Fuzzy Systems in Random Environment. Asymptotics and Some Economic Applications[1]

Manuel Landajo
Unit of Statistics and Econometrics, Department of Applied Economics,
University of Oviedo,
33006 Oviedo (Spain).
e-mail: landajo@uniovi.es

Abstract

Model-free regression capabilities of fuzzy systems are studied in this paper. Under general conditions, fuzzy systems constructed by the usual expedient of combining statistical data and expert knowledge, are capable of learning arbitrary regression surfaces and their derivatives to any arbitrary finite order. Expert knowledge provides bounds which usefully constrain the estimation processes, potentially avoiding ill-posed solutions and meaningless fuzzy rules. Results include least squares and conditional quantile regressions. An application to quantile–based profitability forecasting is included.

Keywords: fuzzy systems, nonparametric statistics, large samples, least-squares learning, LAD learning, quantile regression, economic modelling, ROA.

[1] Version: June 10, 2007. An early version of this paper was presented in a seminar at the European Soft Computing Centre, Mieres, Spain. The author gratefully acknowledges helpful suggestions by participants at the session.

Contents

1 Introduction and motive

The paradigm of fuzzy systems (FSs) has developed in the last four decades, with both a remarkable amount of theoretical research and an increasing variety of applications in multiple fields. FSs have their foundation in principles of fuzzy logic. However, many useful properties of FSs can be studied by conventional (i.e., mathematical and statistical) techniques. Hence, a number of recent contributions (e.g., Kosko, 1991; Wang and Mendel, 1992; Mao *et al.*, 1997; the list is far from complete) have shown that FSs are universal approximators in a number of function spaces of practical relevance. This gives a clear-cut mathematical meaning to the vague notion of "flexibility" of this kind of models, providing an analytical foundation to the intuition that a FS with a sufficiently high number of rules is, in principle, capable of mimicking arbitrary functions. Of course, the above universal approximation (UA) property —which is also shared by many other mathematical devices, such as algebraic/trigonometric polynomials, wavelets and artificial neural networks— only ensures that good approximators for the mappings of interest exist, although usually does not provide any means to construct them. This last issue has been a concern for both theorists and practitioners, who rapidly became aware of the usefulness of automatic or semiautomatic methods for fuzzy-rule building.

This drives us to the other source of flexibility of FSs, namely, their ability to integrate different sources of information (expert knowledge/statistical data). In many practical applications, both sources of knowledge are combined in order to *efficiently* construct an FS with optimal —or, at least, adequate— performance. A vast array of literature has appeared on the subject, and a number of statistical/neural learning mechanisms —including, e.g., batch and on-line algorithms for rule fine-tuning, and cross-validation-based techniques for model selection and evaluation of model performance— have been adapted, and often specifically designed, to enhance the performance of FSs (e.g., Takagi and Sugeno, 1985; Sugeno and Tanaka, 1991; Kosko, 1991, 1992; Wang, 1994; Jang, 1993; Watanabe and Imaizumi, 1997; among many others). Endowed with these algorithms, FSs provide a means to integrate (possibly imprecise and subjective) expert knowledge and statistical information in order to achieve higher performance than available by using purely subjective information.

In contrast to the study of UA capabilities, which is now a well-developed field, the analysis of the statistical properties of FSs endowed with these "training" devices has remained a largely unexplored area. In fields such as engineering FSs are required to coexist with other approaches whose statistical properties are well established. The successful performance of FSs in those environments evidently indicates that FSs must also have good statistical properties, although formal mathematical proofs for this fact were lacking until very recently. Many fuzzy theorists (e.g., Kosko, 1992; Wang, 1994) soon became aware of the above facts.[2] Although statistical considerations are not explicitly present when a practitioner is trying to build a FS in a practical application, the ingredients he/she uses for such a purpose (i.e., a data set, expert information, a training algorithm) may be easily nested into a statistical framework. An advantage of the approach we propose in this paper is that it permits us to analyze FSs in terms of the same methodological framework which includes, e.g., state space models, artificial neural networks and many other paradigms used in the fields where FSs find applications. The relevant ideas closely parallel the statistical theory developed for artificial neural networks (ANNs).[3]

The reader may argue at this point that, certainly, FSs are very different from ANNs. Neural networks are just computational black-boxes, scarcely user-friendly and interpretable, and (at most) with a limited ability to incorporate expert knowledge. As pointed out in literature, the theoretical goal of learning *only* from data, as required for ANNs and other computational black-boxes, often seems unrealistic, essentially because of the enormous amounts of data which may be required (for a thorough discussion see Geman *et al.*, 1992). As to be detailed in next sections, the ability of FSs to incorporate expert knowledge (i.e., *constraints* on the learning process) can turn the learning task much easier. Typically, ANNs (and most nonparametric regression devices) only are consistent estimators for general regression surfaces under rather stringent conditions (basically, the permitted complexity of the ANNs must increase with sample size at a sufficiently slow rate). As to be seen below, in the case of FSs, expert knowledge permits consistent learning under much milder conditions. This peculiarity of FSs, much stressed in literature (e.g. Nguyen *et al.*, 1996), drives us far

[2] L.X. Wang emphasized this issue with particular insight:

> "'Fuzzy systems are constructed and justified based on fuzzy logic; very little is known about their statistical properties when they are used in a random environment. (...) Up to now, it seems that the successful application of fuzzy logic systems is to the control of industrial processes where random noise always exists. Therefore, knowing the statistical properties of various kinds of fuzzy logic systems is important."(Wang, 1994, chapter 7)

[3] A very general theory concerning the statistical properties of ANNs is now available (e.g., White, 1989; 1990; Kuan and White, 1994; Chen and Shen, 1998).

away from the world of computational black boxes and *purely* nonparametric methods.[4] In this paper we elaborate on the above arguments. We shall argue that the expedient of combining statistical data and expert knowledge is not only a practically convenient device, but also endows FSs with useful statistical properties. In particular, we will show that FSs are capable to learn consistently arbitrary regression surfaces.

The rest of the paper has the following structure. We start in Section 2 with a brief review of universal approximation capabilities of FSs. The basic elements of FS-based learning in stochastic contexts are outlined in Section 3. Section 4 contains the theoretical results of the paper: first, the classical problem of least squares learning, and then we focus on nonparametric conditional quantile regression (including median regressions as a particular instance). Our results permit consistent estimation of functions (and, when required, their derivatives to any finite order). In Section 4 some simulation results and an application in the field of profitability forecasting are presented. Mathematical proofs are collected in Appendix.

2 Fuzzy systems and universal approximation capabilities. A brief excursion

Many (but not *all*) of the available results on UA properties of FSs refer to the class of additive FS. These have very close cousins in mathematics, under the form of convolution operators. For a function $f : \mathbb{R}^d \to \mathbb{R}$ the following convolution-based approximant can be defined:

$$\hat{f}_\sigma(x) = \int_{\mathbb{R}^d} f(\mu)\rho\left(\frac{\mu - x}{\sigma}\right) d\mu \tag{1}$$

with ρ being a probability density function and $\sigma > 0$ being a smoothing parameter. $\hat{f}_\sigma(x)$ approximates $f(x)$ by the simple expedient of computing a local average of values of f around x. Classical results on function approximation by using convolutions are well-known

[4] It is in this precise sense that FSs, and in particular additive FSs, differ from computational black-boxes such as kernel regressions and radial basis functions, which may even be functionally equivalent (in the sense of being constructed on the basis of the same classes of functions). The capability of FSs to process expert information would point out some potential commonalities of FSs with Bayeasin statistics, although in the latter field subjective information has a very specific nature (prior beliefs), and is processed in a very formalized way (i.e., by applying Bayes Theorem).

(e.g., Devore and Lorentz, 1993). The above convolution may be discretized as follows:

$$\tilde{f}_{m,\sigma}(x) = \frac{\sum_{j=1}^{m} f(\mu_j)\rho\left(\frac{\mu_j - x}{\sigma}\right)}{\sum_{j=1}^{m} \rho\left(\frac{\mu_j - x}{\sigma}\right)} \tag{2}$$

with $\{\mu_j, j = 1, \ldots, m\}$ being a grid of points in the domain of f. The functional form (2) basically coincides with that of additive FSs.[5] Classes of functions obtained by discretizing a convolution operator inherit the approximation properties of the original operator. This is the basis for most of the mathematical proofs of UA properties for additive FSs, and has permitted researchers to derive UA theorems for FSs in many function spaces of practical relevance (e.g., the instances provided in Section 1 above). Since it is required in the following sections, we adopt the following definition for the UA property.

Definition (UA property). Let Θ be a space of functions $\mathbb{X} \to \mathbb{R}$ endowed with a metric d, and let Θ_m be a class of FSs with m rules such that $\Theta_m \subset \Theta$. The sequence $\{\Theta_m\}$ is said to satisfy the UA property in Θ when, for any $\theta^* \in \Theta$, a sequence of FSs $\{\theta_m^* \in \Theta_m\}$ exists such that $lim_{m\to\infty} d\left(\theta_m^*, \theta^*\right) = 0$. □

The above definition encompasses general (additive and non-additive) FS structures. The statistical learning capabilities studied in this paper are referred to generic classes of FSs satisfying a suitable UA property. In particular, we shall consider UA properties based on uniform approximation in spaces of continuous functions. (Details are provided in the following sections.)

The discretization (2) provides a straightforward method to build FSs, although sometimes a sufficiently fine grid of data $\{(x_i, f(x_i))\,|i = 1, \ldots, n\}$ may be unavailable, and very often, instead of a deterministic setting such as that in Figure 1 below, we may have to cope with something less favorable, as in Figure 2, i.e., a set of observations $\{(x_i, y_i)|\, i = \ldots, n\}$ with the x_i values irregularly spread along the domain of f, and the observed values possibly being corrupted by some kind of noise ε_i, i.e., $y_i = f(x_i) + \varepsilon_i$. Effective methods to filter noise and provide an accurate approximant to f are required, and the naive quasi-interpolation scheme (2) may be problematic (e.g., a careful control of the smoothing parameter σ may

[5] The same structure is also used within kernel regressions (e.g., the classical Nadaraya-Watson estimator).

