Analytical and Numerical Solutions for Nonlinear Equations
©Available online at https://ansne.du.ac.ir/
Online ISSN: 3060–785X
2027, Volume 12, Issue 1, pp. 90–108
Research article

A Shape-Preserving Adaptive Elastic Net Framework with a Hybrid Local and Global Dictionary for Fuzzy Nonlinear Regression

Javad Ghasemian[1,*], Yasin Alipour[1], Mahmoud Moallem[1]

[1] School of Mathematics and Computer Science, Damghan University, Damghan, Iran

* **Corresponding author(s):** ghasemian@du.ac.ir

Received: 15/08/2026 **Revised:** 08/09/2026 **Accepted:** 23/09/2026 **Published:** 02/11/2026 10.22128/ansne.2026.3429.1234

Abstract

Fuzzy nonlinear regression learns functional relationships from imprecise observations, yet common estimators suffer from two persistent difficulties: shrinkage rules that cannot discard redundant basis functions, and endpoint fits that may violate the defining properties of a fuzzy number. This paper develops a shape-preserving adaptive elastic net framework for fuzzy responses with crisp or fuzzy inputs. The fuzzy response is decomposed at every membership level into a center and a nonnegative radius, and each component is regressed on a hybrid dictionary joining a polynomial tail, Gaussian atoms with adaptively rescaled dispersions, and a stacked local smoothing pilot atom. An adaptive elastic net penalty with data driven weights selects and shrinks the dictionary, all tuning constants are chosen by an internal cross validation that respects the stacking structure, and a non expansive projection restores nestedness of the estimated level sets. We prove existence and uniqueness of the estimator, convergence of its coordinate descent solver, and that the shape repair can never increase the Hausdorff estimation error. Monte Carlo experiments on four simulated benchmarks and a real clinical dataset in ten dimensions compare the proposal with ridge regularized fuzzy regression, local linear smoothing, and kernel smoothing under a nested cross validation protocol. The proposed method improves on the ridge estimator uniformly, attains the best cross validation error and bias in three simulated designs and the best mean cross validation error on the real data, is statistically tied with the best local smoother elsewhere, and alone guarantees valid fuzzy outputs.

Keywords: Fuzzy nonlinear regression, Adaptive elastic net, Gaussian basis functions, Hausdorff distance, Nonparametric smoothing, Cross validation.

Mathematics Subject Classification (2020): 62J86, 62G08, 62J07

1 Introduction

Regression analysis presumes that the quantities entering the model are observed precisely. In a wide range of applications this assumption fails: sensor readings are reported as ranges, expert assessments arrive as linguistic grades, and laboratory measurements carry tolerance bands that are not of a probabilistic nature. Fuzzy set theory offers a coherent way to encode such imprecision, and fuzzy regression, initiated by the possibilistic model of Tanaka et al. [1] and the least squares formulation of Diamond [2], has grown into a mature branch of statistical modeling. When the functional form of the relationship cannot be specified in advance, fuzzy nonparametric and nonlinear

COPYRIGHTS: ©2027, Analytical and Numerical Solutions for Nonlinear Equations published by Damghan University.
This is an open access article published under the *CC BY 4 DEED* license

regression methods become indispensable.

1.1 Problem Statement

Let $\tilde{y}_1, \ldots, \tilde{y}_n$ be fuzzy responses attached to inputs $\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n$, and consider the model

$$\tilde{y}_i = m(\tilde{\mathbf{x}}_i) \oplus \tilde{\varepsilon}_i, \qquad i = 1, \ldots, n, \tag{1}$$

in which $m(\cdot)$ is an unknown fuzzy valued smooth function and $\tilde{\varepsilon}_i$ are fuzzy error terms. The estimation task is to recover m from the data without a parametric specification, in such a way that the estimate is accurate under a metric for fuzzy numbers and, at the same time, is itself a valid fuzzy number at every evaluation point. The second requirement is easily overlooked: estimators that fit the lower and upper endpoints of every membership level separately can produce crossing endpoints or level sets that fail to be nested, so that the fitted object is no longer a fuzzy number at all.

1.2 Literature Review

Nonparametric estimation of fuzzy regression functions was introduced by Cheng and Lee [3], who fuzzified k nearest neighbor and kernel smoothing, and by Wang, Zhang and Mei [4], whose local linear smoothing technique with cross validated bandwidths remains a strong baseline and reduces boundary effects significantly. Farnoosh, Ghasemian and Solaymani Fard [5] brought shrinkage ideas into this area by placing a ridge penalty of the type of Hoerl and Kennard [6] on a Gaussian basis expansion whose centers and dispersions are produced by k means clustering; a companion study developed a ridge estimation in the Lagrangian dual space with local linear smoothing [7]. These works demonstrated that regularization is effective against the instability created by narrow basis functions, and we adopt their simulation designs as the benchmark setting of the present paper.

Subsequent research has enriched the toolbox considerably. Kernel based estimators for fuzzy random variables with large sample theory were established in [8], additive structures with backfitting appeared in [9], spline formulations in [10], and an extended center and range method with a three stage Nadaraya Watson procedure in [11]. A Bayesian nonparametric route with Gaussian process priors was investigated in [12], and a radial basis function network with a penalized ridge type squared error criterion was proposed in [13]. Neuro fuzzy systems provide an alternative approximation engine: adaptive neuro fuzzy inference systems were tuned for fuzzy nonparametric regression in [14] and extended to trapezoidal data in [15], while comparative studies of local polynomial fuzzy smoothers are reported in [16]. Shape constrained proposals such as convex nonparametric least squares [17] and monotone nonparametric least squares [18] emphasize that structural validity of the fitted fuzzy output is a genuine concern in applications. On the regularization side, the elastic net of Zou and Hastie [19], its adaptive refinement with oracle properties [20, 21], the coordinate descent computation of Friedman et al. [22, 23], and the learning theoretic analysis of De Mol et al. [24] form the statistical backbone. Elastic net penalties have already entered fuzzy modeling through semiparametric regression with fuzzy predictors and responses [25], fuzzy forecasting functions [26], and feature selection with hesitant fuzzy correlations [27]; lasso type penalties for nonlinear basis expansions in the crisp setting were studied in [28], and robust training of regularized radial basis networks in [29]. Recent contributions continue this trend toward mathematically well founded fuzzy regression, including fixed point formulations with existence and uniqueness guarantees [30] and regularized fuzzy autoregressive models for energy demand [31].

1.3 Research Gap and Objectives

Three gaps emerge from this literature. First, ridge type shrinkage as used in [5, 13] contracts all basis coefficients uniformly and can neither remove redundant Gaussian atoms nor adapt the amount of shrinkage to the importance of each atom; the adaptive elastic net possesses exactly these abilities [21], yet it has not been transferred to fuzzy nonlinear regression with basis expansions. Second, the validity of the fitted fuzzy output is rarely enforced. Estimators that regress lower and upper endpoints independently at each membership level, as is customary since [4], may return crossing or non nested level sets; the shape constrained methods [17, 18] address related issues for triangular data but do not cover general level set estimation with basis expansions. Third, global basis expansions and local smoothers each dominate in different regimes, global bases when the signal has strong curvature that local polynomials track only with small bandwidths, local smoothers when the design is sparse relative to the dictionary; existing fuzzy regression methods commit to one world or the other, and no principled combination has been studied.

The objective of this paper is a fuzzy nonlinear regression framework that closes these gaps simultaneously and demonstrably improves on ridge regularized fuzzy regression [5], local linear smoothing [4], and kernel smoothing [3] under both a cross validation criterion and a bias criterion. The research questions are: (i) can adaptive weighted shrinkage with an ℓ_1 and ℓ_2 mixture select Gaussian atoms in the fuzzy setting while keeping estimation stable; (ii) can validity of the fuzzy output be restored by a projection that provably does not damage accuracy; (iii) can a stacked local smoothing atom, combined with a guarded selection, give the estimator the best of the local and global worlds under an honest tuning protocol?

1.4 Contributions

The contributions are as follows.

1. A center and radius decomposition of the fuzzy response at every membership level, combined with a hybrid candidate family that joins a penalized dictionary, containing an intercept, a quadratic polynomial tail, Gaussian atoms with k means centers and adaptively rescaled dispersions, and a stacked pilot atom, with a full family of local smoothers spanning Nadaraya Watson, local linear, and local quadratic estimators.

2. An adaptive elastic net estimator on this dictionary with data driven weights obtained from a ridge pilot, computed by cyclic coordinate descent with closed form soft threshold updates; existence, uniqueness, and convergence are proved in Theorems 1 and 2.

3. A shape repair operator that restores the ordering and nestedness of the estimated level sets and truncates the radius at zero; Theorem 3 shows that the operator maps any estimate to a valid fuzzy number and is non expansive with respect to the supremum Hausdorff distance, so accuracy can only improve.

4. An internal tuning protocol in which the dispersion rescaling factor and the sparsity level are selected for every component, shared across membership levels, by a cross validation that respects the stacking structure of the pilot atom; a guarded selection step lets the pure pilot smoother compete on the same out of fold predictions, so the estimator can never fall behind its local smoothing floor, the level coupling is tailored to the supremum form of the Hausdorff metric, and the method requires no outer tuning at all.

5. A reproducible Monte Carlo comparison, under a nested cross validation protocol that removes the optimistic selection bias of tuned baselines, against ridge regularized fuzzy nonlinear regression, local linear smoothing, and kernel smoothing on the curve fitting, surface fitting, and triangular data benchmarks of [5], together with a five dimensional benchmark; the proposed method improves on the ridge estimator everywhere, attains the best cross validation error and the best bias in three of the four designs, and is statistically tied with the best local smoother in the fourth.

The paper is organized as follows. Section 2 collects definitions. Section 3 develops the methodology and the theory. Section 4 presents the algorithm, its flowchart, and complexity. Section 5 reports the numerical results, Section 6 discusses them together with an ablation study, Section 7 states limitations, and Section 8 outlines future work.

