Proc. of the European Safety and Reliability Conference (ESREL 2026)
Edited by José C. Matos, Paulo B. Lourenço, Michael Beer, Daniel Oliveira,
Edoardo Patelli, Hélder S. Sousa, Rui A. Silva and Monica Santamaria
 Published by Research Publishing, Singapore.
doi: 10.3850/ESREL2026061419_esrel26-p26374-cd

Zonotopic representation of multi-variable regression with interval dependent variables

Matthew McCann
Civil and Environmental Engineering, University of Strathclyde, United Kingdom.
E-mail: matthew.mccann@strath.ac.uk

Marco de Angelis
Civil and Environmental Engineering, University of Strathclyde, United Kingdom.
E-mail: marco.de-angelis@strath.ac.uk

This paper extends existing work on imprecise regression by elucidating the set-valued representation of the computational problem of training regression models with interval-valued data. We frame the problem in a zonotopic sense to clearly expose the computational cost and limitations of training imprecise regression models with intervals. We focus on the regression problem where the independent predictor variables are precisely measured while the dependent response variable has uncertainty. Unlike classical regression, the dependent variable here is an interval-valued random variable. We demonstrate that ordinary least-square regression can be reduced to a multi-dimensional Minkowski sum of line segments, each corresponding to an individual data interval. This means the problem is closed to zonotope representation and the uncertainty in the form of intervals can be propagated precisely. Recall that intervals are a crude but efficient method for computing with imprecise probability. The zonotopic representation enables the deconstruction of the feasible set of regression coefficients using affine transformations of the unit hypercube. This avoids the exponential enumeration of vertices characteristic of convex-hull propagation in multi-variable regression. Consequently, the proposed zonotopic regression provides an exact and computationally efficient alternative to the expensive vertex-based computation of tight bounds and a preferable option to numeric-grade linear programming and overly conservative naive interval bounds. Additional advantages include the compact storage of uncertainty structures in generator matrices and a clear geometric interpretation of parameter uncertainty when working with interval-valued data for the dependent variable. Since multilinear regression is closed under affine transformations, this representation can be extended to nonlinear and neural-network regression models providing a framework for zonotopic uncertainty propagation. The method's efficiency is demonstrated on a variety of data affected by imprecision.

Keywords: Zonotopes, Imprecise Regression, Uncertainty Propagation, Interval-Valued Data.

1. Introduction

Regression models are a fundamental statistical tool in engineering and science, used in model validation, sensitivity analysis, predictive modelling, engineering seismology, and risk analysis, just to mention a few. Classical regression models assume precise measurement of dependent values, which can be unrealistic in domains with imprecise measurements leading to epistemic uncertainty. This is important as any measurement without its uncertainty is completely meaningless (Reckhow, 1994). Intervals are sets of real values that lie between two endpoints and is the crudest representation of epistemic uncertainty. Intervals work by specifying bounds for the uncertainty in the measurements, which is conceptually straightforward but introduces some quantification challenges (Ferson et al., 2007). One major challenge with the use of intervals within regression is the dependency problem, in which the dependence amongst interval variables is not taken into account leading to overly wide bounds (Černý et al., 2013). The dependency problem leads to the so called wrapping effect in regression analysis, whereby the regression coefficients make up a non-rectangular relation. Therefore, disregarding this relation for the regression coefficients leads to overly wide expectation bands because the dependence expressed by the non-rectangular relation between the coefficients is lost (Schollmeyer and Augustin, 2015). Considering separate intervals for the regression coef-

