PERIODICA POLYTECHNICA SER. EL. ENG. VOL. 43, NO. 2, PP. 109–126 (1999)

ESTIMATION OF CONDITIONAL QUANTILE USING NEURAL NETWORKS

Piotr KULCZYCKI* and Henrik SCHIØLER**

*Systems Research Institute
Polish Academy of Sciences
Newelska 6
PL-01-477 Warszaw, Poland
**Department of Control Engineering
Aalborg University
Fredrik Bajers vej 7
DK-9220 Aalborg Ø, Denmark

Received: Dec. 6, 1999

Abstract

The problem of estimating conditional quantiles using neural networks is investigated here. A basic structure is developed using the methodology of kernel estimation, and a theory guaranteeing consistency on a mild set of assumptions is provided. The constructed structure constitutes a basis for the design of a variety of different neural networks, some of which are considered in detail. The task of estimating conditional quantiles is related to Bayes point estimation whereby a broad range of applications within engineering, economics and management can be suggested. Numerical results illustrating the capabilities of the elaborated neural network are also given.

Keywords: neural networks, conditional quantile, kernel estimators, time-optimal control.

1. Introduction

For statistical purposes, distributions of random variables are most often reported through characteristic parameters describing their fundamental features. Moments, especially mean value and variance, constitute a well known example of such quantities. Another group of characteristics are the positional parameters, namely quantiles and their functions (FISZ, 1963), which are more directly connected to the distribution function by relating certain points to its assumed values. Frequently the median (quantile of order 0.5) is treated as the mean, and the quantile deviation, i.e. the difference between quantiles of order 0.75 and 0.25, can be interpreted similarly to variation. Also special quantiles, such as quadriles, deciles and percentiles often appear in statistical applications.

If auxiliary variables are available, conditional probability distributions may be defined. Consequently, their characteristic parameters, e.g. conditional quantiles, are given as functions of those auxiliary variables. When standard distributions are encountered (for example, if all variables are jointly Gaussian) or in general when the conditional characteristics are linear functions, the problem of estimation is thoroughly investigated, and a variety of methods can be applicable.

The situation is severely complicated if distributions are far from standard, and the conditional characteristics are nonlinear functions with an unknown structure. In this case nonparametric methods including neural networks may prove to be useful, and the precise purpose of this paper is to constructively design a neural network applicable for estimating conditional quantiles in the general nonstandard situation.

Neural networks have in recent years developed into powerful tools for solving optimisation problems within, for example, classification, estimation and forecasting. For the majority of cases, the applied neural networks, from a statistical point of view, solve conditional estimation problems. The celebrated Back Propagation Error algorithm used for training Feed Forward Neural Networks is shown to be a special case of gradient optimisation in the sense of mean squared error (RUMELHART and MCCLELLAND, 1986). Feed Forward Neural Networks are in paper (WHITE, 1990) analysed for consistent estimation of conditional expectation functions, which optimise expected squared error. Optimal classification is concerned with the problem of classifying, on the basis of feature measurements, a set of objects, while obtaining a minimal probability of misclassification. This problem is equivalent to conditional estimation, and it is shown in work (RUCK et al, 1990) that Feed Forward Neural Networks estimate the optimal discriminating function, which is the conditional class probability, when trained with the Back Propagation Error Algorithm. In all of the above cases, some sort of optimisation or training algorithm is applied adjusting initially random network parameters optimally w.r.t. average loss functions on a finite set of training data. A more constructive way to follow is indicated in paper (SPECHT, 1988), where a Probabilistic Neural Network for classification based on kernel estimators is investigated, as well as by articles (SPECHT, 1991; SCHIØLER and HARTMANN, 1992), in which a similar line is followed for proposing neural networks estimating conditional expectation functions. From a certain point of view, this strategy is the basis for suggesting a large class of different neural network architectures, including among others Localised Receptive Fields (MOODY and DARKEN, 1989) and Counter Propagation Networks (NIELSEN, 1987). In this paper such a constructive strategy is pursued in order to design a Feed Forward Neural Network capable of estimating conditional quantiles.

The paper is organised as follows. In Section 2 the mathematical preliminaries of Bayes estimation for the special case, where the associated loss function is partially linear, will be described. Such a loss function introduces the need for estimating the underlying distribution function as well as its quantiles. A kernel based estimator for estimating the above quantities will be developed in Section 3, where an appropriate neural network interpretation is also given. The non parametrical kernel estimator suffers from the fact that no meaning could be associated to the specific model parameters, i.e. they do not represent verbally expressible facts as in similar model structures like f.ex. fuzzy logic or neurofuzzy modelling, where model structure and parameters have a corresponding linguistic representation. For such meaning in order to be associated, a model reduction is strongly required. In the end of Section 3 such a model reduction is suggested along with a scheme for setting model parameters to make the reduced model approximating the former sufficiently close. In Section 4 a numerical example verifying the theoretical

considerations, whereas a conclusion will be provided in the last section.

2. Mathematical Preliminaries

Consider a real random variable w with a distribution P_w, and a number $p \in (0, 1)$. Any real number q_w fulfilling the following inequalities:

$$P_w((-\infty, q_w]) \geq p, \tag{1}$$

$$P_w([q_w, \infty)) \geq 1 - p \tag{2}$$

is said to be a quantile of order p (FISZ, 1963). If the distribution function F_w is continuous and strictly monotonous, the quantile of order p is uniquely defined by the formula

$$F_w(q_w) = p. \tag{3}$$

Thus the quantile divides the real space into two parts, having probabilities p and $1 - p$ of containing realisations of the random variable w. The quantile of order 0.5 is simply the median; quantiles of orders 0.25, 0.5 and 0.75 are called quadriles; quantiles of orders 0.1, 0.2, . . ., 0.9 deciles, and orders 0.01, 0.02, . . ., 0.99 designate percentiles.