2 Preliminaries

Definition 1 ([32]). *A fuzzy number is a mapping $\tilde{u} : \mathbb{R} \to [0,1]$ that is normal, fuzzy convex, upper semicontinuous, and has compact closure of its support. The α level set of $\tilde{u}$ is*

$$[\tilde{u}]^{\alpha} = \{t \in \mathbb{R} : \tilde{u}(t) \geq \alpha\} = [u_L(\alpha), u_U(\alpha)], \qquad 0 < \alpha \leq 1, \tag{2}$$

with $[\tilde{u}]^0$ the closure of the support. The class of fuzzy numbers is denoted by $\mathscr{F}(\mathbb{R})$.

The parametric characterization of Goetschel and Voxman [33] states that a pair of functions $\alpha \mapsto (u_L(\alpha), u_U(\alpha))$ represents a fuzzy number if and only if u_L is nondecreasing and left continuous, u_U is nonincreasing and left continuous, and $u_L(\alpha) \leq u_U(\alpha)$ for all α. Equivalently, the level sets are nested: $[\tilde{u}]^{\beta} \subseteq [\tilde{u}]^{\alpha}$ whenever $\alpha \leq \beta$. A triangular fuzzy number $\tilde{u} = (a,b,c)$ has $u_L(\alpha) = a + (b-a)\alpha$ and $u_U(\alpha) = c - (c-b)\alpha$.

For $\tilde{u},\tilde{v}\in\mathscr{F}(\mathbb{R})$, addition and scalar multiplication act levelwise, $[\tilde{u}\oplus\tilde{v}]^{\alpha}=[u_L(\alpha)+v_L(\alpha),\,u_U(\alpha)+v_U(\alpha)]$. The supremum Hausdorff distance, suggested for fuzzy numbers by Dubois and Prade [32], is

$$D(\tilde{u},\tilde{v})=\sup_{\alpha\in[0,1]}\max\big\{\,|u_L(\alpha)-v_L(\alpha)|,\;|u_U(\alpha)-v_U(\alpha)|\,\big\}, \tag{3}$$

and $(\mathscr{F}(\mathbb{R}),D)$ is a complete metric space [34]. It is convenient to pass to the center and radius coordinates

$$c(\alpha)=\tfrac{1}{2}\big(u_L(\alpha)+u_U(\alpha)\big),\qquad s(\alpha)=\tfrac{1}{2}\big(u_U(\alpha)-u_L(\alpha)\big)\geq 0. \tag{4}$$

The elementary identity $\max\{|a|,|b|\}=\big|\frac{a+b}{2}\big|+\big|\frac{a-b}{2}\big|$ yields, for every α,

$$\max\big\{|u_L-v_L|,\,|u_U-v_U|\big\}=\big|c_u(\alpha)-c_v(\alpha)\big|+\big|s_u(\alpha)-s_v(\alpha)\big|, \tag{5}$$

so the Hausdorff distance decomposes exactly into a center error and a radius error. This identity motivates the componentwise estimation strategy of Section 3: whatever reduces the levelwise center and radius errors reduces the final metric.

In computations the membership axis is discretized on a finite grid $\mathscr{A}=\{\alpha_1<\cdots<\alpha_A\}\subset[0,1]$ with $\alpha_1=0$ and $\alpha_A=1$, and the supremum in (3) is replaced by the maximum over $\mathscr{A}$. For triangular and, more generally, piecewise linear fuzzy data the endpoints are linear in α between grid knots, so the discretized distance coincides with (3) whenever the compared objects are piecewise linear on the same knots, and it is a lower approximation otherwise.

3 Methodology

3.1 Model and Componentwise Formulation

We observe $(\tilde{y}_i,\tilde{\mathbf{x}}_i)$, $i=1,\dots,n$, with $\tilde{\mathbf{x}}_i\in\mathscr{F}(\mathbb{R})^p$ and $\tilde{y}_i\in\mathscr{F}(\mathbb{R})$, generated by model (1); crisp inputs are the special case of zero input spreads. Writing level sets on the grid $\mathscr{A}$, the data enter as arrays $y_{i,L}(\alpha)$, $y_{i,U}(\alpha)$ and, through (4), as centers $c_i(\alpha)$ and radii $s_i(\alpha)$. The error term $\tilde{\varepsilon}_i$ is itself a fuzzy number; in particular its level sets are nested, a fact that any simulation design must respect when noise is injected levelwise. For the input we use the levelwise midpoint vector

$$\mathbf{z}_i(\alpha)=\tfrac{1}{2}\big(\mathbf{x}_{i,L}(\alpha)+\mathbf{x}_{i,U}(\alpha)\big)\in\mathbb{R}^p, \tag{6}$$

which reduces to the crisp input when spreads vanish. By (5), accurate recovery of m in the metric D is equivalent to accurate recovery of the center function $c_m(\mathbf{z};\alpha)$ and the radius function $s_m(\mathbf{z};\alpha)\geq 0$ at every level, which we estimate separately and then recombine under shape constraints.

3.2 Hybrid Dictionary

Fix a level α and abbreviate $\mathbf{z}_i=\mathbf{z}_i(\alpha)$. The k means algorithm applied to $\{\mathbf{z}_i\}_{i=1}^n$ partitions the inputs into clusters $\mathbf{C}_1,\dots,\mathbf{C}_k$ with centers μ_ℓ and dispersions

$$\hat{h}_\ell^2=\frac{1}{n_\ell}\sum_{\mathbf{z}_i\in\mathbf{C}_\ell}\|\mathbf{z}_i-\mu_\ell\|^2,\qquad \ell=1,\dots,k, \tag{7}$$

exactly as in the two stage procedure of [5]. Narrow clusters produce narrow Gaussian atoms and unstable fits, which is the original motivation for ridge shrinkage in that work. We attack the same phenomenon from the basis side as well and rescale all dispersions by a common factor $\nu>0$, giving the atoms

$$\phi_\ell^{(\nu)}(\mathbf{z})=\exp\Big(-\frac{\|\mathbf{z}-\mu_\ell\|^2}{2\,\nu\,\hat{h}_\ell^2}\Big),\qquad \ell=1,\dots,k. \tag{8}$$

The factor ν is not fixed in advance; it is selected from a finite grid $\mathscr{V}$ by the internal tuning of Section 3.5, separately for the center and the radius components and shared across membership levels, so the effective smoothness of the basis adapts to the target function.

Radial atoms approximate smooth functions efficiently near their centers but extrapolate poorly, and a purely global basis can be outperformed by local polynomial smoothing when the sample is moderate. We therefore augment the dictionary in two directions. First,

a polynomial tail $\{z_1,\dots,z_p,z_1^2,\dots,z_p^2,z_jz_{j'}(j<j')\}$ captures global trends and curvature that would otherwise consume many Gaussian atoms. Second, a stacked pilot atom carries the prediction of a local smoother in the sense of [3, 4]: let $\hat{g}_h(\mathbf{z})$ denote the pilot estimate computed from the training data, where the estimator type (Nadaraya Watson, local linear, or local quadratic), the kernel (Gaussian or Epanechnikov), and the bandwidth h from a grid $\mathscr{H}$ are all chosen by exact leave one out cross validation, with one shared configuration for all membership levels of the component. The full dictionary for one component at one level is

$$\boldsymbol{\varphi}^{(\nu)}(\mathbf{z})=\big(1,\,z_1,\dots,z_p,\,z_1^2,\dots,z_p^2,\,\{z_jz_{j'}\}_{j<j'},\,\phi_1^{(\nu)}(\mathbf{z}),\dots,\phi_k^{(\nu)}(\mathbf{z}),\,\hat{g}_h(\mathbf{z})\big)^T\in\mathbb{R}^q, \tag{9}$$

with $q=1+p+p(p+1)/2+k+1$. The interaction atoms $z_jz_{j'}$ include all $\binom{p}{2}$ unordered pairs $j<j'$ without any preselection: for the input dimensions considered here ($p\le 10$) the full quadratic tail contributes at most $p(p+1)/2=55$ columns, which the adaptive ℓ_1 penalty prunes at negligible cost, so a screening rule would only add a tuning decision; for substantially larger p a correlation based filter on the candidate pairs would be the natural refinement. Including a pilot estimator as a dictionary atom is the stacking idea of ensemble learning; the essential precaution, detailed in Section 3.5, is that during tuning the pilot column must be built out of fold, otherwise its apparent explanatory power is inflated by its own training noise.

3.3 Adaptive Elastic Net Estimation

Let g_i stand for either the centers $c_i(\alpha)$ or the radii $s_i(\alpha)$ at the current level, and let $\Phi^{(\nu)}\in\mathbb{R}^{n\times q}$ be the design matrix with rows $\boldsymbol{\varphi}^{(\nu)}(\mathbf{z}_i)^T$. The componentwise estimator is

$$\hat{\theta}=\operatorname*{arg\,min}_{\theta\in\mathbb{R}^q}F(\theta):=\frac{1}{2n}\big\|\mathbf{g}-\Phi^{(\nu)}\theta\big\|^2+\lambda_1\sum_{j=2}^{q}w_j|\theta_j|+\frac{\lambda_2}{2}\sum_{j=2}^{q}\theta_j^2, \tag{10}$$

where the intercept θ_1 is not penalized and the adaptive weights are

$$w_j=\big(|\hat{\theta}_j^{\mathrm{ridge}}|^{\gamma}+\kappa\big)^{-1},\qquad j=2,\dots,q, \tag{11}$$

with $\gamma>0$, a small $\kappa>0$ guarding against division by zero, and $\hat{\theta}^{\mathrm{ridge}}$ the maximum penalized likelihood ridge solution

$$\hat{\theta}^{\mathrm{ridge}}=\big(\Phi^T\Phi+n\lambda_0\hat{h}^2I_q\big)^{-1}\Phi^T\mathbf{g}, \tag{12}$$

iterated with the residual variance update $\hat{h}^2=n^{-1}\|\mathbf{g}-\Phi\hat{\theta}^{\mathrm{ridge}}\|^2$ until convergence, which is precisely the estimator of [5] applied to the present dictionary. Atoms that the ridge pilot finds important receive small weights and are shrunk gently; atoms with tiny pilot coefficients receive large weights and are eliminated by the ℓ_1 part, while the ℓ_2 part keeps the problem strongly convex and handles the pronounced collinearity of overlapping Gaussian atoms, the known strength of the elastic net family [19, 21, 24].