ficient is equivalent to mapping the bounding box of the feasible non-rectangular joint set onto the expectation bands, which explains the overly wide bounds. One way to eliminate the wrapping effect is to fully characterise the relation of regression coefficients. In least-square regression models, such relation can be expressed as a convex set and thus it is fully characterised by its vertices. The enumeration of these vertices is possible when the dimension of the control (independent) variable is very small (< 3) thanks to convex-hull propagation algorithms, but it gets computationally impractical very quickly as the dimension of the multi-variable regression problem increases. A brute-force approach of checking all 2^n combinations of interval endpoints is also impractical even for problems with very few data. Regression with interval-valued data has a relatively long history within symbolic data analysis, possibilistic statistics, and imprecise probability theory (Augustin et al., 2014; Billard and Diday, 2000; Cattaneo and Wiencierz, 2012; Černý et al., 2013; Billard and Diday, 2002) Early work by Billard and Diday (2000) introduced the center method, which while simple and computationally efficient, discards valuable uncertainty information and may underestimate uncertainty. Despite this, it forms the conceptual basis for subsequent methods (Neto and De Carvalho, 2017, 2008). More recent approaches, such as the Likelihood-based Imprecise Regression framework, LIR, proposed by Cattaneo and Wiencierz (2012), perform inference directly on imprecise observations without collapsing them to point values. Their framework returns a set of plausible regression functions, providing a distribution-free treatment of imprecise data. However, this generality comes at the cost of significant computational complexity, particularly when the data is highly imprecise, resulting in large identification regions. In Černý et al. (2013), the authors study linear regression models with uncertainty affecting both dependent and independent variables. In the special case of precise inputs and interval outputs, they show that the resulting parameter uncertainty set has a zonotopic structure. However, they note that explicit vertex or facet descriptions may have exponential complexity in the number of observations, motivating the use of approximations. The efficient training of regression models with interval-valued response variables has also recently appeared within neural networks (Sadeghi et al., 2019; Tretiak et al., 2023). Sadeghi et al. (2019) present a method to train neural networks with imprecise data making use of the interval-predictor model, IPM. Their method directly learns prediction intervals by minimising a maximum-error loss using minibatch gradient descent, enabling scalability to large datasets. In virtue of the IPM framework, Sadeghi's method naturally supports imprecision in the training data including uncertainty for the independent variables. Tretiak et al. (2023) propose a single-layer interval neural network, SLINN, for regression with interval-valued response variables. The model is trained using a novel interval-valued loss function based on mean squared error, and employs gradient-based optimization adapted to interval arithmetic. To mitigate the dependency problem, they introduce a mean-value form approximation that produces tight, rigorous bounds. The resulting expectation band contains all possible regression functions consistent with the data. This work builds upon the existing literature by proposing a new zonotopic angle to tackle the imprecise regression problem, in the attempt to ground its formulation into solid and intuitive geometric foundations. Zonotopes enable the exact propagation of interval uncertainty through the ordinary least-squares model, while explicitly solving the interval dependency problem found within regression. With zonotopes, we obtain exact and computationally very efficient bounds on the regression coefficients, avoiding explicit vertex enumeration, and resulting in negligible computational overhead beyond standard least-squares estimation. Consequently, the proposed view provides a scalable and theoretically sound alternative for regression problems involving interval-valued dependent variables.

2. Zonotopes

A zonotope is a centrally symmetric convex polytope that can be expressed as the Minkowski sum of a finite number of multidimensional line seg-

ments. A zonotope in $\mathbb{R}^d$ is defined as

$$\mathcal{Z} := \{\, c + G\xi \mid \xi \in [-1,1]^n \,\} \tag{1}$$

which is also simply denoted $\mathcal{Z} := \langle c, G\rangle$, where $c \in \mathbb{R}^{d\times 1}$ is the center and $G \in \mathbb{R}^{d\times n}$ is the generator matrix, whose columns $g_1, ..., g_n$ represent the individual segments. Each generating vector $g_i \in \mathbb{R}^d$ identifies a line segment centered in zero $(-g_i, g_i) := \{g_i\xi_i \mid \xi_i \in [-1,1]\}$, and thus the zonotope is precisely the Minkowski sum of all such line segments. In interval notation the line segment is $(-g_i, g_i) := g_i[\xi_i]$, where $[\xi_i] = [-1,1]$. Denoting $\oplus$ the Minkowski addition, a zonotope is also denoted $\mathcal{Z} = c + (-g_1, g_1) \oplus (-g_2, g_2) \oplus \cdots \oplus (-g_n, g_n)$. Aside from center translations, zonotopes can be seen as the image under a linear map of the unit hypercube. Therefore any linear transformation of a zonotope is also a zonotope, which can be obtained by simple manipulations of the generator matrices. For example, let $\mathcal{Z} \subset \mathbb{R}^d$, $z \in \mathcal{Z}$ be any point in it, and $A \in \mathbb{R}^{l\times d}$ be any matrix, then the linear map Az is in the zonotope $\langle Ac, AG\rangle$. The sum of two zonotopes is a zonotope whose generator is the horizontal concatenation of the two generators: $\mathcal{Z}_1 + \mathcal{Z}_2 = \langle c_1 + c_2, [G_1, G_2]\rangle$. Any scalar multiplication results in expanding $\lambda > 1$ or contracting $0 < \lambda < 1$ the zonotope: $\lambda\mathcal{Z} = \langle \lambda c, \lambda G\rangle$. The difference can be defined as $\mathcal{Z}_1 - \mathcal{Z}_2 = \langle c_1 - c_2, [G_1, -G_2]\rangle$, however zonotope subtraction corresponds to an outer approximation of the Minkowski difference and thus cannot be used as an inverse of the addition. Thus zonotope arithmetic cannot be used to eliminate the dependence issue arising from subtracting the same interval to itself. Since interval arithmetic provides exact bounds on the Minkowski addition of unrelated line segments, the bounding box of a zonotope can be simply and compactly expressed