Assume that the quantile q_w is uniquely defined, and $\{F_w^n\}$ denotes a sequence of continuous and strictly monotonous distribution functions converging pointwise to the function F_w at every point of its continuity. Let the sequence $\{q_w^n\}$ be defined uniquely by

$$F_w^n(q_w^n) = p. \tag{4}$$

It is then readily shown that $\{q_w^n\}$ converges towards the quantile q_w. Such a strategy is followed throughout this paper in the design of neural networks for estimating quantiles.

One important statistical application of quantiles is the problem of Bayes point estimation, which, for the sake of illustration, is considered below. In that case a so-called loss function $l : \mathbb{R} \times \mathbb{R} \to [0, \infty)$, representing losses caused by estimation error, will be defined. Its value $l(W, w)$ is interpreted as the loss incurred when estimating the parameter w by the value W. The Bayes loss function l_B can now be defined as follows:

$$l_B(W) = \int_{\mathbb{R}} l(W, w)\, \mathrm{d}P_w(w). \tag{5}$$

The value $l_B(W)$ simply constitutes the expected loss when estimating the parameter w by the value W. Any real number W_B such that

$$l_B(W_B) = \inf_{W \in \mathbb{R}} l_B(W) \tag{6}$$

is called a Bayes estimator.

In cases where losses depend strongly on the sign of estimation error, the loss function l may be defined by

$$l(W, w) = \begin{cases} -a(W - w) & \text{if} \quad W - w \leq 0 \\ b(W - w) & \text{if} \quad W - w \geq 0 \end{cases}, \tag{7}$$

where a and b are real positive numbers. In this case it is readily shown that the Bayes estimator equals the quantile of order

$$p = \frac{a}{a + b}. \tag{8}$$

In the special case where $a = b = 1$ the function l yields absolute value, and the Bayes estimator constitutes simply the median (LEHMANN, 1983, Chapter 4).

A practical example illustrating the relevance of Bayes estimation is described in paper (KULCZYCKI, 1993), where it has been applied to solve a time-optimal control problem. A parameter w, representing motion resistances in a mechanical system, is estimated by the value W, which appears directly in the equations of a time-optimal feedback controller. If $W > w$, overshoots occur, which increase the time to reach the target proportionally to $W - w$ with a coefficient b. In the case where $W < w$, so-called sliding trajectories appear, also prolonging the reaching period proportionally to $w - W$ with a coefficient a. The Bayes optimal estimator of the parameter w therefore exactly constitutes a quantile of order $\frac{a}{a + b}$. That problem has been solved in paper (KULCZYCKI and SCHIØLER, 1994) using a preliminary version of the neural network presented in this work.

The uncertainty of estimated parameters is in practice often caused by disturbances, some of which might be measured and used for improving the quality of estimation. The mathematical tool supporting this aim is provided by the concept of conditional distribution.

Consider the random variables w and $v = (v_1, v_2, \ldots, v_n)$ defined on a common probability space with a joint distribution P_{wv} on the space $\mathbb{R}^{n+1}$. Then the function $P_{w|v} : \beta(\mathbb{R}) \times \mathbb{R}^n \to [0, 1]$, where $\beta(A)$ denotes hereinafter the class of measurable subsets of the space A, exists (BILLINGSLEY, 1979, Section 33) so that

1. for every $v \in \mathbb{R}^n$, $P_{w|v}(\cdot, v)$ is a probability measure on the space $\mathbb{R}$,
2. for every $A \in \beta(\mathbb{R})$ and $B \in \beta(\mathbb{R}^n)$

$$P_{wv}(A \times B) = \int_B P_{w|v}(A, v)\, \mathrm{d}P_v(v). \tag{9}$$

Eq. (9) defines $P_{w|v}(A, \cdot)$ almost everywhere uniquely w.r.t. P_v, i.e. the particular versions of this function differ only on a zero measure set. The measure $P_{w|v}(\cdot, v)$ is called the conditional probability of the random variable w with respect to v. In

the case where the joint distribution P_{wv} has a density function h_{wv}, a conditional density function $h_{w|v}$ is given as

$$h_{w|v}(w, v) = \frac{h_{wv}(w, v)}{\int_{-\infty}^{\infty} h_{wv}(x, v)\,\mathrm{d}x} \tag{10}$$

for every v where the denominator in the above formula is nonzero. Then the conditional probability $P_{w|v}$ can be found explicitly by

$$P_{w|v}((-\infty, d]), v) = \int_{-\infty}^{\mathrm{d}} h_{w|v}(w, v)\,\mathrm{d}w \tag{11}$$

for every $d \in \mathbb{R}$.

For any $v \in \mathbb{R}^n$ the conditional quantile $q_{w|v}$ is defined analogously to the unconditional case, i.e. formulas (1)–(2) are replaced by their conditional equivalents

$$P_{w|v}((-\infty, q_{w|v}(v)], v) \geq p, \tag{12}$$

$$P_{w|v}([q_{w|v}(v), \infty), v) \geq 1 - p. \tag{13}$$

Analogously, if for some $v \in \mathbb{R}^n$ the distribution function F_v is given as

$$F_v(d) = P_{w|v}((-\infty, d], v) \tag{14}$$

and $\{F_v^n\}$ denotes a sequence of continuous and strictly monotonous functions defining the sequence $\{q_v^n\}$ by the equality

$$F_v^n(q_v^n) = p, \tag{15}$$

then it is readily shown that $\{q_v^n\}$ converges towards the conditional quantile $q_{w|v}(v)$, when the latter value is unique.

Similarly, the conditional Bayes estimator $W_B(v)$ can for every $v \in \mathbb{R}^n$ be given by

$$l_B^*(W, v) = \int_{\mathbb{R}} l(W, v)\,\mathrm{d}P_{w|v}(w), \tag{16}$$

$$l_B^*(W_B(v), v) = \inf_{W \in \mathbb{R}} l_B^*(W, v). \tag{17}$$

As can be seen directly from the above definitions, for the loss function given in *Eq.* (7), the conditional Bayes estimator constitutes a conditional quantile of the order r defined by formula (8).

