1}$, and $\alpha_{2,2}$ in the generalized PFS were estimated by maximizing the likelihood of the data set. Since $\alpha_{1,2}$ and $\alpha_{2,2}$ represented standard deviations, the constraints $\alpha_{1,2} > 0$ and $\alpha_{2,2} > 0$ were imposed. Maximization of the loglikelihood function was performed using the function *fmincon* in MATLAB's optimization toolbox. The standard parameter settings of this function were used, except that the maximum number of function evaluations was increased to 10,000 and that the medium-scale optimization algorithm was chosen instead of the large-scale optimization algorithm. The initial values of $\alpha_{1,1}$ and $\alpha_{2,1}$ were drawn from a uniform distribution between 0 and 10, and the initial values of $\alpha_{1,2}$ and $\alpha_{2,2}$ were drawn from a uniform distribution between 0 and 2.

The experiment was repeated ten times using different initial values of the parameters $\alpha_{j,k}$. In nine experiments, the optimization algorithm found the parameter values $\alpha_{1,1} = 0.0256$, $\alpha_{1,2} = 1.0014$, $\alpha_{2,1} = 9.9770$, and $\alpha_{2,2} = 1.0006$. These values were very close to the values $\alpha_{1,1} = 0$, $\alpha_{1,2} = 1$, $\alpha_{2,1} = 10$, and $\alpha_{2,2} = 1$ that would have resulted in a perfect estimate of the

function $f(x)$ in (6.1). In one experiment, the optimization algorithm found the parameter values $\alpha_{1,1} = -0.2762 \cdot 10^6$, $\alpha_{1,2} = 8.4147 \cdot 10^6$, $\alpha_{2,1} = 457.36$, and $\alpha_{2,2} = 0.0001$. Clearly, these values did not result in a satisfactory estimate of $f(x)$. The results of the experiments indicate that appropriate values for the parameters in a generalized PFS can be obtained by using the ML criterion.

Chapter 7

Conclusions and Future Research

7.1 Conclusions

The main conclusions of this thesis are summarized below. The first conclusion addresses the first research question of the thesis, as formulated in Section 1.3. The other conclusions relate to the second research question of the thesis.

1. In general, fuzzy histograms have the same rate of convergence as ordinary crisp histograms. This means that, in general, fuzzy histograms are statistically quite inefficient. However, a special class of fuzzy histograms, which includes fuzzy histograms that use triangular mfs, has the same rate of convergence as kernel density estimators. This special class of fuzzy histograms therefore is statistically quite efficient. Since fuzzy histograms are computationally much more efficient than kernel density estimators, fuzzy histograms can be used to combine a high level of statistical efficiency with a high level of computational efficiency.

2. The conditional probability method for estimating the probability parameters in a PFS [16, 18, 34] provides estimates that have unsatisfactory statistical properties. As an alternative, the probability parameters in a PFS can be estimated using the criterion of maximum likelihood. Finding maximum likelihood estimates of probability parameters is a convex programming problem.

3. In PFSs for classification tasks, the criterion of maximum likelihood can be used for estimating both the probability parameters and the antecedent parameters. In the experiments described in this thesis, the maximum likelihood method gives good results.

4. In many regression problems, it is questionable whether a PFS with a limited number of rules has a satisfactory approximation accuracy. A generalized PFS may have a better approximation accuracy, but the rule base of such a system may also be more difficult to interpret. Probabilistic fuzzy modeling in regression problems is an important issue for future research.

7.2 Future research

There are two main issues for future research:

1. Concerning fuzzy histograms in a sample space $\mathbb{R}$ that is uniformly partitioned, an important question that has not been addressed in this thesis is what choice of mf μ_A in (2.33) is statistically most efficient. It has been shown that the statistically most efficient choice of mf μ_A must satisfy the condition $P(\mu_A) = 0$, where $P(\mu_A)$ is given by (2.49). However, within the class of all mfs that satisfy this condition, it is not known which mf is statistically most efficient. Furthermore, the small sample properties of fuzzy histograms may be studied in future research. This can be done by using Monte Carlo simulation, which has also been used for studying the small sample properties of other nonparametric density estimators [27, 29].

2. The issue of probabilistic fuzzy modeling in regression problems deserves considerable attention in future research. The claim made in this thesis that in many regression problems a PFS with a limited number of rules does not have a satisfactory approximation accuracy needs to be tested experimentally. If the claim turns out to be correct, then an alternative approach to probabilistic fuzzy modeling in regression problems is needed. The idea of generalized PFSs may then be further elaborated. However, since the interpretation of the rules in a generalized PFS is relatively difficult, other approaches to probabilistic fuzzy modeling in regression problems should also be considered.

Bibliography

[1] J. Abonyi and F. Szeifert. Supervised fuzzy clustering for the identification of fuzzy classifiers. *Pattern Recognition Letters*, 24:2195–2207, 2003.

[2] C.M. Bishop. *Neural networks for pattern recognition*. Oxford University Press, 1995.

[3] C.L. Blake and C.J. Merz. UCI machine learning repository, 1998. `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

[4] H. Chongfu. Demonstration of benefit of information distribution for probability estimation. *Signal Processing*, 80:1037–1048, 2000.

[5] E. Cox. *The fuzzy systems handbook*. Academic Press, second edition, 1999.

[6] H. Dourra and P. Siy. Investment using technical analysis and fuzzy logic. *Fuzzy Sets and Systems*, 127:221–240, 2002.

[7] K.J. Hunt, R. Haas, and R. Murray-Smith. Extending the functional equivalence of radial basis function networks and fuzzy inference systems. *IEEE Transactions on Neural Networks*, 7:776–781, 1996.