$$[\mathcal{Z}] := c \pm \sum_{i=1}^{n} |g_i| \tag{2}$$

or in interval notation $[\mathcal{Z}] := c + \sum_{i=1}^{n} g_i[\xi_i]$.

2.1. *Interval Dependency Problem arising in Regression*

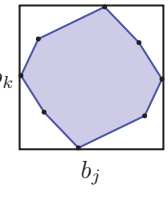
We use zonotopes to eliminate a specific dependency problem arising in regression models with interval dependent variables. This is the case of matrix multiplication in which only the multiplicand column vector has interval-valued components, denoted $A[y]$, where $A \in \mathbb{R}^{d\times n}$, $[y] \in \mathbb{IR}^{n\times 1}$, and $\mathbb{IR}$ is the space of intervals (Neumaier, 1990). It is assumed that the intervals in the column vector $[y]$ display no epistemic dependence, as it is a rigorous albeit conservative model of measurement uncertainty in regression analysis. The row-column products are expanded in the column vector below, where we highlight the interdependence between each component in the form of segments.

$$\begin{pmatrix} a_{11}[y_1] & + & a_{12}[y_2] & + \cdots + & a_{1n}[y_n] \\ a_{21}[y_1] & + & a_{22}[y_2] & + \cdots + & a_{2n}[y_n] \\ \vdots & & \vdots & & \vdots \\ a_{d1}[y_1] & + & a_{d2}[y_2] & + \cdots + & a_{dn}[y_n] \end{pmatrix}$$

segment 1 segment 2 $\cdots$ segment n

The matrix multiplication $A[y]$ evaluates to an interval-valued column vector $[b] \in \mathbb{R}^{d\times 1}$, whose components are interdependent. Given these dependencies, the column vector $[b]$ can be viewed as the sum of n d-dimensional line segments, i.e. a Minkowski sum. From this follows that the dependence between the components of $[b]$ is described by a zonotope.

$$\left.\begin{array}{rcl} \sum_i^n a_{1i}[y_i] & = & [b_1] \\ \sum_i^n a_{2i}[y_i] & = & [b_2] \\ \vdots & = & \vdots \\ \sum_i^n a_{di}[y_i] & = & [b_d] \end{array}\right\}$$



If we express the intervals $[y_i]$ in midpoint-radius notation, $[y_i] = \check{y}_i + r_i[\xi_i]$ then each component becomes $[b_j] = \sum_i^n a_{ji}\check{y}_i + \sum_i^n a_{ji}r_i[\xi_i]$, where $j = 1, ..., d$. Denoting $c_j := \sum_i^n a_{ji}\check{y}_i$ and $g_{ij} := a_{ji}r_i$, it is immediate to see that $[b]$ is a zonotope

whose center is $c = A\check{y}$ and whose generator has columns $g_i = A_{(:,i)}r_i$, so the generator matrix is $G = AD_r$, where $D_r := \text{diag}(r) \in \mathbb{R}^{n\times n}$. The obtained zonotope can be denoted $\mathcal{Z}_b := \{A\check{y} + AD_r\xi \mid \xi \in [-1,1]^{n\times 1}\}$ or simply $\mathcal{Z}_b := \langle A\check{y}, AD_r\rangle$, where $[y] = \check{y} + r[\xi]$.