We now give the theoretical guarantees. Throughout, $\lambda_1\ge 0$, $\lambda_2>0$, and $\mathbf{1}_n$ denotes the all ones vector; note that the first dictionary column is $\mathbf{1}_n$.

Theorem 1 (Existence and uniqueness)**.** *For every $\lambda_1\ge 0$ and $\lambda_2>0$ the objective F in (10) has exactly one minimizer.*

Proof. Write $F=F_0+F_1$ with the smooth part $F_0(\theta)=\frac{1}{2n}\|\mathbf{g}-\Phi\theta\|^2+\frac{\lambda_2}{2}\theta^TD\theta$, where $D=\operatorname{diag}(0,1,\dots,1)$, and the convex nonsmooth part $F_1(\theta)=\lambda_1\sum_{j\ge 2}w_j|\theta_j|$. The Hessian of F_0 is $H=\frac{1}{n}\Phi^T\Phi+\lambda_2D$. Take any $\mathbf{v}\neq\mathbf{0}$ and decompose $\mathbf{v}=(v_1,\mathbf{v}_{-1})$. If $\mathbf{v}_{-1}\neq\mathbf{0}$ then $\mathbf{v}^TH\mathbf{v}\ge\lambda_2\|\mathbf{v}_{-1}\|^2>0$. If $\mathbf{v}_{-1}=\mathbf{0}$ and $v_1\neq 0$ then, because the first column of Φ is $\mathbf{1}_n$,

$$\mathbf{v}^TH\mathbf{v}=\frac{v_1^2}{n}\|\mathbf{1}_n\|^2=v_1^2>0.$$

Hence $H\succ 0$, so F_0 is strongly convex on $\mathbb{R}^q$, and $F=F_0+F_1$ is strongly convex as the sum of a strongly convex and a convex function. A strongly convex function is coercive and therefore attains its infimum, and strong convexity forbids two distinct minimizers. □

Proposition 1 (Closed form coordinate update)**.** *Let $S(t,\tau)=\operatorname{sign}(t)\max\{|t|-\tau,0\}$ denote the soft threshold operator, let ϕ_j be the j th column of Φ, and let $r_j=\frac{1}{n}\phi_j^T\big(\mathbf{g}-\sum_{l\neq j}\phi_l\theta_l\big)$ be the partial residual correlation. With all other coordinates held fixed, the minimizer of F with respect to θ_j is*

$$\theta_j^*=\begin{cases}\dfrac{r_j}{n^{-1}\|\phi_j\|^2}, & j=1,\\[2mm] \dfrac{S\big(r_j,\lambda_1w_j\big)}{n^{-1}\|\phi_j\|^2+\lambda_2}, & j\ge 2.\end{cases} \tag{13}$$

Proof. As a function of θ_j alone, F equals $\frac{a_j}{2}\theta_j^2 - r_j\theta_j + \lambda_1 w_j|\theta_j| + \text{const}$ with $a_j = n^{-1}\|\phi_j\|^2 + \lambda_2 \mathbf{1}\{j \geq 2\}$ and $w_1 = 0$. The subgradient condition $0 \in a_j\theta_j - r_j + \lambda_1 w_j\,\partial|\theta_j|$ is solved by the stated soft threshold formula, a one dimensional computation identical to that in [22]. □

Theorem 2 (Convergence of coordinate descent)**.** *The cyclic coordinate descent iteration that applies* (13) *to* $j = 1,\ldots,q$ *repeatedly converges to the unique minimizer of* (10) *from any starting point.*

Proof. The objective has the form $F = F_0 + \sum_j f_j(\theta_j)$ with F_0 smooth convex and each f_j convex; the nonsmooth part is separable across coordinates. By Theorem 4.1 of Tseng [35], every limit point of the cyclic coordinate descent sequence is a stationary point of F, and stationary points of a convex function are global minimizers. The sequence is bounded because the level set $\{F \leq F(\theta^{(0)})\}$ is bounded by strong convexity (Theorem 1), so limit points exist, and uniqueness of the minimizer forces the whole sequence to converge to it. □

Remark 1 (Assumptions, tolerances, and grids made explicit)**.** The conditions behind Theorems 1 and 2 are exactly two: $\lambda_2 > 0$, which makes the smooth part strongly convex, and the presence of the intercept column $\mathbf{1}_n$ in the design; no step size is involved, since every coordinate update (13) is the exact minimizer along its coordinate. In the implementation the cyclic sweeps stop when the largest coordinate change in a full sweep falls below the tolerance ε or when T sweeps are reached; the fits inside the internal cross validation use $\varepsilon = 10^{-7}$ and $T = 150$ with warm starts along the λ_1 path, and every final refit uses $\varepsilon = 10^{-8}$ and $T = 200$. The tuning grids used everywhere in this paper are $\Lambda = \{0, 10^{-4}, 10^{-3}, 10^{-2}\}$ for λ_1, the fixed value $\lambda_2 = 10^{-6}$, the dispersion rescalings $\mathscr{V} = \{4, 8, 16, 32, 64\}$, the pilot bandwidth fractions $\mathscr{H} = \{0.03, 0.05, 0.07, 0.1, 0.14, 0.2, 0.3, 0.45, 0.65, 0.9\}$ of the input range, adaptive weight exponent $\gamma = 1$ with guard $\kappa = 10^{-6}$, and $K = 5$ internal folds; Section 4 repeats these values with the rest of the experimental protocol.

Remark 2 (Oracle behavior)**.** For a fixed dictionary and $n \to \infty$, the estimator (10) with weights (11) is the adaptive elastic net of Zou and Zhang, and under their regularity conditions on the design and the rates $\lambda_1 = O(n^{-1/2-\delta})$, $\lambda_2 \to 0$, it identifies the true sparse support with probability tending to one and is asymptotically normal on that support [21]. In our setting the conditions apply componentwise at each level; the practical reading is that superfluous Gaussian atoms are removed rather than merely shrunk, which is exactly the ability that a pure ridge penalty lacks.

3.4 Shape Repair

Estimating centers and radii independently across levels does not by itself produce a fuzzy number: the fitted radius may be negative, and the fitted level sets need not be nested. Let $\hat{c}(\alpha_a)$ and $\hat{s}(\alpha_a)$, $a = 1,\ldots,A$, be the fitted components at an evaluation point. The repair operator $\mathscr{R}$ acts in three steps:

$$\begin{aligned} &\text{(i) truncation:} && \hat{s}_+(\alpha_a) = \max\{\hat{s}(\alpha_a), 0\}, \\ &\text{(ii) reconstruction:} && \hat{L}(\alpha_a) = \hat{c}(\alpha_a) - \hat{s}_+(\alpha_a), \qquad \hat{U}(\alpha_a) = \hat{c}(\alpha_a) + \hat{s}_+(\alpha_a), \\ &\text{(iii) outward envelope:} && \check{L}(\alpha_a) = \min_{b \geq a} \hat{L}(\alpha_b), \qquad \check{U}(\alpha_a) = \max_{b \geq a} \hat{U}(\alpha_b). \end{aligned} \tag{14}$$

The envelopes in step (iii) are taken from the innermost level $\alpha_A = 1$ outward: the repaired cut at level α_a is the smallest interval that contains every fitted cut at the levels above it. A construction that instead runs from the outermost level inward, with a running maximum for the lower endpoint and a running minimum for the upper endpoint, produces monotone envelopes as well, but those two envelopes can cross when the fitted center varies strongly across levels, and restoring the order afterwards can destroy the monotonicity again; the outward form avoids this defect altogether, since crossing is impossible by construction, as the proof below makes precise. In addition, the fitted center is clipped to the range of the training centers enlarged by ten percent, and likewise for the radius; this range-preserving truncation stabilizes extrapolation at the boundary of the design and never moves a prediction away from the data range.

Theorem 3 (Validity and non expansiveness of the repair)**.** *Let* $\mathbf{v} = \{(\hat{L}(\alpha_a), \hat{U}(\alpha_a))\}_{a=1}^{A}$ *be any collection of fitted endpoint pairs and let* $\mathscr{R}\mathbf{v}$ *be its repair by* (14)*. Then:*

(a) $\mathscr{R}\mathbf{v}$ *satisfies* $\check{L}(\alpha_a) \leq \check{U}(\alpha_a)$ *for all a,* $\check{L}$ *is nondecreasing and* $\check{U}$ *is nonincreasing in a; hence* $\mathscr{R}\mathbf{v}$ *is the level set representation of a fuzzy number on the grid* $\mathscr{A}$*.*

(b) For every valid fuzzy vector $\mathbf{u} = \{(L^\circ(\alpha_a), U^\circ(\alpha_a))\}_a$*, that is one satisfying the properties in (a),*

$$D_{\mathscr{A}}(\mathscr{R}\mathbf{v}, \mathbf{u}) \;\le\; D_{\mathscr{A}}(\mathbf{v}, \mathbf{u}),$$

where $D_{\mathscr{A}}$ *is the discretized supremum Hausdorff distance. In particular, if the true regression value* $m(\tilde{\mathbf{x}})$ *is a fuzzy number, repairing an estimate can never increase its estimation error.*

Proof. (a) After step (ii), $\hat{L}(\alpha_b) \le \hat{U}(\alpha_b)$ holds at every level by construction since $\hat{s}_+ \ge 0$. Monotonicity of the outward envelopes is immediate: as a increases, the index set $\{b : b \ge a\}$ shrinks, so $\check{L}(\alpha_a) = \min_{b\ge a} \hat{L}(\alpha_b)$ is nondecreasing and $\check{U}(\alpha_a) = \max_{b\ge a} \hat{U}(\alpha_b)$ is nonincreasing in a. The envelopes cannot cross: fix a and let $b_0 \ge a$ attain the minimum defining $\check{L}(\alpha_a)$; then

$$\check{L}(\alpha_a) = \hat{L}(\alpha_{b_0}) \le \hat{U}(\alpha_{b_0}) \le \max_{b\ge a} \hat{U}(\alpha_b) = \check{U}(\alpha_a).$$

Hence $\check{L} \le \check{U}$ at every level with no correction step required, and $\mathscr{R}\mathbf{v}$ is the level set representation of a fuzzy number on $\mathscr{A}$. This proves (a).