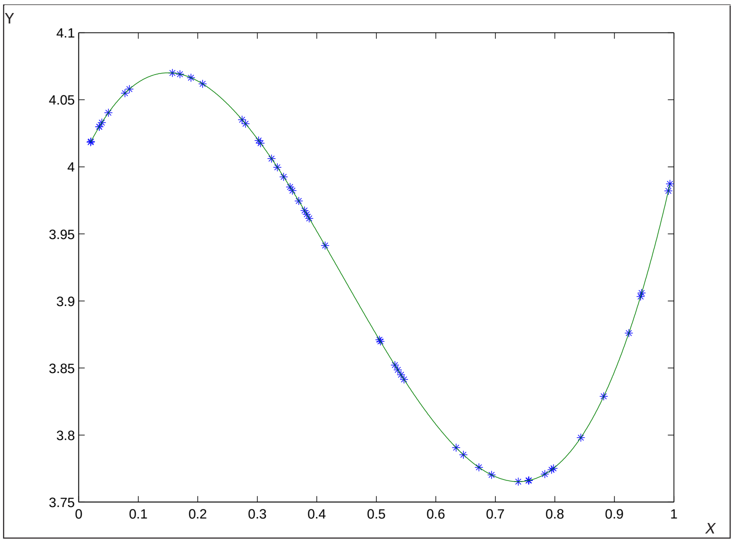

Figure 1: A learning problem in noiseless environment. (Asterisks denote observed values. The underlying mapping is indicated by the continuous line.)

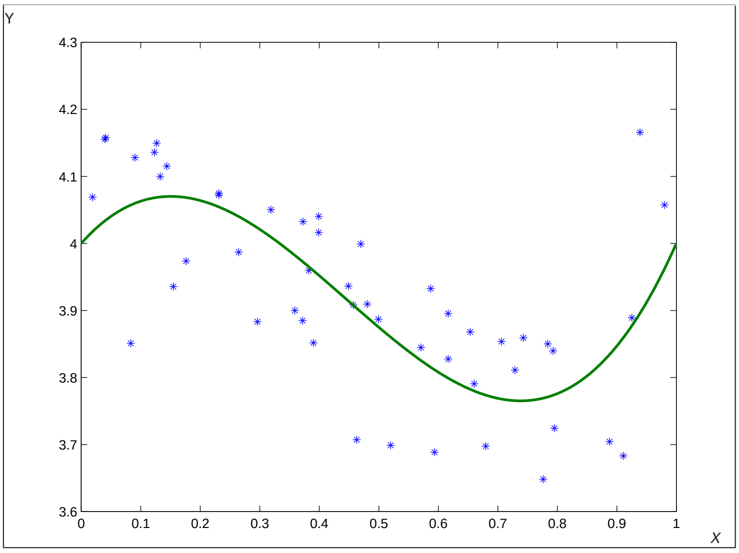

Figure 2: A learning problem in noisy environment.

be required).[6] This drives us to the next section.

3 The ingredients of learning. Connections with sieve extremum estimation

The elements of the learning problem are as the follows:

- **The data generating process**, i.e., a random mechanism which is assumed to generate (crisp) data to be used by the modeler.[7]
- **The mapping to be learned.** It is a *non-random* characteristic of the observed system (e.g., a regression surface which permits optimal prediction/control in any sense).
- **A finite data set**, generated by the system under study.
- **Expert knowledge**, under the form of *crisp* constraints provided by the researcher. In particular, we only assume that some bounds are available that constrain the permitted values and the oscillatory behavior of the fitted FSs.[8]
- **A fuzzy system structure**, i.e., a class of FSs having a suitable UA property.
- **A learning paradigm**. It includes the goal function to be optimized (algorithmic details are not relevant at this level).

The result of the above modelling effort is a fuzzy-rule-based system which fits data in the above specified sense, also satisfying the set of constraints provided by expert knowledge. We will analyze the behavior of the whole mechanism as both sample size and model complexity (indexed by the number of fuzzy rules) grow to infinity. We will show that, under

[6] An additional inconvenience is that fuzzy rule-based systems built by using the "one-datum-one-rule" approach are likely to be unnecessarily complex, since close observations basically providing the same information may be grouped together into a single fuzzy rule, in order to achieve a "parsimonious" approximation to the desired mapping. Typically, FSs have a number of rules much lower than the size of the samples used in their construction.

[7] True randomness is not required. The results are valid for deterministic —e.g., chaotic— systems for which the ergodic distributions required in the following sections exist.

[8] The bounds must be *crisp*, although they are permitted to be more or less loose, depending on the available knowledge. The approach altogether differs from Bayesian methods, where *a priori* information is *probabilistic* in nature.

general conditions, the resulting sequence of FSs converges to the desired regression surface. This is much stronger a result than the typical argument in parametric statistics (e.g., classical linear regression with finite dimension parameter), since the regression surface is now a generic mapping. Expert knowledge is required to provide some qualitative information on the function to be learned, but the requirements are much less demanding than in classical statistics, where a very precise knowledge of the problem —embodied in a parametric specification with only a small number of free parameters to be estimated— is assumed.

3.1 The method of sieves. Basic statistical background

A key issue is that a FS is only an approximate model, in the sense that the true object to be learned is a generic function (e.g., $y = x^a$) which would generally require an *infinite* set of fuzzy rules to be represented, and of course only FSs with a *finite* number of rules can be constructed on the basis of a finite data set. FSs have been seen as approximate models for more complex systems since the earliest contributions (Zadeh, 1973). On the other hand, even when the mapping to be learned can be represented by a finite FS, as the available data set is finite, the FSs we can construct are only estimates of the true (population) mapping.[9] The emphasis of fuzzy modelling on finding "simple" structures to achieve representations of complex phenomena also appears in other branches of mathematics (e.g., approximation theory) and statistics (e.g., nonparametric estimation). A connection which is of interest to us is that of FSs with the statistical method of sieves, proposed by U. Grenander (1981).[10] Not surprisingly, sieve estimation methods apply ideas which are very close (in spirit) to the rationale of FSs. The former provide methods to obtain consistent estimators in "too large" parameter spaces, whereas the latter were conceived as approximate models for complex systems for which a "complete" model is unavailable.

In the statistics literature many estimators are obtained as solutions to optimization problems. Typically, we wish to estimate or *learn*, on the basis of a given data set, a parameter

[9] In statistical jargon, the effect of using a finite set of fuzzy rules to approximate a function which requires an infinite set of rules to be represented is *bias*. The effects of finite sample size had to do with *variance* of the estimates.

[10] The basic idea of the method of sieves consists of replacing the whole parameter space by an adequately chosen sequence (called *sieve*) of "simpler" parameter spaces which have the UA property in the whole parameter space (e.g., Geman and Hwang, 1982). Under general conditions, estimates computed on the chosen sieve are consistent for the desired parameter (typically, a generic regression surface). Most classes of nonparametric estimators used in statistics may be seen as sieve estimators. These include modelling paradigms whose origin was far from statistics, such as artificial neural networks (e.g., White, 1990).

θ^* which is a point in a suitable parameter space Θ. Very often θ^* solves a population optimization problem, such as

$$\min_{\theta \in \Theta} Q(\theta) \tag{3}$$

for a suitably chosen criterion $Q(\cdot)$ (e.g., expected squared error). An *extremum estimator* $\hat{\theta}_n$ is defined as an (approximate and measurable) minimizer of the sample criterion $Q_n(\cdot)$, i.e.,

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} Q_n(\theta) \tag{4}$$

with $Q_n(\cdot)$ converging in probability to $Q(\cdot)$ as the sample size n grows to infinity (in the following sections we shall use " $\xrightarrow{P}$ " to denote convergence in probability under the measure P). Very often the above optimization (4) is not carried out over the whole parameter space, but on a suitably chosen subset Θ_m, whose complexity (e.g., the dimension of the subset) is indexed by m. If $\{\Theta_m\}$ has the UA property in Θ, the sequence of approximating spaces $\{\Theta_m\}$ is called a *sieve*, and the approximate minimizer $\hat{\theta}_n^m = \arg\min_{\theta \in \Theta_m} Q_n(\theta)$ is called a *sieve extremum estimator*. Under general conditions sieve estimators are consistent for θ^* as $n \to \infty$.

In a recent paper (Landajo, 2004) we applied the sieve framework to analyze some learning capabilities of fuzzy systems in stochastic environment. The basic idea consists of permitting the number of fuzzy rules to increase with the available sample size (see Figure 3 below). In next section we improve on these results. We obtain consistency in terms of a stronger norm (involving approximation of derivatives), and extend previous least-absolute-deviation learning results in order to permit estimation of conditional quantiles. We shall borrow ideas and results from the statistical literature on ANN models (basically, Gallant and White, 1992) and nonparametric econometrics (in particular, Newey and Powell, 2003).

4 Model-free learning by using fuzzy systems

We shall consider estimation in parameter space which is both relatively simple and sufficiently general to include a number of interesting practical situations. In particular, we assume that the mapping θ^* to be learned is an element of Θ, a subset of the space $C^r(\mathbb{X})$ of $\mathbb{X} \to \mathbb{R}$ mappings with continuous partial derivatives up to the r-th order (Θ is a func-

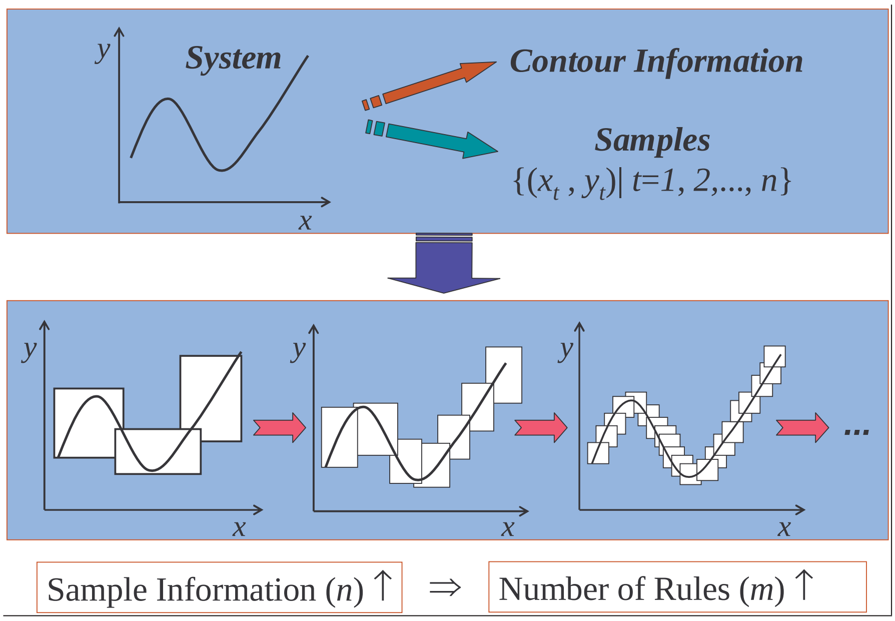

Figure 3: Ingredients of learning in fuzzy-rule-based sieves.

tion space, instead of the usual finite dimension spaces of parametric statistics). We follow the standard notation, and for any multi-index $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)$ we define the order of derivation as $|\boldsymbol{\alpha}| = \alpha_1 + \cdots + \alpha_d$, and denote by $D^{\boldsymbol{\alpha}}\theta(\cdot)$ the corresponding partial derivative of $\theta(\cdot)$ —the usual convention $D^{\mathbf{0}}\theta(\cdot) = \theta(\cdot)$ is applied—. We endow Θ with the metric $d(\cdot, \cdot)$ induced by the norm $\|\theta\| = \max\limits_{0 \leq |\alpha| \leq r} \sup\limits_{x \in int(\mathbb{X})} |D^{\boldsymbol{\alpha}}\theta(x)|$,[11] for some $r \geq 0$ (i.e., for any $\theta, \theta' \in \Theta$, $d(\theta, \theta') = \|\theta' - \theta\|$).

