

# RPNN: Structural modeling robust to outliers in input and output variables

Vladyslav Shaposhnyk<sup>1,2</sup>, Alessandro E.P. Villa<sup>2,4</sup> Tetyana Aksenova<sup>1,3</sup>

<sup>1</sup>*Institute for Applied System Analysis, State Technical University “Kyivskyy Politechnichnyy Instytut”, Kyiv, Ukraine*

<sup>2</sup>*Neuroheuristic Research Group, Grenoble Institute of Neurosciences, Université Joseph Fourier, Grenoble, France*

<sup>3</sup>*RTRA, Foundation “Nanoscience at the limits of Nanoelectronics”, Grenoble, France*

<sup>4</sup>*Neuroheuristic Research Group, Information Science Institute, University of Lausanne, Switzerland*

vshaposhnyk@ujf-grenoble.fr, tetiana.aksenova@cea.fr, avilla@neuroheuristic.org

**Abstract.** *The robust regression analysis works on data affected by deviations from a general assumption of normality. There are number of stable and robust methods in the field of linear regression analysis. In contrast the robust structural modeling is still under active development.*

*This paper describes a novel algorithm designed to solve a task of optimal polynomial model selection on multivariate data sets in presence of outliers in both input and output variables. On one side it is based on GMDH-type Polynomial Neural Network (PNN), which gives an universal model structure identification thanks to the evolving adaptively synthesized bounded network. From the other side the algorithm is based on application of MM-estimator, which allows achieving robustness to outliers in both input and output data sets. Previous version of Robust PNN was addressed to the modeling of the data with outliers in output variables only.*

*Enhanced RPNN was developed and tested on the artificial data set resulted from the simulation polynomials up to third degree. The Gaussian noise as well as outliers was added to the data. RPNN demonstrated robustness to outliers in both input and output variables (20% of outliers) and good accuracy of the automatic structure syntheses as well as of the parameters estimation.*

## Keywords

polynomial neural network, robust regression, non-linear regression, gm-estimators, structure selection

## 1 Introduction

Consider the non-linear regression model:  $y_i = f(\mathbf{x}; \beta_{\mathbf{o}}) + \varepsilon_i; i = 1, \dots, n$ , where  $f(\cdot)$  is a non-linear model function,  $\mathbf{x} = \{x_1, \dots, x_m\} \in \mathbb{R}^m$  is a vector of explanatory variable,  $\beta_{\mathbf{o}} \in \mathbb{R}^p$  is a vector of model parameters,  $y$  is a dependent variable, and  $\varepsilon$  is an error term. As it is known, outliers in the explanatory and/or dependent variables can have great impact on model selection  $f(\cdot)$  and on model parameters  $\beta_{\mathbf{o}}$  estimation resulting in totally wrong results if one applies classical statistical methods. In order to overcome those negative effects of outliers in model parameters estimation number of well developed outlier robust estimators with high break-down point [5, 4] exists. Least median squares [9], S-estimators [8], MM-estimators [10] are most known ones, but most of them consider case with linear model. On the other hand, wide set of GMDH-derived [6] algorithms exist aimed for non-linear structure and model selection task, having their core based on linear regression approaches.

In this paper, we present novel GMDH-type Polynomial Neural Network (PNN) algorithm, which gives an universal model structure and parameters identification and remain robust to outliers in explanatory and dependent variables with high break-down point.

## 2 Methods

Given  $\mathbf{X}$  is  $n \times m$  data-set of explanatory variables ( $n$  experiments with  $\mathbf{x} \in \mathfrak{R}^m$ ), and  $\mathbf{y}$  is  $n \times 1$  vector of dependent variable realizations. One would like to find a functional relation  $f(\cdot)$  between  $\mathbf{X}$  and  $\mathbf{y}$  within a family of multinomial functions of argument  $\mathbf{x}$  and parameters  $\beta$  not higher than  $p_{max}$  order and consists not more than  $t_{max}$  terms, such that:  $y_i \approx f(\hat{\mathbf{x}}; \hat{\beta}) + \hat{\varepsilon}_i$ . Taking into account possible presence of outliers in explanatory variable  $\mathbf{x}$  and/or in dependent variable  $\mathbf{y}$ . The core of proposed algorithm is based on PNN algorithm for model and parameters selection, described in details in [2], and outlier resistant improvement based on techniques used in MM-estimators [10].

Classical M-estimator, proposed by Huber is based on minimization of following criteria:  $\min_{\beta} \sum_{i=1}^n \rho(r_i/\hat{\sigma})$ , where residuals  $r_i = y_i - f(\mathbf{x}, \beta)$ , and  $\rho(\cdot)$  is symmetric residuals weight function with single minimum at zero. Which makes it robust to outliers in dependant variable  $\mathbf{y}$ , but it is still not robust to outliers in explanatory variable  $\mathbf{x}$ . In MM-estimators it is proposed to introduce correction function, which depends on explanatory variable estimation:

$$\min_{\beta} \sum_{i=1}^n w(\mathbf{x}; \hat{\mu}; \hat{\sigma}_x) \rho(r_i/\hat{\sigma}) \quad (1)$$

where  $w$  additional penalty function for those points which are considered as outliers in explanatory variable. The problem 1 is solved in iterative way and in general case  $w(\cdot)$  should be recalculated at each iteration, but we decided to sacrifice the algorithm's precision in favor of calculation speed and thus we compute correction weights  $\mathbf{w}_x$  only once and after we use it in all runs of iterative re-weighted least squares (IWLS) for linear regression. The robust location  $\hat{\mu}_x$  and scale  $\hat{S}$  (pair-wise algorithm was used in sake of speed) of data-set  $\mathbf{X}$  are estimated, distance measure is calculated (which is Mahalanobis distance), and then distances are converted to weights for IWLS in accordance with selected break-down point. Final equation for  $\mathbf{w}_x$  is as follows:  $w_i = \min\{1, \sqrt{\chi_{m-1}^{-2}(\tau)/d_m^2(\mathbf{x})}\}$  for  $\forall i = \{1, \dots, n\}$ , where  $\tau$  - selected outlier quantity threshold,  $\chi_{m-1}^{-2}(\tau)$  inverse of cumulative density function for  $\chi^2$  distribution with  $m - 1$  degrees of freedom,  $d_m^2(\mathbf{x}) = (\mathbf{x} - \hat{\mu}_x)^T \hat{S}^{-1}(\mathbf{x} - \hat{\mu}_x)$  Mahalanobis distance.