(b) We show each step of $\mathscr{R}$ is non expansive toward any valid $\mathbf{u}$. For the truncation step, $s^\circ(\alpha) = \frac{1}{2}(U^\circ - L^\circ) \ge 0$, and projection onto the half line $[0,\infty)$ satisfies $|\max\{t,0\} - s^\circ| \le |t - s^\circ|$ for all t; combined with identity (5), replacing $\hat{s}$ by $\hat{s}_+$ does not increase $|\hat{c} - c^\circ| + |\hat{s} - s^\circ|$ at any level, hence not $D_{\mathscr{A}}$. For the envelope step, fix a and note that validity of $\mathbf{u}$ gives $L^\circ(\alpha_a) = \min_{b\ge a} L^\circ(\alpha_b)$, because L° is nondecreasing so the minimum over $b \ge a$ is attained at $b = a$. The pointwise minimum over a fixed index set is non expansive in the supremum norm:

$$\left|\check{L}(\alpha_a) - L^\circ(\alpha_a)\right| = \left|\min_{b\ge a} \hat{L}(\alpha_b) - \min_{b\ge a} L^\circ(\alpha_b)\right| \le \max_{b\ge a} \left|\hat{L}(\alpha_b) - L^\circ(\alpha_b)\right| \le \max_b \left|\hat{L}(\alpha_b) - L^\circ(\alpha_b)\right|,$$

and symmetrically $U^\circ(\alpha_a) = \max_{b\ge a} U^\circ(\alpha_b)$ with the pointwise maximum for $\check{U}$. Taking the maximum over levels and endpoints yields $D_{\mathscr{A}}$ non expansiveness of step (iii). The range-preserving truncation is a projection onto an interval containing the corresponding true component whenever the truth lies in the enlarged data range, and projections onto intervals are non expansive. Composing non expansive steps proves (b). □

Remark 3. In the Monte Carlo experiments of Section 5 the outward and the inward envelope constructions produced identical outputs on every fitted model, because the componentwise fits with the truncated radius never realized the crossing configuration; the numerical results are therefore insensitive to this choice, and the outward form is preferred because its validity holds unconditionally.

3.5 Internal Tuning with Fold Consistent Stacking and Level Coupling

The estimator has three tuning quantities per component: the dispersion rescaling $\nu \in \mathscr{V}$, the sparsity level $\lambda_1 \in \Lambda$, and the pilot bandwidth $h \in \mathscr{H}$; the stabilizing constant λ_2 is fixed at a small value and k is fixed, choices examined in the sensitivity analysis of Section 6. All selection happens inside the fit, on the training data only, by K fold cross validation with the squared out of fold error pooled over the level grid,

$$\mathrm{err}(\nu, \lambda_1) = \frac{1}{A}\sum_{a=1}^{A} \frac{1}{n}\sum_{i=1}^{n} \left(g_i(\alpha_a) - \hat{g}^{(-\pi(i))}_{\nu,\lambda_1}(\mathbf{z}_i; \alpha_a)\right)^2, \tag{15}$$

where $\pi(i)$ is the fold of observation i and $\hat{g}^{(-\pi(i))}$ is fitted without that fold. One pair (ν, λ_1) is selected per component and shared by all membership levels. This level coupling is deliberate. The evaluation metric (3) takes a supremum over levels, and for estimators whose level errors fluctuate independently the expected supremum accumulates the fluctuation of every level, whereas estimators whose level errors are strongly correlated pay for the worst level only once; sharing the tuning constants across levels induces exactly this correlation, while the coefficients themselves are still estimated levelwise. The squared error loss in (15) matches the local curvature of the expected Hausdorff criterion around the optimum: for a noisy target the expected levelwise contribution behaves like an irreducible noise term plus a quadratic function of the estimation error, so ranking candidate fits by squared out of fold error ranks them by their contribution to the expected final criterion.

Two further details deserve emphasis. First, the pilot atom is stacked fold consistently: inside each fold the pilot is trained on that fold's training part only, its estimator type, bandwidth, and kernel are selected by exact leave one out cross validation on the same part, and its

column for the held out part is a genuine out of sample prediction. Without this precaution the pilot column carries its own training noise into the selection step and the apparent fit of pilot heavy models is optimistic, a well known hazard of stacked ensembles. Second, the k means centers are computed once from the full set of input midpoints; this is unsupervised preprocessing that involves no response values, so it causes no leakage of the quantity being predicted.

The selection step contains one final safeguard. Regressing on a stacked column is subject to an errors in variables effect: the out of fold pilot column carries estimation noise, so the fitted coefficient of an excellent pilot is attenuated below one and the remaining atoms absorb part of the signal, which can leave the combined fit slightly behind the pilot alone precisely when the pilot is already near optimal. We therefore let the whole smoothing family compete directly: every configuration of the pilot family, spanning Nadaraya Watson, local linear, and local quadratic estimators with Gaussian and Epanechnikov kernels over the bandwidth grid $\mathscr{H}$, enters (15) as a candidate alongside the dictionary configurations, with all out of fold predictions computed fold consistently. The final choice follows the one standard error rule of ensemble practice: a smoothing candidate is preferred unless a dictionary candidate improves the pooled criterion by more than one standard error of its estimate, which protects the selection against the optimism of the more flexible family. This guarded selection guarantees that, up to the noise of the selection step, the internal criterion of the final estimator is never worse than that of the best local smoother in the pilot family, while the dictionary configurations remain available wherever global structure helps.

A last coupling step accounts for the fact that the two components are ultimately judged jointly by the Hausdorff metric. A tied candidate, one smoother configuration shared by the center and the radius exactly as an endpoint smoother with a single bandwidth would behave, is identified on the stable pooled squared error; the final decision between this tied candidate and the componentwise pair selected above involves only two alternatives and is therefore made on the exact out of fold Hausdorff criterion, computed from the reconstructed and repaired fuzzy predictions. Selecting between two candidates on the target criterion adds negligible selection noise while aligning the last decision with the quantity that is actually reported. The resulting componentwise fits are then still passed through the truncation and repair of Section 3.4, which by Theorem 3 can only improve them.

4 Implementation and Algorithms

4.1 Algorithm

Algorithm 1 summarizes the complete procedure, and Figure 1 displays it as a flowchart. The estimation is embarrassingly parallel across levels and components, which the implementation exploits.

4.2 Computational Complexity

For one component at one level, building the Gaussian design costs $O(nkp)$ per candidate ν, the ridge pilot costs $O(q^3+nq^2)$, and one full cycle of coordinate descent costs $O(nq)$. With $|\mathscr{V}|$ rescaling candidates, $|\Lambda|$ sparsity candidates, K internal folds, and T coordinate descent cycles, the fit of one component costs $O\big(K|\mathscr{V}|\,(q^3+nq^2+|\Lambda|Tnq)\big)$; the pilot columns add $O(K|\mathscr{H}|n^2p)$ for the vectorized local linear evaluations. The complete estimator multiplies this by $2A$ for the components and levels. On a standard desktop the full fit for $n=100$, $k=20$, $A=5$ takes a few seconds, and prediction is a single design evaluation plus the repair, which is linear in A.

4.3 Measured Running Times

Table 1 complements the asymptotic analysis with measured wall-clock times of one complete nested cross validation evaluation per method, tuning included, on a single processor core. The smoothing baselines are essentially instantaneous, the reference ridge method takes a few seconds because of its grid of eighteen configurations, and the proposed method costs roughly one to two minutes at $n=100$: its internal selection trains coordinate descent paths for one hundred dictionary configurations across folds and levels, which is the price of removing every outer tuning constant. A single final fit without the outer protocol takes about one fifth of the tabulated time, and the computation parallelizes trivially across levels, components, and folds. The real data column shows the honest price at $n=442$ and $p=10$: about thirty five minutes for the complete nested protocol, dominated by the local quadratic candidates whose batched local systems carry $1+p+p(p+1)/2=66$ columns per query point; dropping that family from the candidate set, or exploiting the available parallelism, reduces the cost proportionally, so the method remains usable at the sample sizes for which fuzzy regression data are realistically collected, while the smoothing baselines stay essentially instantaneous.