4.1 Least squares learning

The data generating process (DGP) is as follows: we consider a data set $D = \{(x_t, y_t,)\,|\, t = 1, \ldots, n\}$, with $x_t \in \mathbb{X} \subset \mathbb{R}^d$ (a row vector), and $y_i \in \mathbb{R}$. D is assumed to be a finite realization of a stochastic process $\{(X_t, Y_t)\,|\, t = 1, 2, \ldots\}$, with values of Y_t generated by the following regression structure:

$$Y_t \quad = \theta^* (X_t) + \sigma^* (X_t)\, \varepsilon_t, \qquad t = 1, ..., n; \;\; n = 1, 2, ... \tag{5}$$

with $\theta^*(\cdot)$ being the mapping of interest (a conditional expectation), which is assumed to belong to the parameter space Θ. $\sigma^*(\cdot) : \mathbb{X} \rightarrow \mathbb{R}_+$ is the square root of the conditional

[11] $int(\mathbb{X})$ denotes the set of interior points of $\mathbb{X}$.

variance of Y_t given values of X, i.e. $Var\left(Y_t|X_t = x\right) = \sigma^*(x)^2$, and ε_t is a random noise term with null expectation. We impose the following catalogue of assumptions:

Assumption 1. (**i**) The underlying probability space $(\Omega, \mathbb{F}, P)$ is complete, with the process $\{(X_t, \varepsilon_t) \ | t = 1, 2, ...\}$ being independent identically distributed (i.i.d.) and the random sequences $\{X_t | \, i = 1, 2, \ldots\}$ and $\{\varepsilon_t | \, i = 1, 2, \ldots\}$ being mutually independent. In addition, (**ii**) the error process has expectation $E(\varepsilon_t) = 0$ and $Var(\varepsilon_t) = \sigma_\varepsilon^2 < \infty$, and (**iii**) $\mathbb{X} \subset \mathbb{R}^d$ is a convex compact set with nonempty interior, and the marginal distribution of X_t, denoted as $P_X\left(\cdot\right)$, satisfies the requirement that $P_X(\mathcal{O}) > 0$ for any nonempty open $\mathcal{O} \subset \mathbb{X}$.

Assumption 2. (**i**) $\theta^* \in \Theta$, a compact subset of $C^r(\mathbb{X})$, endowed with the norm $\| \cdot \|$. In addition, (**ii**) $\sigma^*(\cdot)$ is a continuous $\mathbb{X} \rightarrow \mathbb{R}_+$ mapping.

Assumption 3. (**i**) $\{\Theta_m\}$ is a sequence of compact subsets of Θ satisfying the UA property with respect to $\| \cdot \|$. In addition, (**ii**) $\hat{m}$ is an $\Omega \rightarrow \mathbb{N} = \{1, 2, \ldots\}$ mapping which satisfies $\hat{m} \xrightarrow{P} \infty$ as $n \rightarrow \infty$.

Assumption 4. (**i**) θ^* is the unique minimizer on Θ of the function $Q : \Theta \rightarrow \mathbb{R}_+$ defined as follows: $Q(\theta) = \Sigma + \int\limits_{\mathbb{X}} \left[\theta^*(x) - \theta(x)\right]^2 P_X(dx)$, with $\Sigma = \sigma_\varepsilon^2 \int\limits_{\mathbb{X}} \sigma^*(x)^2 P_X(dx)$. □

We will analyze the limiting properties of the class of FS-based sive considered in Assumption 3 above when applied to estimate θ^* by using least squares learning.

Theorem 1. Under Assumptions 1 to 4, let $\hat{\theta}_{\hat{m}}$ be a solution to the problem

$$\min_{\theta \in \Theta_{\hat{m}}} Q_n(\theta) \equiv n^{-1} \sum_{t=1}^{n} [Y_t - \theta(X_t)]^2. \tag{6}$$

Then $\left\| \hat{\theta}_{\hat{m}} - \theta^* \right\| \xrightarrow{P} 0$ as $n \rightarrow \infty$. □

Comments

Assumption 1 defines a generic least-squares regression context. The compactness requirement in Assumption 2 embodies the expert knowledge *a priori* available. Some instances are provided below.

Model complexity $\hat{m}$ in Assumption 3 may be chosen by any deterministic rule —e.g., by any increasing function of n which grows to infinity with sample size—, but it is also permitted to be data-driven. The only technical requirement is that as n grows to infinity the selected number rules also goes to infinity with probability approaching one. This includes standard criteria for model selection, such as cross-validation (e.g., Sugeno and Tanaka, 1991) or complexity penalization mechanisms (e.g., some kind of information criterion, such as Akaike's AIC, or Schwarz's BIC), which can be applied to a range of permitted model complexities $m \in \{m_{n-}, \ldots, m_{n+}\}$, with $m_{n-} \to \infty$ as $n \to \infty$.[12] The above general-purpose result may be particularized to many especial contexts of interest.

Example 1 (Function estimation in a SISO case). We may take a bounded interval $X = [a, b]$ (with $a < b$) and let Θ be the space of continuous $[a, b] \to \mathbb{R}$ mappings, endowed with supnorm, i.e., $\|\theta\| = \sup_{x\in[a,b]} |\theta(x)|$. Compactness is achieved by imposing suitable bounds on the values of the mapping to be learned and its variation. For instance, for *a priori* known bounds $B_1, B_2 < \infty$ we may have $sup_{x\in[a,b]} |\theta(x)| \leq B_1$ and $|\theta(x_1) - \theta(x_2)| \leq B_2 |x_1 - x_2|$ for any $x_1, x_2 \in [a, b]$.[13] Constants B_1 and B_2 embody the available expert knowledge, and are imposed both on Θ and on the sieved FSs we construct to approximate θ^*. For instance, a very simple sieve we may chose may be based on additive FSs with symmetric triangular membership functions. We may select as Θ_m the class of FSs with the following structure:

$$g_m(x, \boldsymbol{\delta}) = \frac{\sum\limits_{j=1}^{m} \delta_j T\left(\frac{x-\mu_j}{\sigma}\right)}{\sum\limits_{j=1}^{m} T\left(\frac{x-\mu_j}{\sigma}\right)} \tag{7}$$

[12] Measurability issues have not been considered in this paper. In the case when $\hat{m}$ is a deterministic function of sample size, e.g., $\hat{m} = \left[n^{1/3}\right]$, with $[\cdot]$ denoting the integer part function, Theorem 2.2 in White and Wooldridge (1991) applies to ensure that measurable minimizers of the sample criterion function exist. When $\hat{m}$ is random (e.g., obtained by cross-validation), general results in Stinchcombe and White (1992) may be applied to ensure measurability or almost-measurability.

[13] Arzelá-Ascoli Theorem ensures that with these bounds Θ is a totally bounded subset of the space of continuous $[a, b] \to \mathbb{R}$ mappings endowed with supnorm (e.g., Adams, 1975, Chapter 1).

with $x \in [a,b]$, $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m) \in \mathbb{R}^m$, $m \geq 1$, and $T(z) = 1 - \min(1, |z|)$. For simplicity we may consider an equispaced triangular array of centers $\mu_j = a + (b-a)(j-1)/(m-1)$, and $\sigma = (b-a)/(m-1)$.[14] Once $\hat{m}$ is chosen, e.g., by cross-validation, only $\boldsymbol{\delta}$ needs to be computed, and the abstract optimization problem (6) reduces to

$$\min_{\boldsymbol{\delta} \in B_{\hat{m}}} \sum_{t=1}^{n} \left[y_t - g_{\hat{m}}\left(x_t, \boldsymbol{\delta}\right)\right]^2 \tag{8}$$

with $B_{\hat{m}} = \left\{ \boldsymbol{\delta} | \, |\delta_j| \leq B_1, j = 1, 2, ..., \hat{m}, \text{and} \, |\delta_j - \delta_{j-1}| \leq B_2 \left|\mu_j - \mu_{j-1}\right|, j = 2, ..., \hat{m} \right\}$, which is just a constrained linear least squares problem that can be easily solved by quadratic programming techniques. Although the above example is rather simplistic, the sieve is intuitive and the estimators are easy to compute.

Example 2 (Simultaneous estimation of a function and its derivatives, SISO case). The above scheme may be extended to least squares learning of the derivatives of θ^* up to the r-th order. In the SISO case with $X = [a,b]$, sufficient conditions are as follows:

$$\begin{aligned} \Theta \quad = \quad & \Big\{ \theta : [a,b] \to \mathbb{R} \mid \max_{0 \leq |\alpha| \leq r} \sup_{x \in (a,b)} |D^\alpha \theta(x)| \leq B_1, \\ & |D^r \theta(x_1) - D^r \theta(x_2)| \leq B_2 \left|x_1 - x_2\right|, x_1, x_2 \in [a,b] \Big\}, \end{aligned} \tag{9}$$

for known (finite) B_1 and B_2. The above set is compact in the space of r times continuously differentiable functions on $[a,b]$ endowed with norm $\|\theta\| = \max\limits_{0 \leq |\alpha| \leq r} \sup\limits_{x \in (a,b)} |D^\alpha \theta(x)|$ (see Adams, 1975, Chapter 1). By applying least squares learning to a FS-based sieve (a suitable UA property is required, see Landajo *et al.*, 2001), with the set of constraints (9) imposed on Θ_m, a consistent estimator for θ^* is obtained.