In the time-optimal control problem mentioned above as an application example (KULCZYCKI, 1993), the vector v contains disturbances possibly influencing the value of motion resistance, such as temperature or target position. After measuring the observations $\hat{v} = (\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_n)$ of these quantities, the Bayes estimator $W_B(\hat{v})$ yields the minimum expected reaching time when applied to the feedback controller equations.

3. Neural Networks for Estimating Conditional Quantiles

Feed Forward Neural Networks are most frequently trained by applying some sort of optimisation procedure, e.g. Back Propagation, in order to set weights and offsets optimally with respect to some objective function defined for a finite sample of training data. In any case this function equals the average value of some loss function on the available set of data. Thus for any reasonable set of assumptions the objective function constitutes an estimator of the expected loss function, i.e. the Bayes loss function. This situation has been investigated in papers (WHITE, 1990; RUCK et al, 1990) for probability of misclassification and squared error.

In this section, a neural network for estimating conditional quantiles, which after training is applicable to all possible values of the quantile order p, will be elaborated. The reasoning follows the constructive line of works (SPECHT, 1988, 1990; SCHIØLER and HARTMANN, 1992) and is based on the theory of kernel estimation, which will be introduced shortly below.

Let $\{w_i\}$ in the following be a sequence of identically distributed random variables with a common density h_w. For any $m \in \mathbb{N}\backslash\{0\}$ and $r > 0$ the density estimator $h_w^{m,r}$ can be defined by

$$h_w^{m,r}(w) = \frac{1}{mV(r)} \sum_{i=1}^{m} \phi\left(\frac{w - w_i}{r}\right), \tag{18}$$

where the volume function V is expressed as

$$V(r) = \int_{-\infty}^{\infty} \phi\left(\frac{w}{r}\right) \mathrm{d}w \tag{19}$$

and the kernel function ϕ obeys

$$\lim_{r\to 0} \frac{1}{V(r)} \int_{-\infty}^{\infty} h(w)\phi\left(\frac{w-d}{r}\right) \mathrm{d}w = h(d) \tag{20}$$

for any bounded continuous density function h. The above estimator has been investigated in paper (PARZEN, 1962) for the case when the sequence $\{w_i\}$ consists of i.i.d. (independent identically distributed) random variables with a common continuous density function h_w. For $r \to 0$, and $m \cdot r \to \infty$ as $m \to \infty$, the function $h_w^{m,r}$ is shown to be a pointwise consistent estimator of the density h_w. By interpreting the kernel function ϕ as the nonlinearity of a neuron, and the sequence $\{w_i\}$ as a set of observations serving as training data, it has been demonstrated in paper (SPECHT, 1991) how this estimator exhibits properties equivalent to neural networks. From a computational point of view it possesses a massively parallel structure, which allows for high speed implementation on dedicated hardware; functionally, it is capable of learning general probabilistic information from measured data. It should be pointed out, however, that the number of neurons in the network defined from formula (18) equals the number of data in the training set, and that learning takes

place more or less by memorising data. In that respect the network provides no data compression, which might be introduced by interpreting *Eq.* (18) only as a paradigm defining the structure of the network, and by replacing all constants with trainable parameters. Such an approach will be pursued in the following.

In papers (SPECHT, 1991; SCHIØLER and HARTMANN, 1992) the estimator (18) was transformed to compute conditional expectation functions. Here this transformation is directed towards estimators of conditional distribution functions, which are applied for the estimation of conditional quantiles.

In the multivariable case the training data is a finite sequence of the form $\{(w_i, v_i)\}$, where v_i denotes an observation of some observable explanatory variable. In that case the multivariable density estimator $h_{wv}^{m,r}$ can be given as

$$h_{wv}^{m,r}(w, v) = \frac{1}{mV(r)} \sum_{i=1}^{m} \phi\left(\frac{w - w_i}{r}\right) \cdot \phi\left(\frac{v - v_i}{r}\right). \tag{21}$$

The function $h_{wv}^{m,r}$ defines a measure $P_{wv}^{m,r}$ which hopefully provides an acceptable estimator of the measure P_{wv}. A conditional distribution estimator $P_{w|v}^{m,r}$ can be obtained by subjecting $h_{wv}^{m,r}$ to a transformation analogous to the one defined by *Eqs.* (10) and (11), i.e.

$$P_{w|v}^{m,r}((-\infty, d], v) = \frac{\displaystyle\int_{-\infty}^{\mathrm{d}} h_{wv}^{m,r}(w, v)\,\mathrm{d}w}{\displaystyle\int_{-\infty}^{\infty} h_{wv}^{m,r}(w, v)\,\mathrm{d}w}, \tag{22}$$

which leads to the following closed form expression

$$P_{w|v}^{m,r}((-\infty, d], v) = \frac{\displaystyle\sum_{i=1}^{m} S\left(\frac{d - w_i}{r}\right) \cdot \phi\left(\frac{v - v_i}{r}\right)}{\displaystyle\sum_{i=1}^{m} \phi\left(\frac{v - v_i}{r}\right)}, \tag{23}$$

where S denotes the antiderivative of the function ϕ, i.e.

$$S(d) = \int_{-\infty}^{\mathrm{d}} \phi(w)\,\mathrm{d}w. \tag{24}$$

A scaled Gaussian density may be proposed as a candidate for the function ϕ, namely

$$\phi(d) = \exp(-d^2). \tag{25}$$

This function exhibits all properties required here except that its antiderivative is not computable in a closed form expression. Therefore the function S can be chosen not according to *Eq.* (24), but as a function exhibiting equivalent properties and

computable in a closed form expression. The well known sigmoid function then constitutes a natural choice, i.e.

$$S(d) = \frac{1}{1 + \exp(-d)}. \tag{26}$$

The above elaboration is based on kernel estimation of a joint density function h_{wv}, and leads to an estimator $P^{m,r}_{w|v}$ of the conditional distribution $P_{w|v}$; it serves here merely as motivation to formula (23), which is more generally valid than indicated above. In fact, from definitions (25) and (26) it can be shown by fairly standard means on a very mild set of assumptions only that the function $P^{m,r}_{w|v}$ consistently estimates the distribution $P_{w|v}$, as stated precisely in the following theorem, which is proved in the appendix. For simplicity, the theorem is stated and proved for w and v being one dimensional, but the result is straightforwardly generalised to arbitrary dimensions.