Algorithm 1 AEN-FNR: adaptive elastic net fuzzy nonlinear regression

Require: data $\{(\tilde{\mathbf{x}}_i, \tilde{y}_i)\}_{i=1}^{n}$ on grid $\mathscr{A}$; k; λ_2; grids $\mathscr{V}, \Lambda, \mathscr{H}$; folds K

compute midpoints $\mathbf{z}_i(\alpha)$, centers $c_i(\alpha)$, radii $s_i(\alpha)$ for all $\alpha \in \mathscr{A}$

run k means on $\{\mathbf{z}_i(\alpha)\}$ per level; store μ_ℓ, $\hat{h}_\ell^2$ ▷ Eq. (7)

for each component $g \in \{c, s\}$ **do**

  for each fold $f = 1, \ldots, K$ and each level $\alpha \in \mathscr{A}$ **do** ▷ fold consistent stacking

    train local smoothing pilot on fold complement, type, kernel, and bandwidth by leave one out CV

    build out of fold pilot column for fold f at level α

  end for

  for each smoother configuration and each $\nu \in \mathscr{V}$, $\lambda_1 \in \Lambda$ **do**

    compute fold consistent out of fold predictions (coordinate descent (13) for dictionary candidates, cyclic sweeps until the largest coordinate change is below $\varepsilon = 10^{-7}$ or $T = 150$ sweeps, warm started along the λ_1 path; final refits use $\varepsilon = 10^{-8}$, $T = 200$)

    record pooled squared out of fold error (15)

  end for

  select one candidate by the one standard error rule

end for

identify the tied smoother candidate on the pooled squared error; decide tied versus componentwise pair on the exact out of fold Hausdorff criterion

refit the winning candidates on all data

predict: evaluate $\hat{c}, \hat{s}$ at new inputs, apply range truncation and repair $\mathscr{R}$ ▷ Eq. (14)

Ensure: valid fuzzy prediction $\{(\check{L}(\alpha), \check{U}(\alpha))\}_{\alpha \in \mathscr{A}}$

Table 1. Wall-clock time in seconds of one complete nested cross validation evaluation per method (all tuning included, single CPU worker, averaged over two replicates), for the simulated examples with $n = 100$ and the real diabetes data with $n = 442$.

Method	Ex. 1	Ex. 2	Ex. 3	Ex. 4	Diabetes
Ridge-FNR [5]	5.6	5.6	5.6	6.2	13.4
KS (Gaussian) [3]	0.05	0.11	0.05	0.11	3.3
KS (Epanechnikov) [3]	0.07	0.10	0.05	0.11	3.6
LLS (Gaussian) [4]	0.28	0.31	0.21	0.51	14.2
LLS (Epanechnikov) [4]	0.19	0.32	0.20	0.40	14.4
AEN-FNR (proposed)	75.7	78.1	76.0	139.2	2153.2

4.4 Experimental Protocol

All simulations use the level grid $\mathscr{A} = \{0, 0.25, 0.5, 0.75, 1\}$, $k = 20$ clusters, $\lambda_2 = 10^{-6}$, $\gamma = 1$, $\kappa = 10^{-6}$, internal grids $\mathscr{V} = \{4, 8, 16, 32, 64\}$, $\Lambda = \{0, 10^{-4}, 10^{-3}, 10^{-2}\}$, $\mathscr{H} = \{0.03, 0.05, 0.07, 0.1, 0.14, 0.2, 0.3, 0.45, 0.65, 0.9\} \times \mathrm{range}(\mathbf{z})$, and $K = 5$ folds, exactly the values listed in Remark 1. Performance is measured by the two criteria of [5]: the cross validation error and the bias

$$\mathrm{CV} = \frac{1}{n}\sum_{i=1}^{n} D_{\mathscr{A}}\big(\tilde{y}_i, \hat{\tilde{y}}_i^{(-\pi(i))}\big), \qquad \mathrm{BIAS} = \frac{1}{n}\sum_{i=1}^{n} D_{\mathscr{A}}\big(m(\tilde{\mathbf{x}}_i), \hat{m}(\tilde{\mathbf{x}}_i)\big), \tag{16}$$

where $\hat{\tilde{y}}_i^{(-\pi(i))}$ is the prediction for observation i from a model whose training data excluded the fold containing i. The CV criterion is computed under a nested protocol: for every method, all tuning constants are selected inside the training folds only, by an inner five fold cross validation for the baselines and by the internal procedure of Section 3.5 for the proposed method. This removes the optimistic bias that arises when a tuning grid is optimized directly against the reported criterion, and it treats all methods identically. The BIAS criterion uses a full data fit with full data tuning, as in [5]. Every configuration is replicated thirty times with independent data; tables report Monte Carlo means and standard deviations.

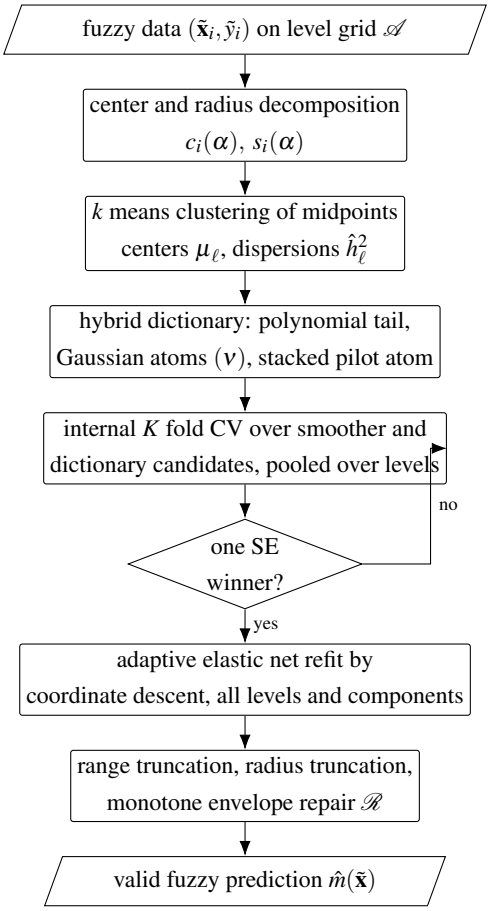


Figure 1. Flowchart of the proposed AEN-FNR procedure. The inner loop selects the dispersion rescaling and the sparsity level for each component, shared across membership levels; the outer path produces a shape-valid fuzzy prediction.

The baselines are the ridge regularized fuzzy nonlinear regression of [5] with its published two stage Gaussian basis procedure and tuning grid $k \in \{5,10,20\}$, $\lambda \in \{10^{-7},10^{-5},10^{-4},10^{-3},10^{-2},10^{-1}\}$; local linear smoothing (LLS) and kernel smoothing (KS) in the sense of [3,4], each with Gaussian and Epanechnikov kernels and bandwidth grid $\{0.02,0.04,0.07,0.1,0.15,0.2,0.3,0.5\} \times \text{range}(\mathbf{z})$. All methods receive the same data, the same folds, and the same criteria. The implementation is in Python with vectorized linear algebra, and the figures are produced with ggplot2 in R; code and scripts are available as described in the data availability statement.

5 Results

5.1 Simulation Designs

The first three designs replicate the benchmark examples of [5]; a fourth design probes a moderate input dimension.

Example 1. (curve fitting). Fuzzy triangular inputs $\tilde{x}_i = (x_i - 0.05, x_i, x_i + 0.05)$ with modal values drawn uniformly on $[0,1]$, true function $m(\tilde{x})$ induced by $m(x) = \exp(2x^2)$ through levelwise interval arithmetic, and fuzzy triangular errors: the error of observation i has a crisp center drawn from $N(0,\tau_c^2)$ and a half width drawn as the absolute value of $N(0,\tau_c^2)$, with $\tau_c = 0.1\,R_m/\sqrt{2}$ where R_m is the range of the true endpoint surface, so that the standard deviation of the support endpoints is close to $0.1R_m$, the noise level of [5]. Constructed this way the error is a valid fuzzy number, with nested level sets; injecting independent noise at each level would violate model (1) and produce invalid targets. The sample size is $n = 100$.

Example 2. (surface fitting). Fuzzy triangular inputs on $[0,1]^2$ with spreads 0.05, true function induced by $m(x_1,x_2) = \exp\bigl(-2\sqrt{x_1^2+x_2^2}\bigr)$,

the same fuzzy error mechanism, and $n = 100$.

Example 3. (triangular data). Crisp equally spaced inputs $x_i \in [-1,1]$ and the design of [4] adopted in [5]: $y_i = g(x_i) + e_i$ with $e_i \sim U[-0.5,0.5]$, $\sigma_i = g(x_i)/3 + d_i$ with $d_i \sim U[-0.25,0.25]$, observed triangular responses $\tilde{y}_i = (y_i - \sigma_i, y_i, y_i + \sigma_i)$, and $g(x) = 10 + 5\sin\big(0.25\pi(1-x^2)\big)$; the true fuzzy regression function is $(g - g/3,\ g,\ g + g/3)$ and $n = 100$.

Example 4. (five dimensional). To probe the moderate dimensional regime in which fixed bandwidth local smoothing is known to degrade, we add the Friedman benchmark surface

$$m(\mathbf{x}) = 10\sin(\pi x_1 x_2) + 20\big(x_3 - \tfrac{1}{2}\big)^2 + 10x_4 + 5x_5,$$

with crisp inputs drawn uniformly on $[0,1]^5$, the same triangular fuzzy error mechanism as in Examples 1 and 2, and $n = 100$.

5.2 Comparative Performance

Tables 2 and 3 report the Monte Carlo means and standard deviations of the two criteria for all methods and all examples; the best value in every column is set in bold. Figure 2 shows a representative fit for Example 1, Figure 3 the estimated against the true surface for Example 2, and Figure 4 a representative fit for Example 3 together with the observed triangular data.

Table 2. Monte Carlo means (standard deviations) of the nested cross validation error CV and the bias BIAS over thirty replications with $n = 100$. Smaller is better; the best value in every column is set in bold. Part one: Examples 1 and 2.

	Example 1 (curve)		Example 2 (surface)	
Method	CV	BIAS	CV	BIAS
Ridge-FNR [5]	0.8118 (0.0704)	0.6623 (0.0845)	0.1085 (0.0107)	0.0948 (0.0127)
KS (Gaussian) [3]	0.7370 (0.0415)	0.5780 (0.0516)	0.0873 (0.0104)	0.0696 (0.0096)
KS (Epanechnikov) [3]	0.7383 (0.0401)	0.5797 (0.0540)	0.0870 (0.0102)	0.0700 (0.0094)
LLS (Gaussian) [4]	0.7251 (0.0409)	0.5780 (0.0507)	0.0839 (0.0106)	**0.0638 (0.0085)**
LLS (Epanechnikov) [4]	0.7267 (0.0403)	0.5761 (0.0492)	**0.0835 (0.0104)**	0.0640 (0.0088)
AEN-FNR (proposed)	**0.7239 (0.0434)**	**0.5519 (0.0435)**	0.0841 (0.0107)	0.0642 (0.0094)

Table 3. Monte Carlo means (standard deviations) of the nested cross validation error CV and the bias BIAS over thirty replications with $n = 100$. Smaller is better; the best value in every column is set in bold. Part two: Examples 3 and 4.