Remark. The bounds B_1 and B_2 restrict the estimation process to a relatively "small"

[14] The vector $(\mu_1, \ldots, \mu_m)$ and the spread σ were assumed fixed *a priori* in this sieve. The conditions of Theorem 1 permit them to be random. This includes two-step procedures, with a preliminary stage which finds suitable values for the antecedents of the rules, and a second step which gives suitable values for the consequents. Conditions required for the first step are mild: it is only required that the antecedent fuzzy sets provide an increasingly fine fuzzy covering of the pattern space $\mathbb{X}$, with probability approaching 1 as $n \to \infty$. These conditions hold, for instance, for the simple choice $\mu_j = x_j$, i.e., when the values of X for the first m observations in the sample are taken as centers for the antecedents of the fuzzy rules. Of course, a suitable clustering mechanism may generally provide a better choice of centers and spreads which avoids redundant rules.
A fully nonlinear-least-squares procedure to obtain (given $\hat{m}$) all parameters simultaneously is also permitted (see Landajo, 2004).

parameter space. Of course, if sufficiently rich knowledge is available, these bounds may be very tight, but in the absence of precise knowledge very loose bounds are enough to ensure consistency (although in this case very large sample sizes may be required). The expedient of replacing unconstrained learning by constrained estimation ensures that $\hat{\theta}_{\hat{m}}$ has a number of desirable regularity properties. First, it avoids ill-posed solutions in certain inverse problems (see Newey and Powell, 2003). Secondly, it may greatly reduce the risk of obtaining meaningless fuzzy rules. As observed by Guillaume (2001), the effects of applying unconstrained learning schemes (e.g., standard backpropagation) to problems with natural constraints can be potentially serious, greatly affecting the semantic aspects of fuzzy systems constructed on the basis of statistical/neural devices.

4.2 Least absolute deviation learning. Nonparametric quantile regressions

Linear quantile regression (LQR) is now a standard technique which has been successfully applied to a number of research tasks in the field of economics, and more specifically, financial economics (see Fitzenberger *et al.*, 2001, and references therein). LQR provides a generalization to classical least absolute deviation (LAD) linear regression. In the SISO case, given a finite sample $\{(x_t, y_t) \,|\, t = 1, \ldots, n\}$ drawn from the studied population, where it is assumed that the q-th conditional quantile of Y given $X = x$ has the form $Q^q(x) = \beta_0^q + \beta_1^q x$, an estimate for the free parameters of the quantile model is obtained as a solution to the asymmetric least absolute deviation (ALAD) problem:

$$\min_{(\beta_0^q, \beta_1^q)} \sum_{t=1}^{n} \ell_q \left[y_t - \beta_0^q - \beta_1^q x_t \right] \tag{10}$$

where $\ell_q[u] = q \max\{u, 0\} + (1 - q) \max\{-u, 0\}$. The above problem may be solved by linear programming techniques (e.g., Koenker, 2000). LQR offers some well-known advantages over least-squares-based linear modelling, its robustness to skewed tails and departures from normality being the most relevant (e.g., Mata and Machado, 1996). In addition, under very general conditions, the asymptotic distribution of the vector of estimated coefficients is multivariate normal, which permits standard inferences to be carried out.

A number of nonparametric quantile regression tools have also been proposed in statistics and econometrics literature. Among them, contributions by Yu and Jones (1998) —who used kernel estimators—, White (1992) —proposing multilayer-feed-forward neural networks—, and Koenker *et al.* (1994), who devised quantile series estimators based on regularized B-splines, can be highlighted. FSs may also be used to consistently estimate the q-th order conditional quantile of Y given X. Minor modifications of the framework we considered for least-squares regressions permit this to be proven.

4.2.1 General background

The basic framework generic quantile regression is as follows. Assuming that the relationship between the variable to forecast (Y) and a set of regressors (X) is studied, for any $q \in (0,1)$ the q-th quantile of Y conditional on X is a mapping Q^q which assigns to each $x \in X$ a number $y = Q^q(x)$ such that the conditional probability $P(Y \leq y|X = x) = q$. Conditional quantiles provide an alternative (but *complete*) description of the probability distribution of Y conditional on X. A typical object of interest is the conditional median, which coincides with the 50% quantile. Other conditional quantiles, especially those related to very high or very low q values may be useful in order to highlight certain patterns of the studied relationship more related to the behavior of the "best" or the "worst" behaved individuals (in terms of Y value) in the studied populations (e.g., Hendricks and Koenker, 1992). Hence, when the analysis is extended in order to include other quantiles, a considerably richer picture emerges. This is in contrast with classical regression methods, in the sense that they focus on estimating a single regression line which summarizes some *central* aspects (e.g., conditional expectations/medians) of the studied relationships.

Conditional quantiles have a number of useful features (e.g., Koenker and Basset, 1978). First, they naturally provide prediction intervals, without relying on any strong assumption (e.g., Gaussianity). In particular, as $P(Q^{\alpha/2}(x) \leq Y \leq Q^{1-\alpha/2}(x)|X = x) = 1-\alpha$, the interval $\left[Q^{\alpha/2}(x, Q^{1-\alpha/2}(x)\right]$ provides a $(1-\alpha) \times 100$ prediction interval for Y given $X = x$. Conditional quantiles inherit some nice properties of marginal quantiles. In particular, for $Z = \Phi(Y)$ with $\Phi(\cdot)$ being any continuous increasing mapping, the $q-$th quantile of Z may be simply obtained as $Q^q_Z = \Phi\left(Q^q_Y\right)$. An analogous property holds for conditional quantiles. This implies that monotonic transformations of the response variable can be carried out without affecting the basic structure of the problem. (As well known, for least squares regression such a property only is available for affine transformations.)

4.2.2 FS-based learning of conditional quantiles

We now assume that the data set to be processed, $D = \{(x_t, y_t) \,|\, t = 1, \ldots, n\}$, is a realization of an i.i.d. sequence, with the behavior of Y conditional on vector X being summarized by the following regression-type structure:

$$Y_t = \mu^*(X_t) + \sigma^*(X_t)\varepsilon_t \tag{11}$$

where $\mu^*(x)$ is the conditional median of Y_t given $X_t = x$, i.e., $\mu^*(x) = Me(Y_t|X_t = x)$, ε_t is an error component with null median, and $\sigma^*(\cdot)$ is a conditional spread function (variances are not required to be finite). The above structure directly gives the form of conditional quantile of Y_t given $X_t = x$:

$$\theta^*_q(x) = \mu^*(x) + \sigma^*(x)Q^q_\varepsilon \tag{12}$$

where Q^q_ε is the q-th quantile of ε_t. This gives an alternative expression for (11):

$$Y_t = \theta^*_q(X_t) + \sigma^*(X_t)U^q_t \tag{13}$$

with $U^q_t = \varepsilon_t - Q^q_\varepsilon$ being an error term with q-th quantile $Q^q_U = 0$. We impose the following requirements:

Assumption 1'. Conditions (**i**) and (**iii**) in Assumption 1 hold. (**ii**) the error ε_t has $Me(\varepsilon_t) = 0$ and $E\,|\varepsilon_t| < \infty$.

Assumption 4'. (**i**) θ^* is the unique minimizer on Θ of the function $Q : \Theta \to \mathbb{R}_+$ defined as follows: $Q(\theta) = \int\limits_{\mathbb{X}}\int\limits_{\mathbb{R}} \ell_q \left[\sigma^*(x)(\varepsilon - Q^q_\varepsilon) + \theta^*_q(x) - \theta(x)\right] P_X(dx)P_\varepsilon(d\varepsilon)$. □

Theorem 2. Under Assumptions 1', 2, 3 and 4', let $\hat{\theta}^q_{\hat{m}}$ be a solution to the problem

$$\min_{\theta \in \Theta_{\hat{m}}} Q_n(\theta) \equiv n^{-1} \sum_{t=1}^{n} \ell_q \left[Y_t - \theta(X_t)\right]. \tag{14}$$

Then $\left\| \hat{\theta}^q_{\hat{m}} - \theta^*_q \right\| \xrightarrow{P} 0$ as $n \to \infty$. □

4.3 B-spline-based FSs for conditional quantile learning

The above results ensure consistent learning for any class of FSs possessing a suitable UA property, without *a priori* any preference for a particular specification of FS structure. In this subsection we will briefly review a number of practical issues related to implementation of the above framework for quantile regression. Although we shall focus on additive FSs with (*fixed-knot*) B-spline membership functions, which provide a very simple implementation in the SISO cases we study in Section 5, most comments are valid for more general contexts (e.g., other choices of basis functions —such as Gaussians—are permitted, and the extension to MISO systems is straightforward, with the habitual complications associated with the increase of dimensionality). The structure of the spline FS approximators is as follows:

$$g_m(x, \boldsymbol{\delta}) \equiv \sum_{j=1}^{m} \delta_j T_j(x) \tag{15}$$

where $x \in [a, b]$ (a bounded interval), and $T_j(\cdot)$ is the j-th element of a basis of normalized r degree B-splines. For the above basis the local partition of unity property holds, i.e., $\sum_{j=1}^{m} T_j(x) = 1$ for any $x \in [a, b]$. The set $\boldsymbol{C} = \{c_{-r} < \ldots < c_0 = a < c_1 < \ldots < c_{K-1} < b = c_K < \ldots < c_{K+r}\}$ collects the knot sequence. The $K - 1$ knots which are interior to $[a, b]$ are known as basic knots and the other points are named auxiliary knots. We chose the equispaced sequence $c_k = a + k(b - a)/(m - r)$, with $k = -r, \ldots, m$. The dimension of the spline space (15) is m, with $m \in \{r + 1, r + 2, \ldots\}$.[15] Vector $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_m)$ collects all the free parameters (once m and $\boldsymbol{C}$ are determined). Standard UA properties of splines and Theorem 2 above ensure that we can use this class of FSs to estimate the q-th conditional quantile of Y given X.[16]

4.3.1 Computational issues

Fixed-knot additive structures provide the remarkable advantage of reducing the generic ALAD problem (14) to a linear programming problem which can be solved efficiently by

[15] With this definition the case $m = r + 1$ corresponds to polynomials of degree r.