[8] D. Husmeier and J.G. Taylor. Predicting conditional probability densities of stationary stochastic time series. *Neural Networks*, 10:479–497, 1997.

[9] J.-S.R. Jang and C.-T. Sun. Functional equivalence between radial basis function networks and fuzzy inference systems. *IEEE Transactions on Neural Networks*, 4:156–159, 1993.

[10] J.-S.R. Jang, C.-T. Sun, and E. Mizutani. *Neuro-fuzzy and soft computing*. Prentice Hall, 1997.

[11] M.C. Jones. Discretized and interpolated kernel density estimates. *Journal of the American Statistical Association*, 84:733–741, 1989.

[12] G.G. Judge, W.E. Griffiths, R.C. Hill, and T.C. Lee. *The theory and practice of econometrics*. John Wiley, 1980.

[13] N.B. Karayiannis and G.W. Mi. Growing radial basis neural networks: merging supervised and unsupervised learning with network growth techniques. *IEEE Transactions on Neural Networks*, 8:1492–1506, 1997.

[14] U. Kaymak. Fuzzy target selection using RFM variables. In *Proceedings of the Joint 9th IFSA World Congress and 20th NAFIPS International Conference*, pages 1038–1043, 2001.

[15] U. Kaymak. Data and cluster weighting in target selection based on fuzzy clustering. *Lecture Notes in Computer Science*, 2715:568–575, 2003.

[16] U. Kaymak and J. van den Berg. On probabilistic connections of fuzzy systems. In *Proceedings of the 15th Belgian-Dutch Conference on Artificial Intelligence (BNAIC'03)*, pages 187–194, 2003.

[17] U. Kaymak and J. van den Berg. On constructing probabilistic fuzzy classifiers from weighted fuzzy clustering. In *Proceedings of the 13th IEEE International Conference on Fuzzy Systems*, pages 395–400, 2004.

[18] U. Kaymak, W.-M. van den Bergh, and J. van den Berg. A fuzzy additive reasoning scheme for probabilistic Mamdani fuzzy systems. In *Proceedings of the 12th IEEE International Conference on Fuzzy Systems*, pages 331–336, 2003.

[19] V. Kodogiannis and A. Lolis. Forecasting financial time series using neural network and fuzzy system-based techniques. *Neural Computing and Applications*, 11:90–102, 2002.

[20] S. Li. The development of a hybrid intelligent system for developing marketing strategy. *Decision Support Systems*, 27:395–409, 2000.

[21] D. Lowe. Radial basis function networks. In M.A. Arbib, editor, *The handbook of brain theory and neural networks*. MIT Press, 1995.

[22] R.T. McIvor, A.G. McCloskey, P.K. Humphreys, and L.P. Maguire. Using a fuzzy approach to support financial analysis in the corporate acquisition process. *Expert Systems with Applications*, 27:533–547, 2004.

[23] C.A. Peña-Reyes and M. Sipper. A fuzzy-genetic approach to breast cancer diagnosis. *Artificial Intelligence in Medicine*, 17:131–155, 1999.

[24] R. Potharst, U. Kaymak, and W. Pijls. Neural networks for target selection in direct marketing. In K. Smith and J. Gupta, editors, *Neural networks in business: techniques and applications*. Idea Group Publishing, 2002.

[25] T.A. Runkler. Fuzzy histograms and fuzzy chi-squared tests for independence. In *Proceedings of the 13th IEEE International Conference on Fuzzy Systems*, pages 1361–1366, 2004.

[26] L. Sánchez, J. Casillas, O. Cordón, and M. José del Jesus. Some relationships between fuzzy and random set-based classifiers and models. *International Journal of Approximate Reasoning*, 29:175–213, 2002.

[27] D.W. Scott. On optimal and data-based histograms. *Biometrika*, 66:605–610, 1979.

[28] D.W. Scott. Averaged shifted histograms: effective nonparametric density estimators in several dimensions. *The Annals of Statistics*, 13:1024–1040, 1985.

[29] D.W. Scott. Frequency polygons: theory and application. *Journal of the American Statistical Association*, 80:348–354, 1985.

[30] D.W. Scott. *Multivariate density estimation.* John Wiley & Sons, 1992.

[31] S. Siekmann, R. Kruse, J. Gebhardt, F. van Overbeek, and R. Cooke. Information fusion in the context of stock index prediction. *International Journal of Intelligent Systems*, 16:1285–1298, 2001.

[32] A. Takayama. *Mathematical economics.* The Dryden Press, 1974.

[33] W.L. Tung, C. Quek, and P. Cheng. GenSo-EWS: a novel neural-fuzzy based early warning system for predicting bank failures. *Neural Networks*, 17:567–587, 2004.

[34] J. van den Berg, U. Kaymak, and W.-M. van den Bergh. Financial markets analysis by using a probabilistic fuzzy modelling approach. *International Journal of Approximate Reasoning*, 35:291–305, 2004.

[35] J. van den Berg, W.-M. van den Bergh, and U. Kaymak. Probabilistic and statistical fuzzy set foundations of competitive exception learning. In *Proceedings of the 10th IEEE International Conference on Fuzzy Systems*, pages 1035–1038, 2001.

[36] G.C. van den Eijkel. *Fuzzy probabilistic learning and reasoning.* PhD thesis, Delft University of Technology, 1998.

[37] L.A. Zadeh. Probability measures of fuzzy events. *Journal of Mathematical Analysis and Applications*, 23:421–427, 1968.