In regression analysis, b denotes the vector of coefficients. The regression line (or expectation line) is then evaluated on any given control variable $x \in \mathbb{R}^{1\times d}$ where $x = x_1, x_2, ..., x_{d-1}$, using the row column product $x[b] \in \mathbb{R}$. This last multiplication projects the zonotope $\mathcal{Z}_b$ down to a one-dimensional zonotope $\mathcal{Z}_y = \langle xA\check{y}, xAD_r\rangle$ which is the zonotope of the regression line. Since this zonotope has dimension one (same as the dimension of the dependent response variable) taking the bounding box yields no outer approximation. The regression band is thus exactly the bounding box of the regression zonotope

$$[\mathcal{Z}_y] = xA\check{y} + \sum_{i=1}^{n}(xAD_r)_{(:,i)}[\xi_i]. \qquad (3)$$

The cost incurred in computing the regression band for each testing data point is that of doing three row-column multiplications $\mathbb{R}^{1\times d} \times \mathbb{R}^{d\times n} \times \mathbb{R}^{n\times n}$ and an accumulation of n intervals, and it is thus much more efficient than vertex enumeration or linear programming.

3. Multi-linear Regression with Zonotopes

In multi-linear regression the response (or dependent) variable y is related to $d-1$ predictor (or independent) variables $x = x_1, ..., x_{d-1}$. Given a set of n observations $(x_1, y_1), ..., (x_n, y_n)$, the regression model for all $i = 1, ..., n$ is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_{d-1}x_{i,d-1} + \varepsilon_i,$$

which in compact form is $y = X\beta + \varepsilon$, where $X = (x_1^\top, ..., x_n^\top)$, β_j are the regression coefficients and ε_i is a zero-mean constant-variance random variable typically normally distributed, $X \in \mathbb{R}^{n\times d}$ is the matrix of predictor variables whose first column is one, and $y \in \mathbb{R}^{n\times 1}$. The unknown regression coefficients are determined by least-square estimation, which simplifies to

$$\hat{\beta} = (X^\top X)^{-1}X^\top y,$$

provided that the inverse matrix $(X^\top X)^{-1}$ exists, condition that is met if the columns of X are linearly independent. Once $\hat{\beta}$ is known, the expression $\hat{y} = \hat{\beta}x$ evaluates the expected value of y (regression line) given any value of x. When the observed responses are intervals $[y] = [\underline{y}, \overline{y}]$, and expand the value of $\hat{\beta}$ following least-square estimation the regression line becomes

$$[\hat{y}] = x\,(X^\top X)^{-1}X^\top [y],$$

where $x \in \mathbb{R}^{1\times d}$ and $(X^\top X)^{-1}X^\top \in \mathbb{R}^{d\times n}$. We use zonotopes to compute the exact range $[\hat{y}]$. Denoting $A = (X^\top X)^{-1}X^\top$ we use (3) to compute the exact range $[\hat{y}] = [\mathcal{Z}_y] = xA\check{y} + \sum_{i=1}^{n}(xAD_r)_{(:,i)}[\xi_i]$.

3.1. *Nonlinear Regression with Zonotopes*

To extend the proposed method beyond linear models, we adopt a basis expansion approach. By mapping the precise input x into a higher-dimensional feature space, the regression problem remains linear in the parameters whilst still capturing nonlinear relationships within the data. The regression model becomes $y = \Phi\beta + \epsilon$, with Φ representing the matrix of basis functions. Since the independent variables are precise (Tretiak et al., 2023), the transformed problem remains linear in the parameters (Hastie et al., 2009, Chapter 5, pg 139). We denote $A = (\Phi^\top\Phi)^{-1}\Phi^\top$ to proceed like above and create the regression zonotope, where $\Phi \in \mathbb{R}^{n\times M}$ is a univariate polynomial feature map,

$$\Phi = (\phi_0(x),\, \phi_1(x),\, \ldots,\, \phi_{M-1}(x)). \qquad (4)$$

The dimensionality of the parameter space is now M, the number of basis functions, which increases rapidly with the dimension and the degree of the polynomial. For a total-degree polynomial basis of degree p with input dimension d, the number of basis functions is given by $M = \binom{d+p}{p}$. Although tractable for moderate p and d, the rapid growth in M motivates exploration of neural network alternatives. For the multivariate case, we construct the feature map using total-degree tensor products of the univariate monomials.