[16] Consistency and other asymptotic properties of nonparametric spline-based quantile regressions have been established in the statistics literature by He and Shi (1994) and Portnoy (1997), both in the smoothing splines version (as originally proposed by Koenker *et al.*, 1994) and in the naïve —i.e., fixed-knots (linear) splines, without roughness penalty— version (Portnoy, 1997).

specialized routines.[17] A further advantage is that (when available) bounds such as those in (8) and (9) can be easily incorporated, as they just add further linear constraints to the basic (linear) ALAD problem. In this paper we fitted all quantile models by using a so-called "Frisch-Newton" algorithm, namely, a simplified version of the interior point algorithm proposed by Portnoy and Koenker (1997); the algorithm essentially solves the dual of the linear program associated with LQR.[18]

4.3.2 Model selection

As to model complexity selection, cross-validation (Stone, 1974) is a leading choice. An enormous number of variants are available in the literature. The classical leave-one-out version is computationally demanding, and a number of authors have pointed out a considerable sampling variability in cross-validated estimates. Some authors (e.g., Ripley, 1996, Chapter 2) have suggested that a compromise reducing computation and producing more stable estimates —even at the cost of moderate bias— is suitable in many practical applications. In this paper we used 10-fold cross-validation, although other possibilities (in particular, the leave-one-out method) produced very close results. For a generic q-th order quantile and a B-spline fuzzy model with m rules the procedure is as follows. First the data set is randomly split into 10 sheets or data subsets. Then, the 10-fold cross-validated mean error is computed:

$$E_q^{CV}(m) \equiv \frac{1}{10}\sum_{k=1}^{10} E_{q,k}^{*} \tag{16}$$

where

$$E_{q,k}^{*} \equiv \frac{1}{n_k}\sum_{t\in \boldsymbol{S}_k} \ell_q\Big[y_t - \sum_{j=1}^{m} \hat{\delta}_{j(k)} N_j(x_t)\Big] \tag{17}$$

i.e., $E_{q,k}^{*}$ is a weighted mean of the prediction errors on the sheet $\boldsymbol{S}_k$ (whose cardinality is $n_k = \left[\frac{9}{10}n\right]$) of the spline model with parameter set $\hat{\boldsymbol{\delta}}_{(k)} = (\hat{\delta}_{1(k)}, ..., \hat{\delta}_{m(k)})$, which was fitted

[17] The standard linear ALAD problem for the q-th quantile has the following (primal) LP formulation:

$$\min_{(\boldsymbol{\beta},\boldsymbol{u},\boldsymbol{v})} \sum_{t=1}^{n} q u_t + \sum_{t=1}^{n} (1-q) v_t$$

s.t.: $x_t\boldsymbol{\beta}' + u_t - v_t = y_t$; $u_t, v_t \geq 0$; $\boldsymbol{\beta} \in \mathbb{R}^d$; $t = 1, \ldots, n$; $\boldsymbol{u} = (u_1, \ldots, u_n)$ and $\boldsymbol{v} = (v_1, \ldots, v_n)$.

[18] The version used in this paper was written by R. Koenker and is available at http://www.econ.uiuc.edu/∼roger/research/rq/rq.html. Interior point methods (especially when combined with preprocessing) have enabled dramatic efficiency gains over classical simplex algorithms in ALAD fitting (see Koenker, 2000). Regularization or roughness penalty methods have also been proposed for spline-based quantile regression, by Koenker *et al.* (1994).

on the basis of the estimation set with sheet $\boldsymbol{S}_k$ excluded. The above procedure is applied to all the permitted values for m, and the lowest model complexity m^* which minimizes $E_q^{CV}(m)$ over the permitted range is selected. The same expedient is successively applied for each of the studied quantiles.

4.3.3 Linearity testing

Since nonparametric methods are more technically demanding than standard linear models, it is important to have available adequate statistical devices to assess to what extent a nonlinear/nonparametric approach is worth trying in a given practical problem. Linearity tests are a basic tool in this respect.[19] B-spline-based FSs are particularly well adapted to this testing framework. The linearity test we propose exploits the so-called *piecewise polynomial form* of spline models. The idea is straightforward. As well known, the structure (15) has the equivalent form:

$$Q^q(x, \boldsymbol{\beta}) \equiv \beta_0^q + \beta_1^q x + \ldots + \beta_r^q x^r + \sum_{j=1}^{K-1} \beta_{r+j}^q (x - c_j)_+^r \tag{18}$$

where now $\boldsymbol{\beta} = (\beta_0^q, \ldots, \beta_{r+K-1}^q)$. Since the linear model is nested in the above functional form, model (18) is *correctly specified* in the case when the quantile is linear, and the vector $\hat{\boldsymbol{\beta}}$ of unrestricted ALAD estimators converges to $\boldsymbol{\beta} = (\beta_0^q, \beta_1^q, 0, \ldots, 0)$ under linearity, with asymptotic normality ensured under standard conditions. Therefore, we can test for linearity by testing the following parametric restriction on the spline model:

$$\begin{aligned} &H_0 : \beta_2^q = \ldots = \beta_{r+K-1}^q = 0, \text{ against} \\ &H_A : \beta_j^q \neq 0 \;\text{ for some }\; j \geq 2. \end{aligned} \tag{19}$$

The above restriction may be straightforwardly tested by following the minimum distance (MD) approach proposed by Buchinsky (1998), since (19) amounts to a set of linear constraints (represented by an appropriate restriction matrix R) on the parameters of the

[19] The proposed test straightforwardly extends to the least-squares nonparametric regressions considered in the above subsections. MISO systems are also permitted.

quantile equation. The following test statistic may be constructed:

$$d = n\left(\hat{\boldsymbol{\beta}} - R\hat{\boldsymbol{\beta}}^{r}\right)' A^{-1}\left(\hat{\boldsymbol{\beta}} - R\hat{\boldsymbol{\beta}}^{r}\right) \tag{20}$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^{r}$ are respectively the unrestricted (ALAD) and restricted (MD) estimator for $\boldsymbol{\beta}$,[20] the restriction matrix is $R = [r_{i,j}]$, $i = 1, \ldots, r + K$, $j = 1, 2$, with $r_{1,1} = r_{2,2} = 1$ and $r_{i,j} = 0$ otherwise, and $A = \widehat{\Lambda}_{\hat{\boldsymbol{\beta}}}$ is an estimate of $\Lambda_{\hat{\boldsymbol{\beta}}}$ (the latter being the asymptotic covariance of $\hat{\boldsymbol{\beta}}$). Under standard conditions the limiting distribution of $\sqrt{n}\left(\hat{\boldsymbol{\beta}}^{r} - \boldsymbol{\beta}\right)$ is normal, and the asymptotic null distribution of (20) is Chi-squared with $r + K - 2 = m - 2$ degrees of freedom (Buchinsky, *ibidem*).

The above test has asymptotically correct size for any choice of the spline model complexity with $r \geq 2$ and $K \geq 1$ (we adopted the convention of identifying $K = 1$ with a polynomial of degree r). However, the power of the test (i.e., its capability to detect nonlinear patterns) increases with the complexity of (18), since more complex spline models permit us to approximate more complicated nonlinear patterns. A usual recommendation in related literature is undersmoothing, i.e., permitting a model complexity (basically, a K value) somewhat higher than selected by cross-validation or penalization methods (e.g., Hong and White, 1995, in application of spline models in nonparametric testing within the framework of least squares regression; also Racine, 1997). Other protection usually recommended against bias is nesting the null (i.e., linear) model into the nonparametric estimator. This is ensured in our setting by the structure of splines.[21] In the next section we shall apply the above testing strategy in order to assess linearity in a number of practical problems.[22]

[20] The MD estimator is a minimizer of $\tilde{Q}(\boldsymbol{\beta}^{r}) = (\hat{\boldsymbol{\beta}} - R\boldsymbol{\beta}^{r})A^{-1}(\hat{\boldsymbol{\beta}} - R\boldsymbol{\beta}^{r})'$, with R being a restriction matrix and A being a positive definite matrix.

[21] For other classes of FSs not nesting linear models a slight modification of the above testing framework is required. For instance, for additive FSs with Gaussian membership functions (any other class which *does not include* linear functions is permitted), we may consider the following augmented structure:

$$f(x, \beta, \boldsymbol{\delta}) = \beta x + g_m(x, \boldsymbol{\delta}) = \beta x + \frac{\sum\limits_{j=1}^{m} \delta_j T \quad \frac{x-\mu_j}{\sigma}}{\sum\limits_{j=1}^{m} T \quad \frac{x-\mu_j}{\sigma}}$$

with $T(\cdot)$ denoting Gaussian membership function (the array of centers μ_j and centers σ_j is assumed *a priori* fixed). Linearity amounts to a parametric restriction of the form $\delta_1 = \ldots = \delta_m$. With a suitable modification of the restriction matrix R we proceed as in (20).

[22] Other tests, based on comparing conditional predictive ability in out-of-sample forecasting, can also be applied to assess linearity of one or a set of conditional quantiles. We omit details for brevity (see Landajo *et al.*, 2007b).

5 Some Applications

Two applications to SISO modelling are included in this section. We applied cubic-spline-based FSs (i.e., $r = 3$ was set). With this choice, the case when the number of fuzzy rules is $m = 4$ corresponds to a cubic polynomial, and cubic splines properly correspond to $m \geq 5$. For $m = 3$ we reduced the degree to $r = 2$, and a quadratic polynomial is obtained.

5.1 Example 1. A generic quantile learning problem

We considered the following learning problem. Observations were generated by the following system:

$$Y_t = 4 + 3\sin(\pi X_t) + (1 + 2X_t^4)\varepsilon_t$$

with $\{X_t\}$ and $\{\varepsilon_t\}$ being mutually independent i.i.d. sequences, X_t following a uniform distribution with suport in $[0, 1]$ and ε_t following a standard normal density. A sample of size $n = 1,000$ was drawn from the above population. 70% of the data set was used for model fitting and the rest of the sample was set apart for predictive evaluation purposes. We tried to estimate quantiles $q = 0.1, 0.25, 0.5, 0.75 \,\mathrm{and}\, 0.9$ by using the above spline-based FS structure (expert information was not incorporated). We permitted the number of fuzzy rules to range between 2 and 10. Ten-fold cross-validation was used for complexity selection. Figure 4 below shows the data and the population's conditional quantiles, and the FS-fitted quantiles are displayed in Figure 5. The shapes of the estimated quantiles closely resemble those of their population's counterparts. Table 1 shows, for each quantile, the main diagnostics of model fitting and cross-validation, as well as the results of MD linearity tests, for a range of m values (estimates for the asymptotic covariance matrix $\Lambda_{\hat{\beta}}$ were computed by using design matrix bootstrap with $B = 500$ resamples). As expected, very strong evidence against linearity is provided by the tests for all the studied quantiles. Table 1 also reports the mean ALAD prediction errors on the prediction set. Finally, estimates for the L_1 and supnorm distances between estimated and population quantiles appear in the same Table.[23] These distances were approximated by using a grid of $v = 1,000$ points

[23] We were not specifically concerned with learning of derivatives. This goal usually requires higher sample sizes and/or more accurate expert information.