THEOREM 1 *Let P_{wv} be a probability distribution on the space $\mathbb{R}^2$ with an associated distribution function F_{wv}, and define the distribution P_v on the real space by*

$$P_v(A) = P_{wv}(\mathbb{R} \times A). \tag{27}$$

Assume the discrete time random process $z = (w, v) : \Omega \times Z \to \mathbb{R}^2$ to be such that empirical distributions converge to the function F_{wv} at every point of its continuity, i.e.

$$\lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} U(z - z_i(\omega)) = F_{wv}(z) \qquad w.P.1 \tag{28}$$

for every continuity point z of the function F_{wv}, where the mapping $U : \mathbb{R}^2 \to \{0, 1\}$ is given as

$$U(x_1, x_2) = \begin{cases} 1 & \text{if } x_1 \geq 0 \text{ and } x_2 \geq 0 \\ 0 & \text{otherwise} \end{cases}. \tag{29}$$

Also let the conditional distribution function $F_{w|v} : \mathbb{R}^2 \to [0, 1]$ defined by

$$F_{w|v}(d, v) = P_{w|v}((-\infty, d]), v) \tag{30}$$

be continuous at a point (d, v) where v belongs to the support of the mapping P_v.

Then, for the estimator $P^{r,m}_{w|v}$ defined in equation (23) the following is true:

$$\lim_{r\to 0}\left[\lim_{m\to\infty} P^{r,m}_{w|v}((-\infty, d], v)\right] = P_{w|v}((-\infty, d], v) \qquad w.P.1. \ \blacksquare \tag{31}$$

In Theorem 1 only a very general ergodic property has to be fulfilled by the data sequence requiring empirical measures to converge to the limit measure P_{wv}, which is not assumed even locally to possess a density function. The conditional distribution function $F_{w|v}$ defined by *Eq.* (30) needs to be continuous in the point of estimation,

which in most practical cases is fulfilled for almost every value of the explanatory variable v.

Viewed directly as the definition of a neural network, the estimator $P_{w|v}^{m,r}$, as discussed in the beginning of this section, yields a number of neurons equal to the number of data m and consequently a network performing no data compression at all. Compression is generally introduced by replacing the estimator $P_{w|v}^{m,r}$ by its compressed modification $P_C^{n,x,\rho,\beta,\gamma}$ defined by

$$P_C^{n,x,\rho,\beta,\gamma}((-\infty, d], v) = \frac{\sum_{i=1}^{n} \beta_i \cdot S\left(\frac{d - x_i^w}{\rho_i}\right) \cdot \phi\left(\frac{v - x_i^v}{\rho_i}\right)}{\sum_{i=1}^{n} \gamma_i \cdot \phi\left(\frac{v - x_i^v}{\rho_i}\right)}. \tag{32}$$

In this equation n denotes the number of neurons which is considered to be a design parameter restricted by $n \ll m$ in order to ensure a sufficient level of compression. The parameters $x_i = (x_i^w, x_i^v)$, ρ_i, β_i, γ_i are viewed as adjustable weights and offsets, subject to some training procedure projecting the statistical information of the data to the network parameters. The compression described above in general terms is implementable in a variety of ways, two representatives of which are discussed below.

A rather nonconstructive approach utilises the conditional distribution estimator $P_{w|v}^{m,r}$ as the target function in a pure supervised learning scheme, setting all parameters in the estimator $P_C^{n,x,\rho,\beta,\gamma}$ optimally with respect to some measure of distance between the two functions. If all parameters are trained from initially random settings, a long period of training, probably including several restarts, may be anticipated.

A far more constructive method introduces the compression to the kernel estimator $F_{wv}^{m,r}$ given by

$$F_{wv}^{m,r}(w, v) = \int_{-\infty}^{w} \int_{-\infty}^{v} h_{wv}^{m,r}(x, y)\, \mathrm{d}x\, \mathrm{d}y = \frac{1}{m} \sum_{i=1}^{m} S\left(\frac{w - w_i}{r}\right) \cdot S\left(\frac{v - v_i}{r}\right) \tag{33}$$

of the joint distribution function F_{wv}. This leads to the compressed estimator $F_C^{n,x,\rho,\alpha}$ defined as

$$F_C^{n,x,\rho,\alpha}(w, v) = \sum_{i=1}^{n} \alpha_i \cdot S\left(\frac{w - x_i^w}{\rho_i}\right) \cdot S\left(\frac{v - x_i^v}{\rho_i}\right). \tag{34}$$

If the parameter vectors (x_i^w, x_i^v) are generated randomly according to some joint distribution function F^*, the mapping $F_C^{n,x,\rho,\alpha}$ for $\alpha_i = \frac{1}{n}$ and $\rho_i = r$ provides a kernel estimator of the mapping F^*. The strategy therefore is to generate the

parameters (x_i^w, x_i^v) for the function F^* in order to be close to the function F_{wv}. This is accomplished by first drawing these parameters randomly from the training set (w_i, v_i). For this choice, as well as $\alpha_i = \frac{1}{n}$ and $\rho_i = r$, the compressed estimator $F_C^{n,x,\rho,\alpha}$ simply constitutes a kernel estimator of the function F_{wv}, based on a training set with n measurements, which of course is not satisfactory when a far larger number of data is available. Therefore subsequently self organising is imposed on the parameters, adjusting them from their initial random settings to points in the parameter space far less sensitive to the randomness of the initial draw. That is, if the self organising algorithm is efficient, the parameters (x_i^w, x_i^v) are placed so that the different sets of the corresponding Voronoi partition contain approximately equally many points from the training data set. This in turn implies that the random uncertainty of the estimator $F_C^{n,x,\rho,\alpha}$, for $\alpha_i = \frac{1}{n}$ and $\rho_i = r$ as well as for large n and m, is close to that of the function $F_{wv}^{m,r}$. Thus the magnitude of the training set m, and not the number of neurons n, determines the uncertainty of the estimator $F_C^{n,x,\rho,\alpha}$.