3.2. *Algorithm*

Zonotopic regression proceeds as follows:

(1) Construct the zonotope from intervals.
(2) Select basis expansion (nonlinear case only).
(3) Compute regression zonotope.
(4) Obtain exact bounds using Eq. (3).

3.2.1. *Illustrative Example*

For the sake of illustrating the method, we take a simple dataset with three interval-valued observations,

$$y_1 \in [-2, 3], \quad y_2 \in [2, 4], \quad y_3 \in [-5, -2],$$
$$x_1 = -1, \quad x_2 = 2, \quad x_3 = 3$$

The observation zonotope is constructed as described resulting in $\mathcal{Z}y$, which has center c_y and generator matrix G_y.

$$c_y = \begin{pmatrix} 0.5 \\ 3 \\ -3.5 \end{pmatrix}, \quad G_y = D_r = \begin{pmatrix} 2.5 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1.5 \end{pmatrix}$$

With $\mathcal{Z}y$ we perform the first of two row-column products. We first transform the zonotope from observation space into parameter space, this is done because we directly observe y not the parameters β. This first transformation maps the set of all possible observations onto the set of all possible parameters that could have produced them. Firstly we compute $A = (X^\top X)^{-1} X^\top \in \mathbb{R}^{d \times n}$. Then a row-column product with $\mathcal{Z}y$, results in zonotope $\mathcal{Z}_{\hat{\beta}}$.

$$X = \begin{pmatrix} 1 & -1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}, \quad A = \begin{pmatrix} 0.69 & 0.23 & 0.08 \\ -0.27 & 0.08 & 0.19 \end{pmatrix}$$

$$c_{\hat{\beta}} = \begin{pmatrix} 0.77 \\ -0.58 \end{pmatrix}, \quad G_{\hat{\beta}} = \begin{pmatrix} 1.73 & 0.23 & 0.12 \\ -0.67 & 0.08 & 0.29 \end{pmatrix}$$

The second row-column product transforms the parameter zonotope from $\mathbb{R}^d$ (parameter space) into $\mathbb{R}^n$ (prediction space). The resulting regression zonotope $\mathcal{Z}_{\hat{y}}$ encapsulates all possible values $\hat{y} = X_{\text{test}}\,\beta$ that are consistent with the uncertain observations. With the testing set $x_{\text{test}} = 0, 1, 3$

$$X_{\text{test}} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 3 \end{pmatrix},$$

the prediction zonotope is $\mathcal{Z}_{\hat{y}} = \langle c_{\hat{y}}, G_{\hat{y}} \rangle$, where

$$c_{\hat{y}} = \begin{pmatrix} 0.77 \\ 0.19 \\ -0.96 \end{pmatrix}, \quad G_{\hat{y}} = \begin{pmatrix} 1.73 & 0.23 & 0.12 \\ 1.06 & 0.31 & 0.4 \\ -0.29 & 0.46 & 0.98 \end{pmatrix}.$$

The expectation bands follows immediately

$$[\hat{y}] = c_{\hat{y}} + \sum_{i=1}^{n} G_{\hat{y}(:,i)}[\xi_i].$$

For example, for $x = 3$, $\hat{y} \in [-2.69, 0.77]$. This interval is much tighter than the naive interval arithmetic result $\hat{y} \in [-2.92, 4.23]$.