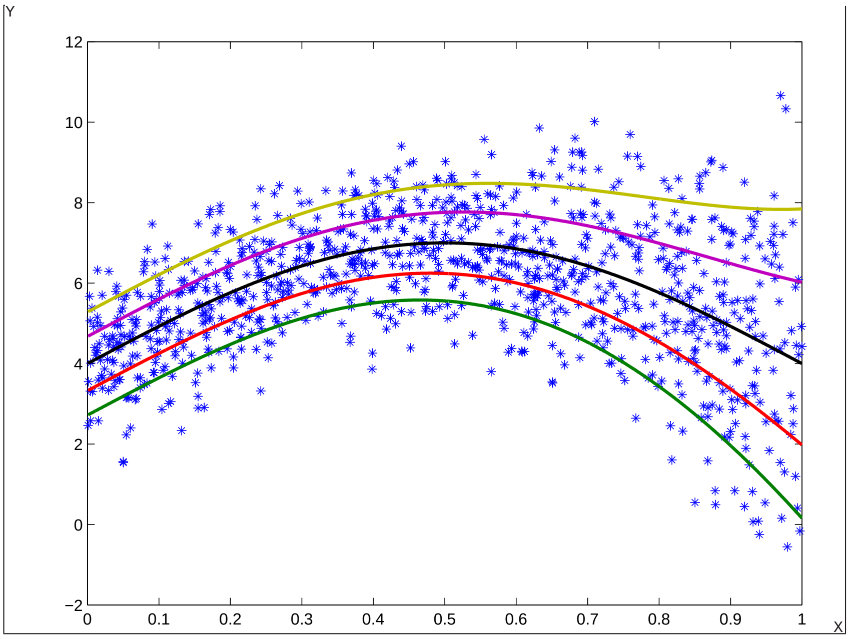

Figure 4: Data vs. population conditional quantiles.

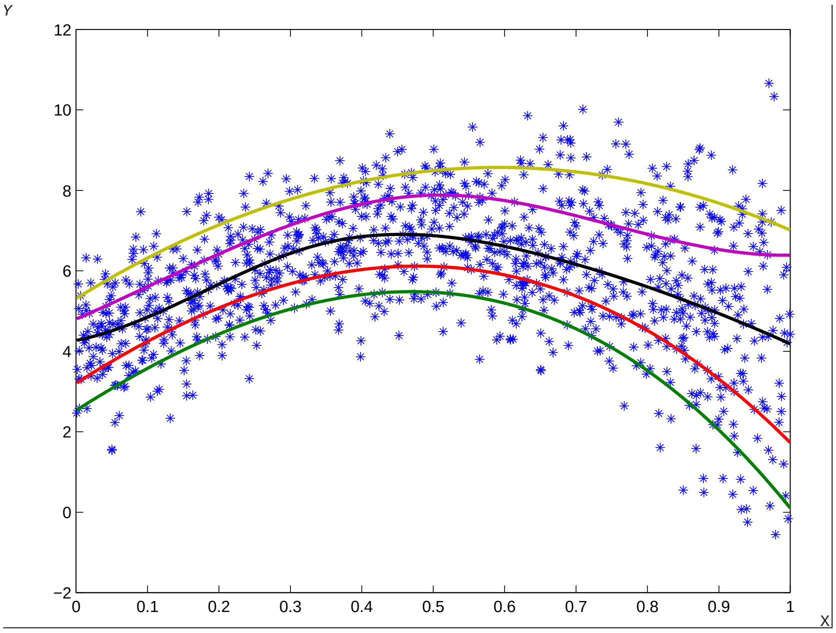

Figure 5: Data vs. spline-based FSs for conditional quantiles.

Table 1: Results of B-spline-based FS models for conditional quantiles in example 1 (“error” means “ℓ_q error)”.

QUANTILE	10%	25%	50%	75%	90%
Model Fitting					
Complexity (m)	4	4	6	5	3
Mean B-spline error	0.242	0.442	0.558	0.446	0.239
Cross-validation					
Mean error	0.245	0.446	0.562	0.449	0.241
Linearity Tests					
$m = 4$ (Cubic polynomial)					
Chi-squared stat.	163.24	248.24	189.40	154.443	111.449
p-value	0.000	0.000	0.000	0.000	0.000
$m = 8$					
Chi-squared stat.	204.05	308.167	207.085	204.592	122.149
p-value	0.000	0.000	0.000	0.000	0.000
$m = 10$					
Chi-squared stat.	184.842	246.538	188.485	151.918	104.616
p-value	0.000	0.000	0.000	0.000	0.000
Prediction					
Mean error	0.249	0.451	0.572	0.464	0.253
Approximation accuracy					
$d_1\left(\hat{\theta}^q_{\hat{m}}, \theta^*_q\right)$	0.070	0.071	0.118	0.069	0.121
$d_\infty\left(\hat{\theta}^q_{\hat{m}}, \theta^*_q\right)$	0.187	0.248	0.274	0.365	0.831

of the form $\tilde{x}_i = 0 + (i-1)/(v-1)$, $i = 1, \ldots, v$, i.e., we calculated:

$$d_1\left(\hat{\theta}^q_{\hat{m}}, \theta^*_q\right) = v^{-1} \sum_{i=1}^{v} \left|\hat{\theta}^q_{\hat{m}}(\tilde{x}_i) - \theta^*_q(\tilde{x}_i)\right|$$

and

$$d_\infty\left(\hat{\theta}^q_{\hat{m}}, \theta^*_q\right) = \max_{i=1,\ldots,v} \left|\hat{\theta}^q_{\hat{m}}(\tilde{x}_i) - \theta^*_q(\tilde{x}_i)\right|.$$

5.2 Example 2. FS-based profitability forecasting

Company performance is usually measured by profitability, which may itself be proxied by using the return on assets (ROA) ratio, defined as the quotient of net profit after taxes to total assets. This ratio is very popular among both academics and financial analysts. Typically, the value of the ROA ratio for a given firm is assessed by comparison with the so-called ‘industry norm’, this being a suitable synthesis (e.g., mean or median) of the values of the ratio in the relevant population (namely, the industry to which the company under scrutiny belongs). The suitability of the above comparison requires that the bivariate relationship between the components of the ratio satisfies a number of statistical properties. In particular, the method requires this relationship to be *linear* and *strictly proportional* (in statistical jargon, this amounts to assuming that the relationship between returns and assets is well summarized by a linear regression model with null intercept). However, a number of authors (e.g., Lev and Sunder, 1979; Whittington, 1980) have pointed out that the assumptions of linearity and strict proportionality rarely hold, which would imply that the use of the ratio form to summarize the relationship between two accounting variables can be generally inadequate.

In a recent work (Landajo *et al.*, 2007a) we relaxed the linearity assumption, and considered ANN-based nonparametric LAD estimation. In Landajo *et al.* (2007b) we addressed the problem from a more general standpoint, in order to consider potential nonlinear features associated with extreme conditional quantiles. Nonparametric regressions provide a flexible implementation of the quantile framework, which permits a differentiated treatment for each quantile. The latter is a particularly useful feature in the case of profitability analysis, because of the potential presence of nonlinear features (e.g., scale economies) which may affect extreme quantiles much more strongly than more 'central' regression lines. In this subsection we apply the FS-based quantile regression approach outlined in Subsection 4.3 above. We used as a benchmark a representative data base from an homogeneous sector, namely, the Spanish book-publishing firms (NACE 2211). Once a number of filters were applied we obtained a final data set made up of 520 firms (details on the data base and underlying economic theory appear in Landajo *et al.*, 2007b).

As the available sample size was relatively small, the number of rules was limited to range between 3 and 6 (this limitation had basically no practical effects, as detailed below). A visual inspection of the estimated conditional quantiles (see Figure 6) reveals the main features of the nonparametric analysis. Basically, the estimated quantiles are remarkably smooth-looking curves, with moderate degrees of complexity. The fitted 'central' quantiles (25%,50%,75%) are clearly linear, although the situation for the extreme quantiles appears to be somewhat different, suggesting moderate departures from linearity.

More elaborate conclusions may be drawn from a careful analysis of the results in Table 2. Results of LQR were also displayed for comparison. First, model fitting results —with 10-

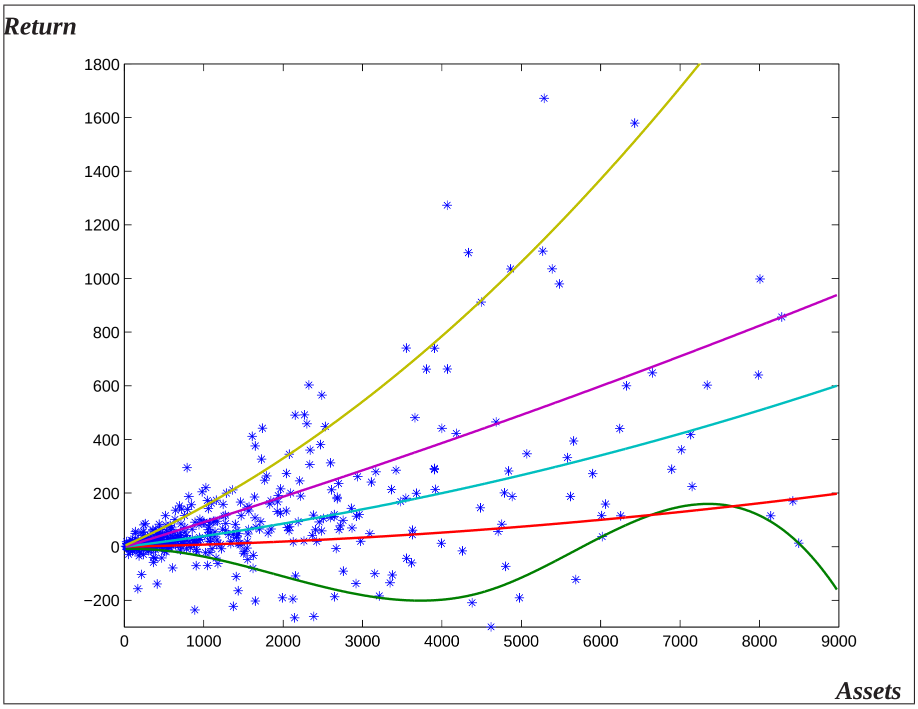

Figure 6: Fitted FS-based estimates for the conditional quantiles in example 2.

fold cross-validated model complexities and mean errors— are displayed. The most evident detail is that cross-validation tended to select remarkably simple models. The complexity $m = 3$, i.e., a quadratic polynomial (hence including linear models as a restriction), is the most frequent choice. For the 90% quantile the chosen complexity is also $m = 3$, although the form displayed in Figure 5 is clearly quadratic, which would be congruent with the shape expected in the presence of economies of scale. A slightly more complicated nonlinear pattern appears at the 10% quantile. The selected complexity is $m = 5$, which corresponds to a cubic spline with 1 basic knot. Figure 5 also shows a slight crossing of the 10% and 25% quantiles (this being clearly a consequence of the few observations on that area).