The smoothing parameters ρ_i might be set to a common value r fulfilling perhaps some limit relation with respect to the numbers m and n in order to ensure consistency when asymptotic properties of the training data sequence are known. Otherwise heuristics are applicable, such as defining ρ_i as the average Euclidean distance between (x_i^w, x_i^v) and its k nearest neighbours among $\{(x_j^w, x_j^v),\ j = 1, 2, \ldots, n,\ j \neq i\}$.

The parameters α_i are all initialised to the value $\frac{1}{n}$ following the above reasoning. Subsequently these parameters are imposed on a supervised training scheme in which the empirical joint distribution F^m defined by

$$F^m(w, v) = \frac{1}{m}\#\{(w_i, v_i) \ :\ w_i \leq w \text{ and } v_i \leq v, i \in \{1, 2, \ldots, m\}, \tag{35}$$

where # denotes the number of elements, serves as a target function. The training can be implemented by minimising the objective function $E^{x,\rho,\alpha}$ given as

$$E^{x,\rho,\alpha} = \frac{1}{m}\sum_{i=1}^{m} l\left(F_C^{n,x,\rho,\alpha}(w_i, v_i), F^m(w_i, v_i)\right) \cdot p(w_i, v_i), \tag{36}$$

where l is an appropriate loss function and p denotes an optional penalty function emphasising accuracy in certain domains. During the supervised training the remaining network parameters x_i^w, x_i^v and ρ_i could be either fixed or subject to training along with the constants α_i. Empirical studies indicate that only an insignificant improvement can be gained by following the second alternative, whereas computational effort is significantly increased.

When training has been completed, the distribution function estimator $F_C^{n,x,\rho,\alpha}$ is transformed into the corresponding conditional distribution $P_C^{n,x,\rho,\beta,\gamma}((-\infty,$

$d] \mid v)$. This can be achieved by the following parameter transformations:

$$\beta_i = \gamma_i = \alpha_i \cdot \rho_i. \tag{37}$$

Finally the conditional quantiles approximation $\hat{q}$ is found from

$$P_C^{n,x,\rho,\beta,\gamma}((-\infty, \hat{q}(v)] \mid v) = p, \tag{38}$$

which can easily be solved numerically.

4. Numerical Example

In this section a numerical example illustrating the performance of the proposed method will be presented. Data are generated artificially as follows:

1. the sequence $\{w_i,\ i = 1, 2, \ldots, 1000\}$ is defined by

$$w_i = \sin(v_i) + (0.1 \cdot v_i^2 + 1) \cdot e_i, \tag{39}$$

2. the sequence $\{v_i, i = 1, 2, \ldots, 1000\}$ is generated as the realisation of independent random variables, all uniformly distributed in the interval $[-5, 5]$,
3. the sequence $\{e_i,\ i = 1, 2, \ldots, 1000\}$ is generated as the realisation of independent random variables, all uniformly distributed in the interval $[-0.5, 0.5]$.

According to the above definitions, the conditional quantile of order p can be found as

$$q_{w|v}(v) = \sin(v) \cdot (0.1 \cdot v^2 + 1) \cdot (p - 0.5). \tag{40}$$

The generated sequence of training data $\{(w_i, v_i)\}$, as well as the theoretical conditional quantiles of order 0.2, 0.5 and 0.8, are depicted in *Fig. 1*.

A neural network of 50 neurons was trained to estimate the conditional quantiles. The network parameters are found in the following manner:

1. The parameters $\{(x_i^w, x_i^v),\ i = 1, 2, \ldots 50\}$ are initially chosen randomly from $\{(w_i, v_i)\}$, and subsequently adjusted by a self organising scheme, as described in the previous section. The initial and final positions of the parameters $\{(x_i^w, x_i^v)\}$ are shown in *Fig. 2*.
2. Each smoothing parameter ρ_i is found as the average distance of the parameters $\{(x_i^w, x_i^v)\}$ to the four nearest neighbours.
3. The parameters $\{\alpha_i,\ i = 1, 2, \ldots, 50\}$ are found by supervised training using the empirical joint distribution computed for training data, i.e. $\{F^m(w_i, v_i),\ i = 1, 2, \ldots, 1000\}$, as target values.

The theoretical conditional quantiles of orders 0.2, 0.5 and 0.8, as well as their estimators, can be seen in *Fig. 3*. The numerical results are judged acceptable to confirm the theoretical considerations carried out earlier in this paper.