4. Experiments

4.1. *Linear Regression Example*

For a holistic comparison with other methods, we first consider a simple linear problem with one independent variable and one interval-valued response variable. The observed data is generated according to a linear model with homoscedastic noise, where the noise magnitude is constant across all values of x, and the observed values are obtained by adding Gaussian noise

$$y_{\text{obs}} = y_{\text{true}} + \epsilon,$$
$$\epsilon \sim \mathcal{N}\big(0.3, 2^2\big),$$
$$[\underline{y}, \overline{y}] = [y_{\text{obs}} - 0.3x,\ y_{\text{obs}} + 0.3x].$$

We generate $n = 10$ observations at $x = 1, ..., 10$. The exact zonotope bands are compared against the exact method of vertices, the outer approximated naive bounds and the inner approximated Monte Carlo bounds. The results in Fig 1 show the overly wide bounds produced by the naive application of interval arithmetic, which highlight the issue of dependency within interval arithmetic. The vertex enumeration method offers an exact solution to this problem. Rather than naively propagating intervals through the regression formula, this approach considers all possible regression lines consistent with the interval data. The extreme predictions of this feasible set occur at the vertices of the coefficients zonogone. By evaluating the regression only at these vertices and taking their envelope, we obtain the tightest possible bounds. While these bounds are exact, this method grows exponentially with the dimension of the

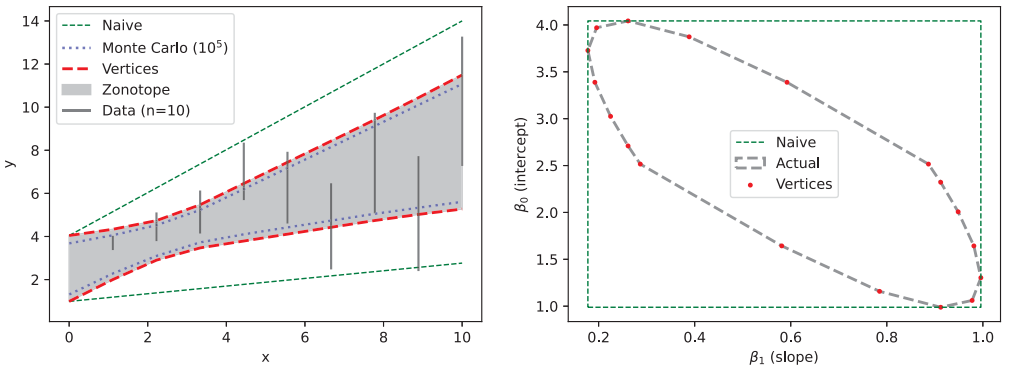


Fig. 1. Left: Imprecise Linear Regression with the computed naive interval bands (thin dashed-line), the zonotope bands (shaded region), the Monte Carlo sampled bands (dotted-line), and the exact bands using vertex enumeration (thick dashed-line) which is an envelope of all feasible regression lines. Right: The naive box (dashed bounding box) compared to the actual feasible parameter set (grey convex shape).

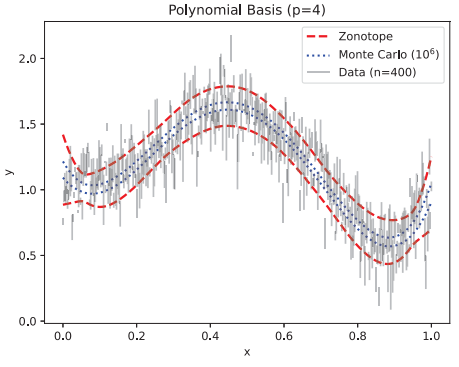


Fig. 2. Nonlinear imprecise regression using a polynomial basis function with the zonotope bands (dashed-line) and the Monte Carlo sampled bands (dotted-line).

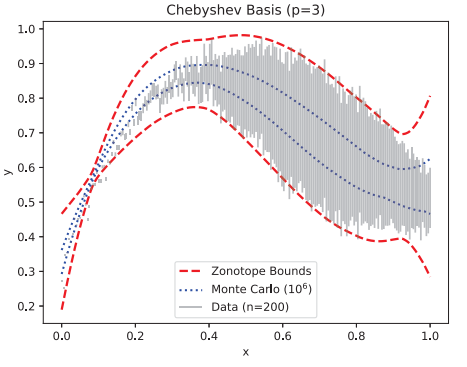


Fig. 3. Nonlinear imprecise regression using a chebyshev basis function with the zonotope bands (dashed-line) and the Monte Carlo sampled bands (dotted-line).

multilinear problem. Zonotopes offer a suitable alternative to this method, which as indicated in Fig. 1 also provide exact bounds but without the exponential cost of enumerating all vertices.