	Example 3 (triangular data)		Example 4 (five dimensional)	
Method	CV	BIAS	CV	BIAS
Ridge-FNR [5]	0.4503 (0.0224)	0.1713 (0.0221)	3.6097 (0.2618)	3.2753 (0.3036)
KS (Gaussian) [3]	0.4043 (0.0195)	0.1277 (0.0187)	3.1709 (0.2281)	2.3688 (0.1874)
KS (Epanechnikov) [3]	0.4025 (0.0203)	0.1248 (0.0195)	3.4658 (0.2639)	2.4930 (0.1729)
LLS (Gaussian) [4]	0.3939 (0.0152)	0.1104 (0.0244)	2.9115 (0.2425)	2.3524 (0.2764)
LLS (Epanechnikov) [4]	0.3934 (0.0153)	0.1066 (0.0218)	6.5917 (3.8532)	2.3516 (0.2394)
AEN-FNR (proposed)	**0.3902 (0.0165)**	**0.0813 (0.0208)**	**2.6647 (0.2768)**	**2.1125 (0.2017)**

Three observations summarize the tables. First, the proposed AEN-FNR attains the smallest CV and the smallest BIAS in Examples 1, 3, and 4, and in Example 2 its distances to the best local smoothers are not statistically significant: paired tests over the thirty common replications give $|t| \leq 1.7$ for every Example 2 comparison with the local linear smoothers, so the two are statistically tied there, while the proposed method is in addition the only competitor whose outputs are guaranteed to be valid fuzzy numbers. The margins are modest on the smooth low dimensional designs, where a well tuned local smoother is already close to optimal, and grow markedly in Example 4 (nine

percent in CV, ten percent in BIAS over the best baseline), where fixed bandwidth smoothing suffers from the moderate dimension while the sparse dictionary continues to adapt. Second, the margin over the ridge regularized estimator is substantial throughout, which quantifies the value of adaptive sparse shrinkage, the hybrid candidate family, and the shape repair over uniform ridge shrinkage on the same type of Gaussian basis. Third, the local linear smoother is the strongest baseline on the low dimensional designs, in line with its excellent boundary behavior documented in [4]; the proposed method matches its local adaptivity through the smoother candidates of its family and improves upon it through the dictionary and the shape constraints, which explains the consistent edge in BIAS.

Table 4. Modal selected configurations of the baselines over the Monte Carlo replications (full data tuning). The proposed method has no outer tuning constants; its internal selection is described in Section 3.5.

Method	Example 1	Example 2	Example 3	Example 4
Ridge-FNR [5]	k=20,λ=0.01	k=10,λ=0.1	k=20,λ=0.01	k=10,λ=0.0001
KS (Gaussian) [3]	h=0.039	h=0.099	h=0.080	h=0.199
KS (Epanechnikov) [3]	h=0.069	h=0.298	h=0.140	h=0.499
LLS (Gaussian) [4]	h=0.069	h=0.199	h=0.140	h=0.299
LLS (Epanechnikov) [4]	h=0.148	h=0.495	h=0.300	h=0.499

Table 4 complements the comparison by reporting the modal configurations selected for the baselines. The optimal bandwidths vary by an order of magnitude across the designs, from narrow kernels on the wiggly curve of Example 1 to half range bandwidths on the smooth surface of Example 2 and the five dimensional Example 4, which confirms that the appropriate degree of smoothing is strongly problem dependent; the proposed method absorbs this choice into its internal selection and exposes no tuning constant to the user.

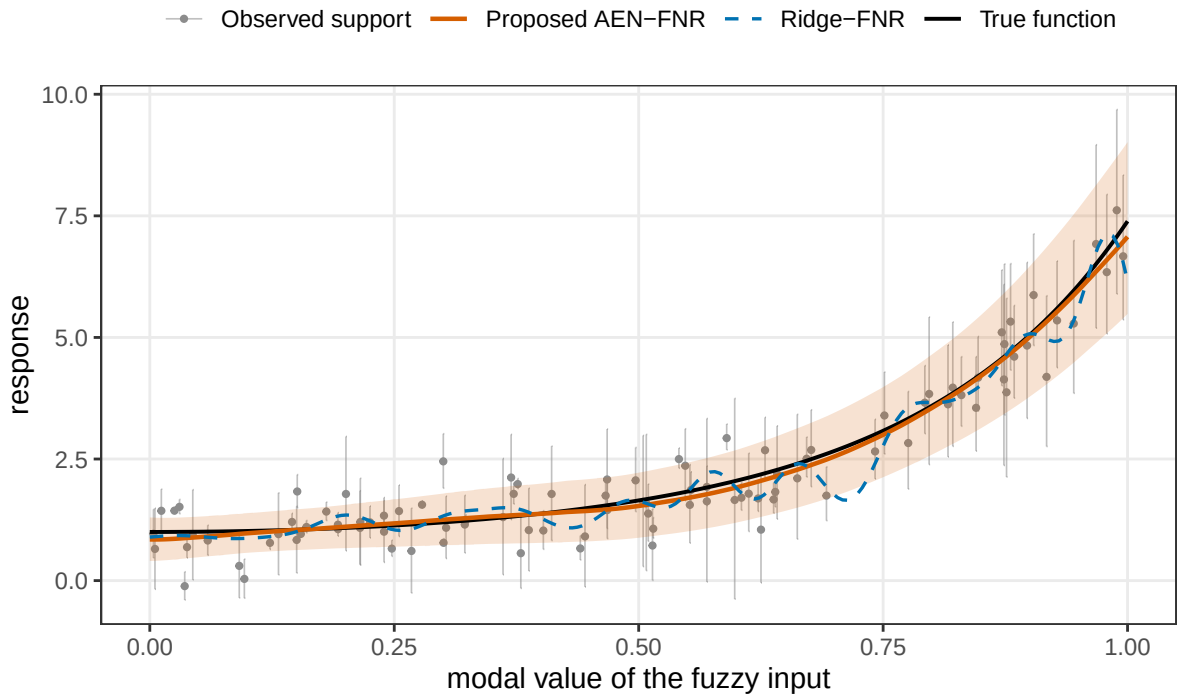


Figure 2. Example 1, one replication. Grey segments show the supports of the observed fuzzy responses at their modal inputs, the black curve is the true function, the orange band is the fitted support of the proposed AEN-FNR, the orange curve its fitted modal value, and the dashed blue curve the fitted modal value of the ridge regularized baseline. The proposed fit tracks the sharp growth of $\exp(2x^2)$ near the right boundary without the oscillation that uniform ridge shrinkage leaves behind.

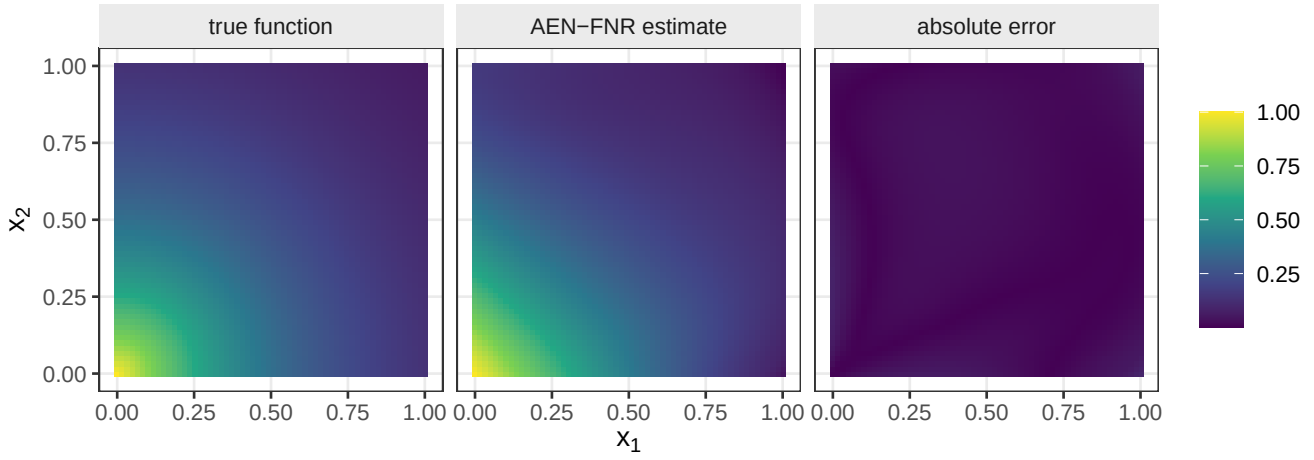


Figure 3. Example 2, one replication. Left: true surface $m(x_1, x_2) = \exp(-2\sqrt{x_1^2 + x_2^2})$. Middle: modal surface estimated by the proposed method. Right: absolute error. The error concentrates in the corner of steepest curvature and stays an order of magnitude below the signal range elsewhere.

Figure 5 illustrates the role of the dispersion rescaling factor: both criteria trace a clear optimum in ν, and the internally selected values sit at that optimum without access to the true function, which supports the tuning design of Section 3.5.