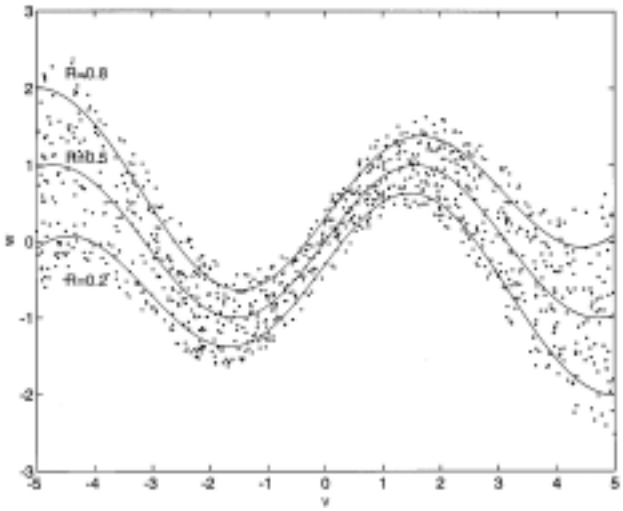

Fig. 1. Applied training data along with theoretical conditional quantiles

5. Conclusion

A neural network for estimating conditional quantiles has been constructed in the present paper. Although the network is designed on the basis of kernel estimation of joint probability density functions, a theory has been presented showing the network to be valid in more general settings, where only the continuity of conditional distribution functions, as well as a very general ergodic property of the training data, is assumed.

The problem of estimating conditional quantiles has been related to Bayes estimation in the case of a special asymmetric loss function, which would be feasible for application within a variety of areas in engineering, as well as science and economics. By estimating conditional quantiles, the neural network designed is applicable to the Bayes estimation problem.

An intermediate network version identical to a kernel estimator of a conditional distribution function may be viewed as a structural paradigm for a class of networks distinguished by size, parameter interpretation and the training algorithms applied to set network parameters. Data compression has been discussed, and a constructive method including both unsupervised as well as supervised learning has been suggested. The compression and training techniques suggested serve as representatives for a broad class of methods which might be applied for setting network parameters. This of course opens up paths for further research on the application of statistical identification methods for estimating optimal parameter

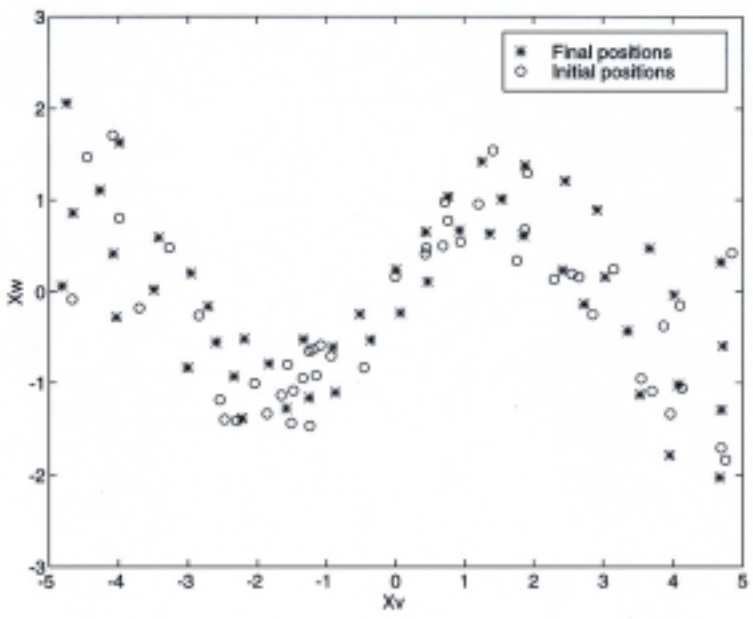

Fig. 2. Locations of 'X' parameters before and after self organising

settings, as well as for selecting network size and structure.

Acknowledgement

The authors would like to thank Professor László T. Kóczy for his creative comments and suggestions.

Appendix

(Proof of Theorem 1)

Assumption (28) guarantees weak convergence of empirical measures to the probability measure P_{wv}. This yields, according to equation (23) and Theorem 29.1 of

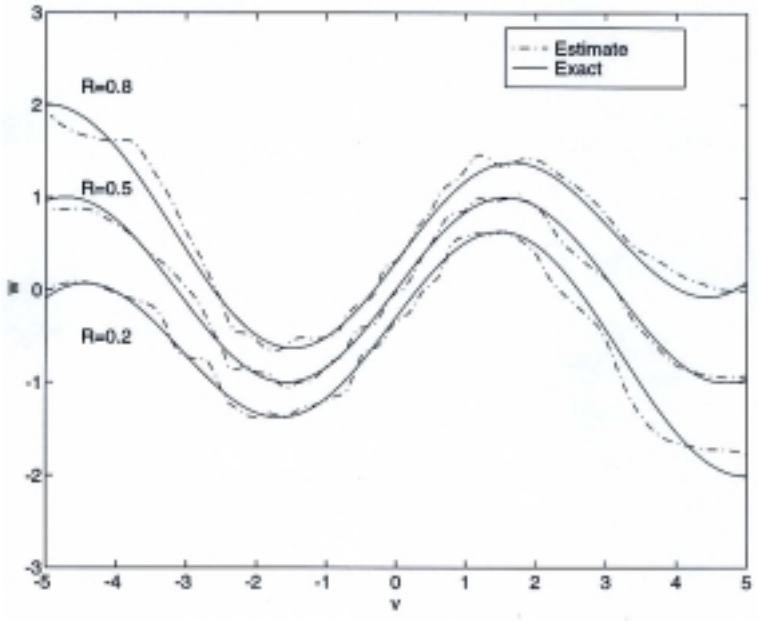

Fig. 3. Theoretical conditional quantiles along with estimators computed by the designed neural network

book (BILLINGSLEY, 1979):

$$F^r_{w|v}(d,v) = \lim_{m\to\infty} P^{r,m}_{w|v}((-\infty,d] \mid v) = \frac{\displaystyle\int_{\mathbb{R}^2} S\left(\frac{d-w}{r}\right)\cdot\phi\left(\frac{v-y}{r}\right) \mathrm{d}P_{wv}(w,y)}{\displaystyle\int_{\mathbb{R}^2}\phi\left(\frac{v-y}{r}\right)\mathrm{d}P_{wv}(w,y)}. \tag{41}$$