4.2. *Nonlinear Regression Example*

The datasets used to demonstrate this method in the nonlinear context are taken from the work done by Tretiak et al. (2023). In these datasets (Eq. 5, 6), r_i represents the interval radius, so our intervals are defined as $[\underline{y}, \overline{y}] = [y - r_i, y + r_i]$.

$$\begin{aligned} y_i &= \sin(2\pi x^2)e^{-x} + 1 + \mathcal{N}(0, 0.15^2) \\ r_i &\sim |\mathcal{N}(0.1, 0.05^2)| \end{aligned} \tag{5}$$

$$\begin{aligned} y_i &= 5xe^{-3x} + 0.25 + \mathcal{N}(0, 0.02^2) \\ r_i &= 0.2\, e^{(-10(x-0.7)^2)} \end{aligned} \tag{6}$$

For the presented examples we make use of two different basis functions, however there is no requirement for these specific functions to be used. Fig. 2 makes use of a polynomial basis, in which the input is first scaled to $[0, 1]$ and the function is given as $\phi_i(x) = x^i, \quad i = 0, ..., p$. For Fig. 3 the Chebyshev basis is used, in which the input is scaled to $[-1, 1]$, and the function is defined as $\phi_i(x) = \mathcal{T}_i(x)$, where $\mathcal{T}_i(x)$ denotes the Chebyshev polynomial of the first kind of degree i. **Computational Restrictions.** The method was

validated on datasets up to $n = 50{,}000$ (linear) and $n = 30{,}000$ (nonlinear), with runtime ≈ 10 seconds. The nonlinear memory overhead (≈ 20 GB) arises from explicit generator storage and is an implementation artefact; column-wise scaling reduces the requirement to under 10 MB. The intrinsic cost stems from basis expansion, whose size grows combinatorially with degree, this is unavoidable, regardless of interval uncertainty. Neural networks can mitigate this by avoiding explicit expansions.

4.3. *Practical Examples*

To demonstrate the effectiveness of the proposed approach, we have applied this method to some real-world datasets, that may represent practical use cases of the approach. The data in these datasets was split into separate training and testing sets according to the 80:20 rule, in which 80% of the data was used for training and 20% was used for testing.

4.3.1. *House Prices Dataset*

For a univariate nonlinear example we make use of the Ames housing dataset (De Cock, 2011), in which we model the relationship between the year the house was built and its sales price. The sales price variable is intervalised using the method described by Tretiak et al. (2023) where the values are intervalised using a uniformly biased approached,

$$r_i \sim \mathcal{N}(15000, 10000^2), \quad b_i \sim \mathcal{U}(-1, 1),$$
$$m_i = y_i + b_i r_i,$$
$$[\underline{y_i}, \overline{y_i}] = [\, m_i - r_i,\ m_i + r_i \,].$$

To fit this data, Fig. 4, we made use of a 3rd degree polynomial basis function.

4.3.2. *Multivariate Regression Dataset*

For the multivariate case we make use of a fuel consumption dataset, the dataset provides over 22,000 entities describing a variety of vehicles between the years 2000 and 2022. For this data we model the relationship between four independent variables (Engine Size, Cylinder Count, Miles Per Gallon, Fuel Consumption) and one dependent variable (Emissions). In order to make use of this dataset, we had to first intervalise the precise values. This was done using the same method to intervalise the values in artifical dataset #1 (Eq. 5). The results of which can be seen in Fig. 5. This result makes use of a 3rd degree Chebyshev basis and with a regularization value of $\lambda = 1 \times 10^{-4}$.

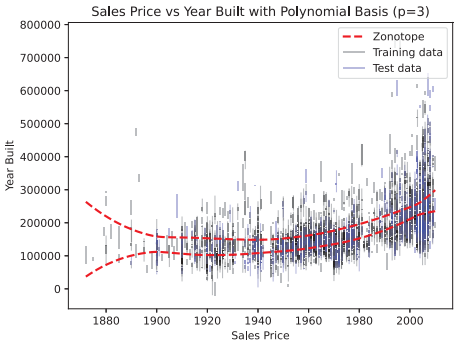


Fig. 4. An example of nonlinear imprecise regression using a polynomial basis function with the zonotope bands (dashed-line) on the Ames Housing Dataset. Black intervals represent training data, Blue intervals represent testing data.