The shapes of the 10% and 90% quantiles appear to correspond to those predicted by economic theory. As to the 10% quantile, the U-shaped curve suggests the presence of the so-called “small firm effect”. Regarding the 90% quantile, the estimated curve suggests that returns tend to grow at increasing rates as assets increase, indicating that only the best performing firms are capable of taking advantage of the benefits of scale.

Table 2 also reports the results of MD linearity testing, for several choices of model complexity (as commented above, a certain degree of overfitting may be desirable to enhance the power of the test against a sufficiently wide range of nonlinear alternatives). We permitted complexities ranging from the quadratic and cubic polynomials up to cubic splines with $K = 6$ (or equivalently, FSs with $3 \leq m \leq 10$ rules). The results roughly support the conclusions from visual analyses. As to the central quantiles, for all the considered nonlinear alternatives the p-values are rather large, indicating that no significant evidence against linearity may be deduced from data. On the contrary, the linearity tests for the 10% quantile strongly suggest the presence of nonlinear patterns, with low p-values under all the alternative nonlinear specifications, going under 0.005 when the permitted model complexity becomes sufficiently large. However, for the 90% quantile, the tests suggest no significant departure from linearity, with large p-values for the most complicated spline models, although when we restrict the search to quadratic and cubic polynomials the linearity hypothesis seems to be in doubt, with moderately low p-values which would be congruent with a low degree polynomial pattern.

Table 2: Linear vs. FS-based quantile models in example 2. Model fitting and cross-validation (mean ℓ_q error), number of fuzzy rules (m), and results of linearity tests.

QUANTILE	10%	25%	50%	75%	90%
Model Fitting					
Complexity	5	3	3	3	3
Mean linear error	26.74	39.72	49.19	44.05	26.91
Mean B-spline error	25.46	39.57	48.93	43.94	25.82
10-fold cross-validation					
Mean linear error	26.85	39.80	49.67	44.15	27.25
Mean B-spline error	26.00	39.76	49.82	44.13	26.37
Linearity Tests					
$m = 3$					
Chi-squared stat.	3.894	0.567	1.021	0.045	2.611
p-value	0.049	0.451	0.312	0.833	0.106
$m = 4$					
Chi-squared stat.	5.393	1.286	2.314	0.071	5.071
p-value	0.067	0.526	0.314	0.965	0.079
$m = 5$					
Chi-squared stat.	11.327	2.961	1.940	2.182	4.902
p-value	0.010	0.398	0.585	0.536	0.179
$m = 6$					
Chi-squared stat.	11.439	2.170	2.953	2.117	5.015
p-value	0.022	0.705	0.566	0.714	0.286
$m = 8$					
Chi-squared stat.	24.056	4.810	1.705	3.924	5.211
p-value	0.001	0.568	0.945	0.687	0.517
$m = 10$					
Chi-squared stat.	21.915	3.631	1.322	4.350	4.008
p-value	0.005	0.889	0.995	0.824	0.856

6 Concluding remarks and further research

The results in this paper show that any class of FSs having suitable universal approximation properties, constructed on the basis of statistical data and expert information provided by the user, possesses model-free learning capabilities analogous to those of the most sophisticated nonparametric statistical techniques. We studied two classical learning paradigms, namely least squares and least absolute deviation. When sufficiently large samples are available, FSs can approximate with high accuracy the mappings of interest as well as their derivatives to any finite order. This provides further support to the application of fuzzy models in random contexts such as those arising in many fields. Knowledge of the statistical properties of FSs may help improve the design and performance of such models (e.g., conventional inference tools can be applied) without having to renounce interpretability (constrained learning may help in this respect).

The topics analyzed in the text provide only a small sample (we focused on consistency) of the statistical properties of FSs that practitioners may exploit in order to enhance the performance of their fuzzy models. It can be expected that FS-based sieves possess other useful properties similar to those of standard nonparametric techniques (e.g., asymptotic normality of the limiting distributions of smooth functionals; see Pagan and Ullah, 1999). The above results remain valid when cross-sectional data are replaced by time series data generated by stationary ergodic processes, and can also be extended to extraction of deterministic components (e.g., trends) in some classes of non-stationary systems. Continuous functionals of the regression surfaces (e.g., average derivatives) can also be consistently estimated, by using the plug-in method.

Appendix. Mathematical proofs.

The results in this paper follow as particularizations of the following general purpose results of Newey and Powell (2003).

Lemma A.1. Suppose **i)** $Q(\theta)$ has a unique minimum on Θ at θ^*; **ii)** $\hat{Q}(\theta)$ and $Q(\theta)$ are continuous, Θ is compact and $\max_{\theta\in\Theta}\left|\hat{Q}(\theta) - Q(\theta)\right| \xrightarrow{P} 0$; iii) $\hat{\Theta}$ are compact subsets of Θ such that for any $\theta \in \Theta$ there exists $\tilde{\theta} \in \hat{\Theta}$ such that $\tilde{\theta} \xrightarrow{P} \theta$. Then $\hat{\theta} = \arg min_{\theta\in\hat{\Theta}}\hat{Q}(\theta) \xrightarrow{P} \theta^*$. □

Proof. See Lemma A1 in (Newey and Powell, 2003). □

Lemma A.2. If **i)** Θ is a compact subset of a space with norm $\|\cdot\|$; **ii)** $\hat{Q}(\theta) \xrightarrow{P} Q(\theta)$ for all $\theta \in \Theta$; **iii)** there is $\delta > 0$ and $B_n = O_P(1)$ such that for all $\theta, \tilde{\theta} \in \Theta$, $\left|\hat{Q}(\theta) - \hat{Q}(\tilde{\theta})\right| \leq B_n \|\theta - \tilde{\theta}\|^\delta$, then $Q(\theta)$ is continuous and $\sup_{\theta\in\Theta} \left|\hat{Q}(\theta) - Q(\theta)\right| \xrightarrow{P} 0$.
Proof. See Lemma A2 in (Newey and Powell, 2003). □

Proof of Theorem 1.

We only have to check that conditions of Lemma A.1 above hold. By assumption Θ is compact and $Q(\cdot)$ has a unique minimum on Θ at θ^*. The role of $\hat{Q}(\cdot)$ is played by our $Q_n(\cdot)$. For fixed $D = \{(x_t, y_t), t = 1, \dots, n\}$, $Q_n(\cdot)$ is easily shown to be continuous with respect to θ, as for any $\theta, \theta' \in \Theta$, it holds

$$|Q_n(\theta') - Q_n(\theta)| \leq n^{-1} \sum_{t=1}^{n} 2|y_t||\theta'(x_t) - \theta(x_t)| + n^{-1} \sum_{t=1}^{n} |\theta(x_t) + \theta'(x_t)|\,|\theta'(x_t) - \theta(x_t)|$$
$$\leq \left(n^{-1} \sum_{t=1}^{n} 2|y_t| + 2\|\theta\| + \|\theta' - \theta\| \right) \|\theta' - \theta\|$$

which converges to zero as $\|\theta' - \theta\| \to 0$. As to the weak uniform law of large numbers (WULLN), Lemma A.2 applies. This requires a WLLN and a stochastic Lipschitz condition to hold for $Q_n(\cdot)$. As to the WLLN, it is straightforwardly derived from the standard decomposition

$$Q_n(\theta) = n^{-1} \sum_{t=1}^{n} \left[\sigma^*(X_t)\varepsilon_t + \theta^*(X_t) - \theta(X_t)\right]^2 = I + II + III$$

with $I = n^{-1} \sum\limits_{t=1}^{n} \sigma^*(X_t)^2 \varepsilon_t^2$, $II = n^{-1} \sum\limits_{t=1}^{n} 2\sigma^*(X_t)\varepsilon_t \left(\theta^*(X_t) - \theta(X_t)\right)$, and $III = n^{-1} \sum\limits_{t=1}^{n} [\theta^*(X_t) - \theta(X_t)])^2$. Since Assumptions 1 and 2 imply $E\left|\sigma^*(X_t)\varepsilon_t\right|^2 < \infty$, Kolmogorov's LLN for i.i.d. sequences (together with independence of X_t and ε_t) give that $I \xrightarrow{P} \Sigma \equiv \sigma_\varepsilon^2 \int\limits_{\mathbb{X}} \sigma^*(x)^2 dP_X(x) < \infty$ as $n \to \infty$. As to II, an analogous argument gives $II \xrightarrow{P} 0$. Similarly, boundedness of Θ and the LLN for i.i.d. sequences gives $III \xrightarrow{P} \int\limits_{X} \left[\theta^*(x) - \theta(x)\right]^2 dP_X(x)$. As to the Lipschitz condition, it is straightforwardly derived. For any $\theta, \theta' \in \Theta$

$$|Q_n(\theta') - Q_n(\theta)| \leq n^{-1} \sum_{t=1}^{n} 2|\varepsilon_t|\sigma^*(x_t)|\theta'(x_t) - \theta(x_t)|$$
$$+ n^{-1} \sum_{t=1}^{n} |2\theta^*(x_t) - \theta'(x_t) - \theta(x_t)|\,|\theta'(x_t) - \theta(x_t)| \leq (B_n + 4B)\,\|\theta' - \theta\|$$

where $B_n = n^{-1} \sum_{t=1}^{n} 2|\varepsilon_t|\sigma^*(x_t)$, which is bounded in probability as a consequence of the WLLN, and $B = \sup_{\theta \in \Theta} \|\theta\|$, which is finite by compactness of Θ. Hence, $Q_n(\cdot)$ satisfies a stochastic Lipschitz condition (with exponent 1). Thereof, Lemma A.2 ensures that $Q(\cdot)$ is continuous on Θ, and that a WULLN holds for $Q_n(\cdot)$ on the same set, as required by Lemma A.1 above. The remaining conditions in Lemma A.1 are ensured by identifying $\Theta_{\hat{m}} = \hat{\Theta}$, which is a (random) compact set. The UA property in Assumption 3 above, together with the requirement $\hat{m} \xrightarrow{P} \infty$, ensure the existence of $\tilde{\theta}_{\hat{m}}$ (e.g., we can take $\tilde{\theta}_{\hat{m}} = \arg min_{\theta \in \Theta_{\hat{m}}} \|\theta - \theta^*\|$) which satisfies $\tilde{\theta}_{\hat{m}} \xrightarrow{P} \theta^*$ as $n \to \infty$ for any choice of $\theta^* \in \Theta$. Hence, the requirements of Lemma A.1 are fulfilled and the conclusion of the Theorem straightforwardly follows. □

Remark. Uniqueness of θ^* imposed in Assumption 4 is redundant. The mapping $Q(\theta) = \Sigma + \int\limits_X [\theta^*(x) - \theta(x)]^2 \, dP_X(x)$ only can have a unique minimum over Θ, at θ^*, as any other minimizer θ^{**} should satisfy the equality $\int\limits_X (\theta^*(x) - \theta^{**}(x))^2 \, dP_X(x) = 0$, which implies $\theta^{**} \equiv \theta^*$ (this being a consequence of continuity of θ^{**} and θ^* on $\mathbb{X}$, plus the requirement in Assumption 1 that $P_X(\mathcal{O}) > 0$ for any nonempty open subset of $\mathbb{X}$.