By continuity of the conditional distribution $F_{w|v}$ at the point (d, v), for every $\varepsilon > 0$ a number $\delta > 0$ exists such that

$$|F_{w|v}(d,v) - F_{w|v}(\tilde{d},y)| \le \varepsilon \qquad \text{for } |v-y| < \delta \text{ and } |d-\tilde{d}| \le \delta. \tag{42}$$

The general properties of the function S defined in equation (26), and the fact that the number v is in the support of the measure P_v, implies for fixed δ the existence of a number $r > 0$ yielding

$$1 - S\left(\frac{d}{r}\right) \le \varepsilon \quad \text{for } d \ge \delta \qquad \text{and} \qquad S\left(\frac{d}{r}\right) \le \varepsilon \quad \text{for } d < -\delta \tag{43}$$

as well as

$$\frac{\exp\left(-\frac{3\delta^2}{4r^2}\right)}{P_v\left(B\left(v,\frac{\delta}{2}\right)\right)} \leq \varepsilon. \tag{44}$$

If the number δ is fixed, the function $F^r_{w|v}(d,v)$ can be decomposed in the following way:

$$\begin{aligned} F^r_{w|v}(d,v) &= \frac{\int_{(-\infty,d-\delta]\times B(v,\delta)} \phi\left(\frac{v-y}{r}\right) \mathrm{d}P_{wv}(w,y)}{\int_{\mathbb{R}^2} \phi\left(\frac{v-y}{r}\right) \mathrm{d}P_{wv}(w,y)} + T_2 + T_3 + T_4 + T_5 \\ &= \frac{\int_{B(v,\delta)} F_{w|v}(d-\delta,y)\cdot\phi\left(\frac{v-y}{r}\right) \mathrm{d}P_v(y)}{\int_{\mathbb{R}} \phi\left(\frac{v-y}{r}\right) \mathrm{d}P_v(y)} + T_2 + T_3 + T_4 + T_5 \qquad (45) \\ &= \frac{\int_{B(v,\delta)} F_{w|v}(d,y)\cdot\phi\left(\frac{v-y}{r}\right) \mathrm{d}P_v(y)}{\int_{\mathbb{R}} \phi\left(\frac{v-y}{r}\right) \mathrm{d}P_v(y)} + T_1 + T_2 + T_3 + T_4 + T_5, \end{aligned}$$

where

$$T_1 = \frac{\int_{B(v,\delta)} (F_{w|v}(d-\delta,y) - F_{w|v}(d,v))\cdot\phi\left(\frac{v-y}{r}\right) \mathrm{d}P_v(y)}{\int_{\mathbb{R}} \phi\left(\frac{v-y}{r}\right) \mathrm{d}P_v(y)}, \tag{46}$$

$$T_2 = \frac{\int_{(-\infty,d-\delta]\times B(v,\delta)} \left(S\left(\frac{d-w}{r}\right)-1\right)\cdot\phi\left(\frac{v-y}{r}\right) \mathrm{d}P_{wv}(w,y)}{\int_{\mathbb{R}} \phi\left(\frac{v-y}{r}\right) \mathrm{d}P_v(y)}, \tag{47}$$

$$T_3 = \frac{\int_{(d-\delta,d+\delta]\times B(v,\delta)} S\left(\frac{d-w}{r}\right)\cdot\phi\left(\frac{v-y}{r}\right) \mathrm{d}P_{wv}(w,y)}{\int_{\mathbb{R}} \phi\left(\frac{v-y}{r}\right) \mathrm{d}P_v(y)}, \tag{48}$$

$$T_4 = \frac{\int_{(d+\delta,\infty]\times B(v,\delta)} S\left(\frac{d-w}{r}\right)\cdot\phi\left(\frac{v-y}{r}\right) \mathrm{d}P_{wv}(w,y)}{\int_{\mathbb{R}} \phi\left(\frac{v-y}{r}\right) \mathrm{d}P_v(y)}, \tag{49}$$

$$T_5 = \frac{\int_{B(v,\delta)^C} S\left(\frac{d-w}{r}\right) \cdot \phi\left(\frac{v-y}{r}\right) \mathrm{d}P_{wv}(w,y)}{\int_{\mathbb{R}} \phi\left(\frac{v-y}{r}\right) \mathrm{d}P_v(y)}, \tag{50}$$

while the superscript 'C' denotes the complementary set. Inequalities (42) and (43) directly yield the bounds

$$|T_1| \leq \varepsilon, \tag{51}$$
$$|T_2| \leq \varepsilon, \tag{52}$$
$$|T_4| \leq \varepsilon. \tag{53}$$

By definition of the conditional probability $F_{w|v}(d,v)$, the remaining term T_3 can be rewritten as

$$T_3 = \frac{\int_{(d-\delta,d+\delta]\times B(v,\delta)} S\left(\frac{d-w}{r}\right) \cdot \phi\left(\frac{v-y}{r}\right) \mathrm{d}P_{wv}(w,y)}{\int_{\mathbb{R}} \phi\left(\frac{v-y}{r}\right) \mathrm{d}P_v(y)}$$
$$= \frac{\int_{B(v,\delta)^C} (F_{w|v}(d+\delta,v) - F_{w|v}(d-\delta,v)) \cdot S\left(\frac{d-w}{r}\right) \cdot \phi\left(\frac{v-y}{r}\right) \mathrm{d}P_{wv}(w,y)}{\int_{\mathbb{R}} \phi\left(\frac{v-y}{r}\right) \mathrm{d}P_v(y)}. \tag{54}$$