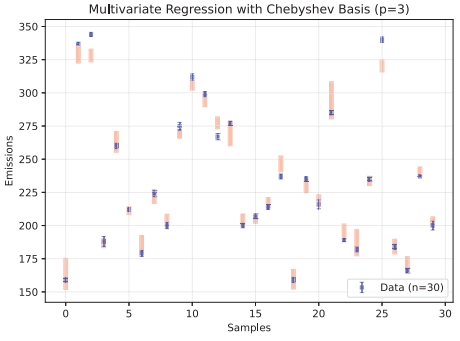


Fig. 5. Multivariate Regression using the Chebyshev basis function, with the zonotope intervals (red-bars) using the first 30 samples from the Testing set of the Fuel Consumption 2002-2022 dataset.

5. Conclusions

We have presented a zonotopic representation of the problem of least-square regression with imprecise response variables, introducing an exact and very efficient method for determining the expectation bands in both univariate and multivariate settings. The method is exact because yields the

same bands as the exact combinatorial method of vertices and it is efficient because it requires only a few row-column products. This zonotopic representation opens up exciting extensions. For instance, the integration within neural network models offers enhancing performance, particularly in nonlinear multivariate contexts. Such extensions could further enable the application of this approach to more complex learning tasks, including time-series forecasting through the use of recurrent architectures. Further research may include extending the proposed methodology to account for uncertainty in the independent variables, thereby broadening its applicability.

Reproducibility

Source code available at: `https://github.com/MattCodes03/Zonotopic-representation-of-multi-variable-regression-with-interval-dependent-variables`

Acknowledgement

This work is funded through the Research Excellence Award of the University of Strathclyde.

References

Augustin, T., G. Walter, and F. P. A. Coolen (2014). *Statistical inference*, Chapter 7, pp. 135–189. John Wiley & Sons, Ltd.

Billard, L. and E. Diday (2000). Regression analysis for interval-valued data. In *Data analysis, classification, and related methods*, pp. 369–374. Springer.

Billard, L. and E. Diday (2002). Symbolic regression analysis. In *Classification, clustering, and data analysis: recent advances and applications*, pp. 281–288. Springer.

Cattaneo, M. E. and A. Wiencierz (2012). Likelihood-based imprecise regression. *International Journal of Approximate Reasoning 53*(8), 1137–1154.

De Cock, D. (2011). Ames, iowa: Alternative to the boston housing data as an end of semester regression project. *Journal of Statistics Education 19*(3).

Ferson, S., V. Kreinovich, J. Hajagos, W. L. Oberkampf, and L. Ginzburg (2007). Experimental uncertainty estimation and statistics for data having interval uncertainty. Technical report, Sandia National Laboratories (SNL), Albuquerque, NM, and Livermore, CA.

Hastie, T., R. Tibshirani, J. Friedman, et al. (2009). The elements of statistical learning.

Neto, E. d. A. L. and F. d. A. T. De Carvalho (2008). Centre and range method for fitting a linear regression model to symbolic interval data. *Computational Statistics & Data Analysis 52*(3), 1500–1515.

Neto, E. d. A. L. and F. d. A. T. De Carvalho (2017). Nonlinear regression applied to interval-valued data. *Pattern Analysis and Applications 20*(3), 809–824.

Neumaier, A. (1990). *Interval methods for systems of equations*. Number 37. Cambridge university press.

Reckhow, K. H. (1994). Importance of scientific uncertainty in decision making. *Environmental Management 18*(2), 161–166.

Sadeghi, J., M. de Angelis, and E. Patelli (2019, October). Efficient training of interval Neural Networks for imprecise training data. *Neural Networks 118*, 338–351.

Schollmeyer, G. and T. Augustin (2015). Statistical modeling under partial identification: Distinguishing three types of identification regions in regression analysis with interval data. *International Journal of Approximate Reasoning 56*, 224–248.

Tretiak, K., G. Schollmeyer, and S. Ferson (2023, April). Neural network model for imprecise regression with interval dependent variables. *Neural Networks 161*, 550–564.

Černý, M., J. Antoch, and M. Hladík (2013, September). On the possibilistic approach to linear regression models involving uncertain, indeterminate or interval data. *Information Sciences 244*, 26–47.