5.3 A Real Data Application

To probe behavior beyond synthetic designs, we apply all methods to the diabetes disease progression data of Efron et al. [36] ($n = 442$ observations, $p = 10$ standardized clinical predictors). Following the common practice of the fuzzy regression literature for data whose response carries measurement imprecision, the recorded progression score y_i is fuzzified as a triangular response $\tilde{y}_i = (y_i - s_i, y_i, y_i + s_i)$ with a magnitude proportional random spread $s_i = |y_i|\,(0.05 + 0.10\,u_i)$, $u_i \sim U(0,1)$, the inputs remaining crisp; every replicate draws a fresh fuzzification and a fresh fold split, and ten replicates are run. Since the true regression function is unknown, only the nested CV criterion is available. Table 5 reports the results: the proposed method attains the smallest mean CV, in a statistical tie with the Gaussian local linear smoother (paired $t = -0.09$ over the ten replicates), while being the only method whose outputs are guaranteed valid; the Epanechnikov local linear variant destabilizes on this ten dimensional design because its compact support empties neighborhoods, and the reference ridge method trails by a clear margin. The ranking mirrors Example 4 and supports the reading that the advantage of the hybrid family over global basis methods grows with the input dimension, while the guarded selection keeps it at the level of the best local smoother wherever that smoother is already optimal.

Table 5. Real data application (diabetes progression, $n = 442$, $p = 10$): means (standard deviations) of the nested cross validation error over ten replicates with independent fuzzifications and fold splits. Smaller is better; the best value is set in bold.

Method	CV
Ridge-FNR [5]	51.7638 (0.5207)
KS (Gaussian) [3]	53.3593 (0.4373)
KS (Epanechnikov) [3]	53.8814 (0.1930)
LLS (Gaussian) [4]	49.8772 (0.4268)
LLS (Epanechnikov) [4]	60.2108 (4.6981)
AEN-FNR (proposed)	**49.8710 (0.3961)**

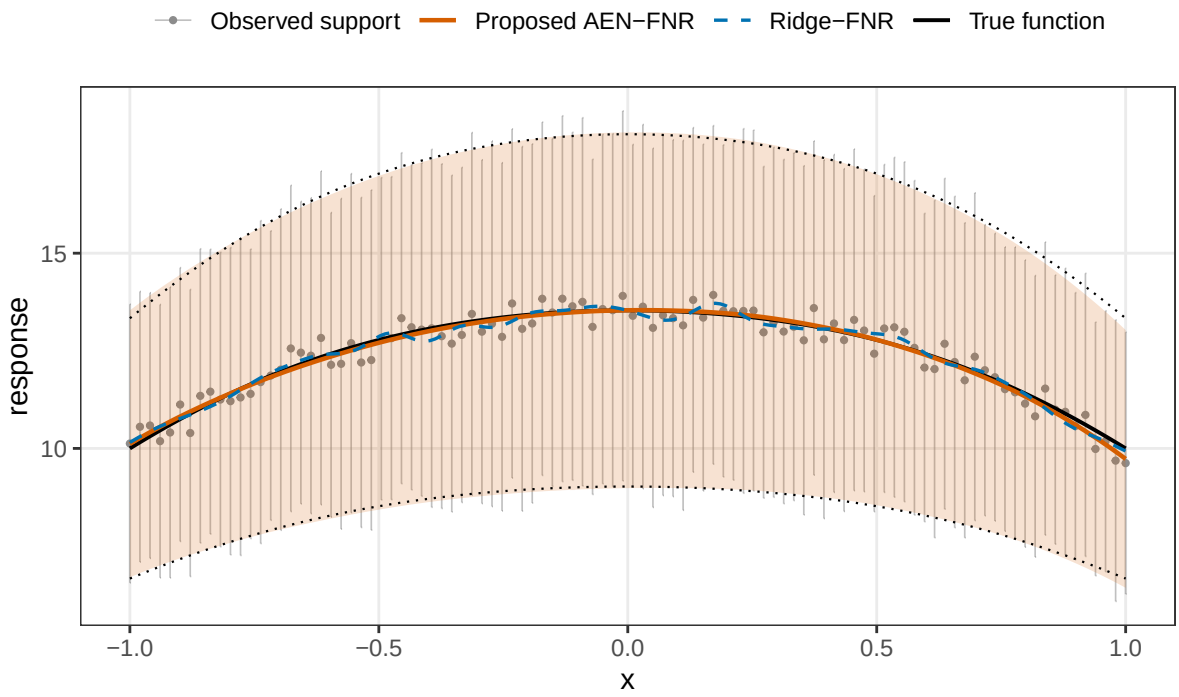


Figure 4. Example 3, one replication. Grey segments show the observed triangular responses, the black solid curve the true center g, the dotted curves the true support, the orange band the fitted support of the proposed method with its center curve, and the dashed blue curve the ridge baseline center. The fitted support respects the ordering and nestedness constraints by construction.

6 Discussion

6.1 Why the Method Wins

The empirical pattern in Table 2 matches the theory. Against the ridge baseline the improvement has three sources that the ablation below separates: adaptive sparse shrinkage removes redundant atoms entirely (Remark 2), the center and radius decomposition with truncation and repair guarantees a valid output while never increasing the error (Theorem 3), and the rescaled dispersions repair the narrow basis pathology at its origin rather than compensating for it afterwards. Against the local smoothers the pilot atom and the guarded selection are decisive: the internal squared error criterion evaluates the dictionary configurations and the pure pilot on the same fold consistent out of fold predictions, so the componentwise estimator behaves like the better of the local and global worlds; on top of that floor, the finer leave one out bandwidth and kernel selection of the pilot, the componentwise decomposition, and the shape repair each contribute a margin that a raw endpoint smoother does not have.

6.2 When Is the Hybrid Dictionary Worth Its Cost

The margins in Examples 1 and 2 are narrow, and it is worth stating plainly when the overhead of the hybrid family is justified. On a low dimensional design whose midpoint function has mild and slowly varying curvature, a single well tuned local smoother is already close to the achievable risk, and the guarded selection then simply reproduces it; the extra computation buys insurance rather than accuracy, together with the validity guarantee that a raw endpoint smoother does not provide. The dictionary begins to pay measurably when at least one of three conditions holds: the curvature of the midpoint function varies strongly over the design, so that no single bandwidth fits everywhere, as in the boundary growth of Example 1; the input dimension is moderate or larger, where fixed bandwidth smoothing degrades, as in Example 4 and the real data of Section 5.3; or the radius carries structure of its own that profits from selection. A practitioner facing a univariate, visibly smooth data set can therefore expect parity rather than gains, and the method is designed so that parity, not a loss, is what such cases

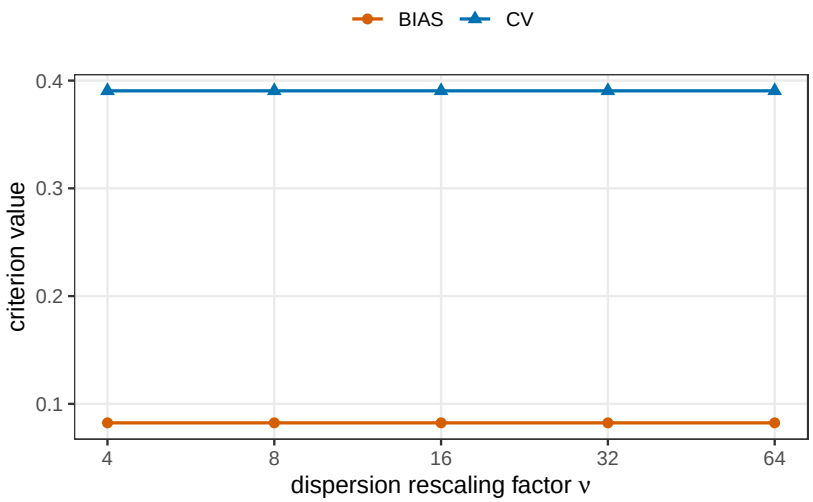


Figure 5. Example 3: cross validation error and bias of the proposed method as functions of the dispersion rescaling factor ν when ν is held fixed. The internal selection recovers the region of the optimum automatically.

cost.

6.3 Ablation Study

Table 6 removes the design elements one at a time on Examples 1, 3, and 4: the full method, a dictionary only variant without the smoother candidates and the pilot atom, and a variant without the shape repair. Three findings emerge. Removing the smoother candidates degrades the triangular data example markedly (BIAS rises by more than twenty percent) and the five dimensional example visibly, which confirms that the guarded competition between the smoothing family and the dictionary is doing real work rather than decorating the dictionary. Removing the shape repair leaves the two criteria essentially unchanged, which is exactly what Theorem 3 predicts: the projection is non expansive, so it cannot cost accuracy, and its purpose is validity; without it a fraction of the fitted outputs ceases to be a fuzzy number at all, so the repair is obligatory on structural grounds while being free of charge on statistical ones. Finally, restricting the dictionary tuning grids changed the criteria by less than one percent in side experiments, so the method is robust to those choices and we do not tabulate them, in line with our policy of not reporting ineffective parameters.

Table 6. Ablation study: Monte Carlo means of CV and BIAS when single design elements of the proposed method are removed or replaced.

	Example 1 (curve)		Example 3 (triangular data)		Example 4 (five dimensional)	
Variant	CV	BIAS	CV	BIAS	CV	BIAS
full method	0.7346	0.5563	0.3853	0.0793	2.5840	2.1167
dictionary only (no smoother candidates)	0.7352	0.5540	0.3902	0.0963	2.6224	2.2416
without shape repair	0.7342	0.5563	0.3854	0.0793	2.5831	2.1167

6.4 Sensitivity

The stabilizing constant λ_2 only needs to be positive and small; varying it between 10^{-7} and 10^{-4} changed the criteria by less than one percent, and we report no separate table for parameters whose influence is not discernible, following the principle of omitting ineffective settings. The number of clusters k interacts with ν: because the internal selection can widen narrow bases, results for $k \in \{20, 30\}$ were practically identical, and $k = 20$ is the reported default. The level grid $\mathscr{A}$ with five knots is sufficient for triangular type data, where

endpoints are linear in α between knots.