Proof of Theorem 2.

We proceed as in Theorem 1. The inequality $|\ell_q(u) - \ell_q(u')| \leq |u - u'|$ gives

$$|Q_n(\theta') - Q_n(\theta)| \leq n^{-1} \sum_{t=1}^{n} |\ell_q[y_t - \theta(x_t)] - \ell_q[y_t - \theta'(x_t)]| \leq$$
$$n^{-1} \sum_{t=1}^{n} |\theta'(x_t) - \theta(x_t)| \leq \|\theta' - \theta\|$$

for any $\theta, \theta' \in \Theta$, which proves Lipschitz continuity of $Q_n(\cdot)$ with respect to θ. Since $E\left|\sigma^*(X_t)U_t^q + \theta_q^*(X_t) - \theta(X_t)\right| < \infty$, the WLLN required by Lemma A.2 follows from Kolmogorov's LLN for i.i.d. sequences. Hence, for any $\theta \in \Theta$, $Q_n(\theta) \xrightarrow{P} Q(\theta)$. The rest of the proof is identical to that of Theorem 1. □

References

R.A. Adams, *Sobolev Spaces.* New York: Academic Press, 1975.

M. Buchinsky, "Recent Advances in Quantile Regression Models. A Practical Guide for Empirical Research," *The Journal of Human Resources*, vol. 33, no. 1, pp. 88-126, 1998.

X. Chen and X. Shen, "Sieve Extremum Estimates for Weakly Dependent Data," *Econometrica*, vol. 66, no. 2, pp. 289-314, 1998.

R.A. DeVore and G.G. Lorentz, *Constructive Approximation.* Berlin: Springer-Verlag, 1993.

B. Fitzenberger, R. Koenker and J.A.F. Machado, "Introduction," *Empirical Economics*, vol. 26, no. 1, pp. 1-5, 2001.

A. R. Gallant and H. White, "On Learning of the Derivatives of an Unknown Mapping With Multilayer Feedforward Networks," *Neural Networks*, vol. 5, pp. 129-138, 1992.

S. Geman, E. Bienenestock, and R. Doursat, "Neural Networks and the Bias/Variance Dilemma," *Neural Computation*, 4 pp. 1-58, 1992.

S. Geman and C. R. Hwang, "Nonparametric Maximum Likehood Estimation by The Method of Sieves," *The Annals of Statistics*, vol. 10, no. 2, pp. 401-414, 1982.

U. Grenander, *Abstract Inference.* New York: John Wiley & Sons, 1981.

S. Guillaume, "Designing Fuzzy Inference Systems from Data: An Interpretability-Oriented Review," *IEEE Transaction on Fuzzy Systems*, vol. 9, no. 3, pp. 426-443, 2001.

X. He and P. Shi, P., "Convergence rate of B-spline estimators of nonparametric conditional quantile functions," *Journal of Nonparametric Statistics*, no. 3, pp. 299-308, 1994.

Y. Hong and H. White, H., "Consistent specification testing via nonparametric series regression," *Econometrica*, vol. 63, no. 5, pp. 1133-1159, 1995.

J. S. R. Jang, "ANFIS: Adaptive-Network-Based Fuzzy Inference System," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 23, no. 3, pp. 665-685, 1993.

R. Koenker, "Galton, Edgeworth, Frisch, and prospects for quantile regression in Econometrics," *Journal of Econometrics*, vol. 95, pp. 347-374, 2000.

R. Koenker and G.W. Basset, "Regression quantiles," *Econometrica*, vol. 46, pp. 33-50, 1978.

R. Koenker, P.T. Ng and S. Portnoy, S., “Quantile smoothing splines,” *Biometrika*, vol. 81, pp. 673-680, 1994.

B. Kosko, *Neural networks and fuzzy systems.* A dynamical systems approach to machine intelligence. Englewood Cliffs, N.J.: Prentice-Hall, 1992.

B. Kosko, “Fuzzy Systems as Universal Approximators,” *IEEE Transactions on Computers*, vol. 43, no. 11, pp. 1329-1333, 1994.

C. M. Kuan and H. White, “Artificial Neural Networks: An Econometric Approach,” *Econometric Reviews*, vol. 13, no. 1, pp. 1-91, 1994.

M. Landajo, “A note on model-free regression capabilities of fuzzy systems,” *IEEE Transactions on Systems, Man and Cybernetics-Part B*, vol. 34, no. 1, pp. 645-651, 2004.

M. Landajo, M. J. Río, and R. Pérez, “A Note on Smooth Approximation Capabilities of Fuzzy Systems,” *IEEE Transactions on Fuzzy Systems*, vol. 9, no. 2, pp. 229-237, 2001.

M. Landajo, J. De Andrés and P. Lorca, “Robust neural modeling for the cross-sectional analysis of accounting information,” *European Journal of Operational Research*, vol. 177, no. 2, pp. 1232-1252, 2007a.

M. Landajo, J. De Andrés, and P. Lorca, “Measuring Firm Performance by Using Linear and Nonparametric Quantile Regressions,” *Journal of The Royal Statistical Society (Series C, Applied Statistics)*, (in press), 2007b.

B. Lev and S. Sunder, “Methodological issues in the use of financial ratios,” *Journal of Accounting & Economics*, vol. 1, no. 6, pp. 187-210, 1979.

Z. H. Mao, Y. D. Li, and X. F. Zhang, “Approximation Capability of Fuzzy Systems Using Translations and Dilations of One Fixed Function as Membership Functions,” *IEEE Transactions on Fuzzy Systems*, vol. 5, no. 3, pp. 468-473, 1997.

J. Mata and J.A.F. Machado, “Firm start-up size: a conditional quantile approach,” *European Economic Review*, vol. 40, no. 6, pp. 1305-1323, 1996.

I. Min and I. Kim, “A Monte Carlo comparison of parametric and non-parametric quantile regressions,” *Applied Economics Letters*, vol. 11, pp. 71-74, 2004.

W.K. Newey and J.L. Powell, “Instrumental variable estimation of nonparametric models,” *Econometrica*, vol. 71, no. 5, pp. 1565-1578, 2003.

H. T. Nguyen, V. Kreinovich, and O. Sirisaengtaksin, "Fuzzy control as a universal control tool," *Fuzzy Sets and Systems*, vol. 80, pp. 71-86, 1996.

A. Pagan and A. Ullah, *Nonparametric Econometrics.* Cambridge University Press, Cambridge, UK.

S. Portnoy, "Local asymptotics for quantile smoothing splines," *The Annals of Statistics*, vol. 25, no. 1, pp. 414-434, 1997.

S. Portnoy and R. Koenker, "The Gaussian hare and the Laplacian tortoise: computability of squared-error vs. absolute-error estimators, with discussion," *Statistical Science*, vol. 12, pp. 279-300, 1997.

J. Racine, "Consistent significance testing for nonparametric regression," *Journal of Business and Economic Statistics*, vol. 15, no. 3, pp. 369-378, 1997.

B.D. Ripley, *Pattern Recognition and Neural Networks.* Cambridge: Cambridge University Press, 1996.

M.B. Stinchcombe and H. White, "Some measurability results for extrema of ranom functions over random sets," *Review of Economic Studies*, vol. 59, pp. 495-512, 1992.

C.J. Stone, "Cross-validatory choice and assessment of statistical predictions (with discussion)," *Journal of The Royal Statistical Society, Series B (Methodological)*, vol. 36, no. 2, pp. 111-147, 1974.

M. Sugeno and K. Tanaka, "Successive identification of a fuzzy model and its applications to prediction of a complex system," *Fuzzy Sets and Systems*, 42, pp. 315-334, 1991.

T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 15, pp. 116-132, 1985.

L. X. Wang, *Adaptive fuzzy systems and control: design and stability analysis.* Englewood Cliffs, N. J.: PTR Prentice-Hall, 1994.

L. X. Wang and J. M. Mendel, "Fuzzy Basis Functions, Universal Approximation, and Orthogonal Least-Squares Learning," *IEEE Transactions on Neural Networks*, vol. 3, no. 5, pp. 807-814, 1992.

N. Watanabe and T. Imaizumi, "On least squares methods in fuzzy modeling," in Proceedings of Seventh IFSA World Congress, M. Mares, Ed., vol. 2, pp. 336-341. Prague: Academia, 1997.

H. White, "Learning in Artificial Neural Networks: A Statistical Perspective," *Neural Computation*, vol. 1, no. 4, pp. 425-464, 1989.

H. White, “Connectionist Nonparametric Regression: Multilayer Feedforward Networks Can Learn Arbitrary Mappings,” *Neural Networks*, vol. 3, pp. 535-549, 1990.

H. White, “Nonparametric estimation of conditional quantile using neural networks,” in *Artificial Neural Networks: Approximation and Learning Theory* (ed. H. White), pp. 191-205. Oxford: Blackwell Publishers, 1992.

H. White and J.M.Wooldridge, “Some Results for Sieve Estimation with Dependent Observations,” in W. Barnett, J. Powell and G. Tauchen, eds., *Nonparametric and Semi-Parametric Methods in Econometrics and Statistics*, pp. 459-493. New York: Cambridge University Press, 1991.

G. Whittington, “Some basic properties of accounting ratios,” *Journal of Business Finance & Accounting*, vol. 7, no. 2, pp.219-232, 1980.

K. Yu and M. Jones, “Local linear quantile regression,” *Journal of the American Statistical Association*, vol. 93, pp. 228-237, 1998.

L. Zadeh, “Outline of a New Approach to the Analysis of Complex Systems and Decision Processes,” IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-3, no. 1, pp. 28-44, 1973.