Therefore the application of formula (42) and the triangle inequality yields

$$|T_3| \leq 2 \cdot \varepsilon. \tag{55}$$

Now obviously

$$F_{w|v}(d,v) = \frac{\int_{\mathbb{R}} F_{w|v}(d,v) \cdot \phi\left(\frac{v-y}{r}\right) \mathrm{d}P_v(y)}{\int_{\mathbb{R}} \phi\left(\frac{v-y}{r}\right) \mathrm{d}P_v(y)} \tag{56}$$

and subtracting equations (56) and (45) it leads to

$$F_{w|v}(d,v) - F^r_{w|v}(d,v) = \frac{\int_{B(v,\delta)^C} F_{w|v}(d,v) \cdot \phi\left(\frac{v-y}{r}\right) \mathrm{d}P_v(y)}{\int_{\mathbb{R}} \phi\left(\frac{v-y}{r}\right) \mathrm{d}P_v(y)}$$
$$-T_1 - T_2 - T_3 - T_4 - T_5. \tag{57}$$

The functions $F_{w|v}(d, v)$ and S are both numerically bounded by 1. Along with inequalities (51)–(53) and (55), this implies

$$\begin{aligned} |F_{w|v}(d,v) - F^r_{w|v}(d,v)| &\leq \frac{\int_{B(v,\delta)^C} F_{w|v}(d,v) \cdot \phi\left(\frac{v-y}{r}\right) \mathrm{d}P_v(y)}{\int_{\mathbb{R}} \phi\left(\frac{v-y}{r}\right) \mathrm{d}P_v(y)} + 5 \cdot \varepsilon + |T_5| \\ &\leq 2 \cdot \frac{\int_{B(v,\delta)^C} \phi\left(\frac{v-y}{r}\right) \mathrm{d}P_v(y)}{\int_{\mathbb{R}} \phi\left(\frac{v-y}{r}\right) \mathrm{d}P_v(y)} + 5 \cdot \varepsilon. \end{aligned} \tag{58}$$

For $V \in B(v, \delta)^C$ obviously

$$\left(\frac{v-y}{r}\right)^2 \geq \frac{\delta^2}{r^2} \tag{59}$$

and for $V \in B\left(v, \frac{\delta}{2}\right)$ similarly

$$\left(\frac{v-y}{r}\right)^2 \leq \frac{\delta^2}{4r^2}. \tag{60}$$

With the function ϕ defined as in *Eq.* (25), the following bounds are obtained from inequalities (59) and (60):

$$\int_{B(v,\delta)^C} \phi\left(\frac{v-y}{r}\right) \mathrm{d}P_v(y) \leq \exp\left(-\frac{\delta^2}{r^2}\right), \tag{61}$$

$$\begin{aligned} \int_{\mathbb{R}} \phi\left(\frac{v-y}{r}\right) \mathrm{d}P_v(y) &\geq \int_{B\left(v,\frac{\delta}{2}\right)} \phi\left(\frac{v-y}{r}\right) \mathrm{d}P_v(y) \\ &\geq \exp\left(-\frac{\delta^2}{4r^2}\right) \cdot P_v\left(B\left(v, \frac{\delta}{2}\right)\right). \end{aligned} \tag{62}$$

This, together with formulas (44) and (58), yields

$$|F_{w|v}(d,v) - F^r_{w|v}(d,v)| \leq 2 \cdot \frac{\exp\left(-\frac{\delta^2}{r^2}\right)}{\exp\left(-\frac{\delta^2}{4r^2}\right) \cdot P_v\left(B\left(v, \frac{\delta}{2}\right)\right)} + 5 \cdot \varepsilon$$

$$= 2 \cdot \frac{\exp\left(-\frac{3\delta^2}{4r^2}\right)}{P_v\left(B\left(v, \frac{\delta}{2}\right)\right)} + 5 \cdot \varepsilon \leq 7 \cdot \varepsilon, \qquad (63)$$

by which Theorem 1 is finally proved.

References

[1] BILLINGSLEY, P. (1979): Probability and Measure. Wiley, New York.

[2] FISZ, M. (1963): Probability Theory and Mathematical Statistics. Wiley, New York.

[3] KULCZYCKI, P. (1993): Time-Optimal Stochastic Positional Control. *Proc. IFAC 12th World Congress*, Vol. 7, pp. 443–448.

[4] KULCZYCKI, P. – SCHIØLER, H. (1994): Parameter Identification by Bayes Decision and Neural Networks. *Proc. 10th IFAC Symposium on System Identification*, Vol. 3, pp. 477–482.

[5] LEHMANN, E. L. (1983): Theory of Point Estimation. Wiley, New York.

[6] MOODY, J. – DARKEN, C. (1989): Fast Learning in Networks of Locally Tuned Processing Units. *Neural Computation*, Vol. 1, pp. 281–294.

[7] NIELSEN, R. H. (1987): Counter Propagation Networks. *Proc. the First IEEE International Conference on Neural Networks*, Vol. 2, pp. 19–33.

[8] PARZEN, E. (1962): On Estimation of a Probability Density Function and Mode. *Annals of Mathematical Statistics*, Vol. 33, pp. 1065–1076.

[9] RUCK, D. W. – ROGERS, S. K. – KABRISKY, M. – OXLEY, M. E. – SUTER, B. W. (1990): The Multilayer Perceptron as an Approximation to a Bayes Optimal Discriminant Function. *IEEE Transactions on Neural Networks*, Vol. 1, pp. 296–298.

[10] RUMELHART, D. E. – MCCLELLAND, J. (1986): Parallel Distributed Processing. MIT Press.

[11] SCHIØLER, H. – HARTMANN, U. (1992): Mapping Neural Network Derived from the Parzen Window Estimator. *Neural Networks*, Vol. 5, pp. 903–909.

[12] SPECHT, D. F. (1991): A General Regression Neural Network. *IEEE Transactions on Neural Networks*, Vol. 2, pp. 568–576.

[13] SPECHT, D. F. (1988): Probabilistic Neural Networks for Classification, Mapping, or Associative Memory. *Proc. IEEE International Conference on Neural Networks*, Vol. 1, pp. 525–533.

[14] WHITE, H. (1990): Connectionist Nonparametric Regression: Multi Layer Feedforward Network can Learn Arbitrary Mappings. *Neural Networks*, Vol. 3, pp. 535–549.