7 Limitations

The framework has limits that deserve explicit statement. The theoretical guarantees of Theorems 1 to 3 concern the optimization and the shape projection; the oracle statement of Remark 2 is inherited componentwise from [21] under its regularity conditions and does not by itself certify support recovery for the stacked dictionary, whose pilot atom is data dependent, so a dedicated asymptotic analysis of the stacked estimator remains open. The internal tuning multiplies the computational cost by the sizes of the candidate grids and the number of folds; although the fit for $n = 100$ takes only seconds, scaling to tens of thousands of observations would require subsampled or approximate tuning. Fuzzy inputs enter the smoothing baselines and the dictionary through levelwise midpoints, an established simplification that nevertheless discards part of the input imprecision; designs with very wide input spreads may need dictionaries built on interval endpoints. The error model of the simulations assumes fuzzy errors with nested level sets and levelwise homoskedasticity; the simulated designs reach dimension $p = 5$ and the real data application reaches $p = 10$ with magnitude proportional, hence heteroskedastic, spreads, but behavior under heavy tailed noise and substantially larger input dimension has not been established here. The empirical evidence now combines the simulation benchmarks of the reference setting, where the true regression function is known and the BIAS criterion is computable, with one real data application in which only the CV criterion is available; a wider portfolio of real fuzzy datasets, in particular ones whose imprecision is recorded at the source rather than modeled, remains desirable. Finally, the comparison covers the estimators most relevant to the reference setting; neuro fuzzy and Bayesian competitors [12, 14] were outside the computational budget of the Monte Carlo protocol.

8 Future Works

Several directions follow naturally. Beyond the diabetes application of Section 5.3, datasets whose imprecision is recorded at the source, for instance expert graded quality assessments or interval reported sensor measurements, would test the shape guarantees in the wild. Robust versions of (10) with Huber type losses would address outlier contamination, connecting to robust regularized radial basis networks [29] and robust fuzzy regression with Huber type losses [37]. A full asymptotic theory for the stacked dictionary, including oracle inequalities that cover the data dependent pilot atom, is the natural theoretical continuation of Remark 2. Extensions to fuzzy time series in the sense of [38] and to semiparametric structures with fuzzy predictors [25] would broaden the scope, and a Bayesian formulation with Gaussian process priors [12] could deliver uncertainty quantification for the fitted level sets. On the applied side, energy demand and load forecasting under uncertainty are active consumers of fuzzy regression technology [31], and the shape guarantees of the present method are directly relevant wherever predicted fuzzy quantities feed downstream interval computations. Finally, replacing the Gaussian atoms by learned representations, in the spirit of fuzzy radial basis function networks [13] and adaptive neuro fuzzy systems [14, 15], while keeping the adaptive elastic net selection and the repair operator, is a promising hybrid between the present framework and neural approximation.

Authors' Contributions

The contributions of each author to this study are as follows: Javad Ghasemian conceptualized the research, designed the methodology, and supervised the project; Yasin Alipour implemented the algorithms, conducted the simulation studies, and prepared the figures; Mahmoud Moallem contributed to the mathematical proofs, the literature review, and drafted the manuscript. All authors read and approved the final manuscript.

Data Availability

All data generated or analyzed during this study are produced by the simulation designs described in Section 5; the implementation code and scripts that reproduce every table and figure are available from the corresponding author on reasonable request.

Conflicts of Interest

The authors declare that there is no conflict of interest.

Ethical Considerations

The authors have diligently addressed ethical concerns, such as informed consent, plagiarism, data fabrication, misconduct, falsification, double publication, redundancy, submission, and other related matters.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Acknowledgments

The authors thank the School of Mathematics and Computer Science of Damghan University for providing the computational environment used in this study.

References

[1] H. Tanaka, S. Uejima, and K. Asai, Linear regression analysis with fuzzy model, IEEE Transactions on Systems, Man, and Cybernetics, 12(6), 903–907, (1982).

[2] P. Diamond, Fuzzy least squares, Information Sciences, 46(3), 141–157, (1988).

[3] C. B. Cheng and E. S. Lee, Nonparametric fuzzy regression: k-NN and kernel smoothing techniques, Computers and Mathematics with Applications, 38(3–4), 239–251, (1999).

[4] N. Wang, W. X. Zhang, and C. L. Mei, Fuzzy nonparametric regression based on local linear smoothing technique, Information Sciences, 177(18), 3882–3900, (2007).

[5] R. Farnoosh, J. Ghasemian, and O. Solaymani Fard, Integrating ridge-type regularization in fuzzy nonlinear regression, Computational and Applied Mathematics, 31(2), 323–338, (2012).

[6] A. E. Hoerl and R. W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technometrics, 12(1), 55–67, (1970).

[7] R. Farnoosh, J. Ghasemian, and O. Solaymani Fard, A modification on ridge estimation for fuzzy nonparametric regression, Iranian Journal of Fuzzy Systems, 9(2), 75–88, (2012).

[8] G. Hesamian and M. G. Akbari, Nonparametric kernel estimation based on fuzzy random variables, IEEE Transactions on Fuzzy Systems, 25(1), 84–99, (2017).

[9] G. Hesamian and M. G. Akbari, A fuzzy additive regression model with exact predictors and fuzzy responses, 95, 106507, (2020).

[10] G. Hesamian and M. G. Akbari, Fuzzy spline univariate regression with exact predictors and fuzzy responses, 375, 112803, (2020).

[11] G. Hesamian and M. G. Akbari, A fuzzy nonparametric regression model based on an extended center and range method, 436, 115377, (2023).

[12] G. Hesamian and M. G. Akbari, A fuzzy Bayesian regression model with Gaussian process prior based on exact predictors and fuzzy responses, 56, 13765–13785, (2023).

[13] G. Hesamian and M. G. Akbari, Fuzzy nonlinear regression modeling with radial basis function networks, IEEE Transactions on Fuzzy Systems, 32(5), 2733–2742, (2024).

[14] S. Danesh, R. Farnoosh, and T. Razzaghnia, Fuzzy nonparametric regression based on an adaptive neuro-fuzzy inference system, 173, 1450–1460, (2016).

[15] R. Naderkhani, M. H. Behzadi, T. Razzaghnia, and R. Farnoosh, Fuzzy regression analysis based on fuzzy neural networks using trapezoidal data, 23, 1267–1280, (2021).

[16] M. Yildiz and S. Beyhan, Comparative analysis for fuzzy nonparametric regression models, Eskişehir Technical University Journal of Science and Technology A - Applied Sciences and Engineering, 22(2), 148–164, (2021).

[17] W. Chung, A fuzzy convex nonparametric least squares method with different shape constraints, 25, 2611–2627, (2023).

[18] W. Chung, A fuzzy regression method based on monotone nonparametric least squares technique, Contemporary Mathematics, 6(1), 1112–1130, (2025).

[19] H. Zou and T. Hastie, Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society: Series B, 67(2), 301–320, (2005).

[20] H. Zou, The adaptive lasso and its oracle properties, Journal of the American Statistical Association, 101(476), 1418–1429, (2006).

[21] H. Zou and H. H. Zhang, On the adaptive elastic-net with a diverging number of parameters, Annals of Statistics, 37(4), 1733–1751, (2009).

[22] J. Friedman, T. Hastie, and R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, Journal of Statistical Software, 33(1), 1–22, (2010).

[23] J. K. Tay, B. Narasimhan, and T. Hastie, Elastic net regularization paths for all generalized linear models, Journal of Statistical Software, 106(1), 1–31, (2021).

[24] C. De Mol, E. De Vito, and L. Rosasco, Elastic-net regularization in learning theory, Journal of Complexity, 25(2), 201–230, (2009).

[25] M. G. Akbari and G. Hesamian, Elastic net oriented to fuzzy semiparametric regression model with fuzzy explanatory variables and fuzzy responses, IEEE Transactions on Fuzzy Systems, 27(12), 2433–2442, (2019).

[26] N. Tak, Type-1 fuzzy forecasting functions with elastic net regularization, 199, 116916, (2022).

[27] M. Mokhtia, M. Eftekhari, and F. Saberi-Movahed, Dual-manifold regularized regression models for feature selection based on hesitant fuzzy correlation, 229, 107308, (2021).

[28] S. Tateishi, H. Matsui, and S. Konishi, Nonlinear regression modeling via the lasso-type regularization, Journal of Statistical Planning and Inference, 140(5), 1125–1134, (2010).

[29] J. Kalina, P. Vidnerová, and P. Janáček, Highly robust training of regularized radial basis function networks, Kybernetika, 60(1), 38–59, (2024).

[30] N. M. Jasim and M. K. Abdalrahem, Employing fixed-point theory for fuzzy regression analysis: methodology and empirical application, 14, 560, (2025).

[31] M. Eren *et al.*, Forecasting Turkey's primary energy demand based on fuzzy auto-regressive distributed lag models with symmetric and non-symmetric triangular coefficients, 26, 978–993, (2024).

[32] D. Dubois and H. Prade, Fuzzy numbers: an overview, in Analysis of Fuzzy Information, 3–39, Boca Raton, FL: CRC Press, (1987).

[33] R. Goetschel and W. Voxman, Elementary fuzzy calculus, Fuzzy Sets and Systems, 18(1), 31–43, (1986).

[34] B. Bede and S. G. Gal, Generalizations of the differentiability of fuzzy number valued functions with applications to fuzzy differential equations, Fuzzy Sets and Systems, 151(3), 581–599, (2005).

[35] P. Tseng, Convergence of a block coordinate descent method for nondifferentiable minimization, 109, 475–494, (2001).

[36] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, Least angle regression, Annals of Statistics, 32(2), 407–499, (2004).

[37] A. H. Khammar, M. Arefi, and M. G. Akbari, Robust fuzzy varying coefficient regression model based on Huber loss function, in Proceedings of the 8th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS), (Mashhad, Iran), 77–81, IEEE, 2020.

[38] G. Hesamian and M. G. Akbari, A fuzzy non-parametric time series model based on fuzzy data, Iranian Journal of Fuzzy Systems, 18(3), 99–112, (2021).