We used modified robust Akaike criteria [3, 7, 1] with additional leverage resistant term in it for model selection:

$$AICr = \frac{1}{n-1} \sum_{i=1}^n \rho(r_i \cdot w_x^2/\hat{\sigma}) + \frac{n+k}{n-k-2} = \frac{1}{n-1} RSSw + \frac{n+k}{n-k-2}. \quad (2)$$

where  $k$  is number of terms in the model. Then we use RSSw, as fit criteria, in IWLS.

The general scheme of algorithm is as follows:

1. Compute robust estimation of location  $\hat{\mu}$  and scale  $\hat{S}$  of explanatory data-set  $\mathbf{X}$ ;
2. Compute weights vector  $\mathbf{w}_x$  basing on  $\hat{\mu}$  and  $\hat{S}$  for later use in iterative MM-estimator;
3. Initialize best models set  $M'_{best} = \emptyset$ ;
4.  $\forall (t, p)$  where  $t \in \{1, \dots, t_{max}\}$  and  $p = p_{max}$ 
  - 4.1. run PNN algorithm having its starting models set  $M'_{start}$  equal to  $M'_{best}$  and with constraints on number of model terms and power equal to  $t$  and  $p$  accordingly;
  - 4.2. Obtain final  $M_{best}$  from PNN algorithm, to which it converged;
  - 4.3. Evaluate models in accordance with AICr criteria, as in equation 2;
  - 4.4. Update set  $M'_{best}$  with those models  $M \in M_{best}$ , which have lower values of AICr criteria and is different by structure from current models in  $M'_{best}$ ;
5. Take model  $M \in M'_{best}$  with minimal value of AICr criteria, which is the best fit model.

As one can see, the algorithm above is aimed for optimal model selection, while raw model search and parameters selection is done by mentioned in step 4.1. PNN algorithm. Enhanced PNN algorithm searches for best fit models for given pair  $(t, p)$  of constraints on model terms and power respectively, correction weight vector  $w_x$ , and list of starting models  $M'_{start}$  in accordance with next steps:

1. Initialize working sets: best models  $M_{best} = M'_{start}$ ; estimation of best models  $\mathbf{X}_{best}$  set in accordance with  $M_{best}$ ; expand working data-set  $\mathbf{X}_{all} = [\mathbf{X}; \mathbf{X}_{best}]$ ;
2.  $\forall \{i, j, k\}$  where  $i, j, k \in \{1, \dots, |\mathbf{X}_{all}|\}$ :
  - 2.1. In accordance with generator function  $G(i, j, k) = \mathbf{x}^i + \mathbf{x}^j \mathbf{x}^k$ , where by  $\mathbf{x}^p$  denoted vector corresponding to the  $p$ -th column of matrix  $\mathbf{X}$ , build a model  $M_{ijk} = \alpha_1 x^i + \alpha_2 x^j x^k$ ;

- 2.2. Reject model  $M_{ijk}$  if it does not fit to the constraints on terms and power  $(t, p)$ ;
- 2.3. Find linear regression coefficients vector  $\alpha = \{\alpha_1; \alpha_2\}$ , which is subject of robust *linear* regression task  $y \approx [\mathbf{x}^i; \mathbf{x}^j \cdot \mathbf{x}^k]^T \cdot \alpha$  and is done via iterative least squares method with use of weight vector  $\mathbf{w}_x$ ;
- 2.4. Calculate robust sum of residuals criteria  $RSSw(r) = \sum_{t=1}^n \rho(r_t(M_{ijk}) \cdot w_{xt}^2 / \hat{\sigma}) = \rho(\mathbf{w}_x([\mathbf{x}^i; \mathbf{x}^j \cdot \mathbf{x}^k]^T \cdot \alpha - \mathbf{y}) / \hat{\sigma})$  for model  $M_{ijk}$ ;
- 2.5. Reject model  $M_{ijk}$  if  $\forall M_i \in M_{best} : RSSw(r(M_{ijk})) \geq RSS(r(M_i))$ ;
- 2.6. Reject model  $M_{ijk}$  if  $\exists M_i \in M_{best} : M_{ijk}$  has same structure as  $M_i$  and  $RSSw(r(M_{ijk})) > RSSw(r(M_i))$ ;
- 2.7. Include model  $M_{ijk}$  into set  $M_{best}$  and model estimation  $\mathbf{x}_{ijk} = \alpha_1 \mathbf{x}^i + \alpha_2 \mathbf{x}^j \mathbf{x}^k = [\mathbf{x}^i; \mathbf{x}^j \cdot \mathbf{x}^k]^T \cdot \alpha$  into set  $\mathbf{X}_{best}$  (replacing model with same structure if exists);
3. Limit set  $M_{best}$ , and  $\mathbf{X}_{best}$  accordingly, by given number of models with lowest RSSw criteria;
4. Update model estimation set  $\mathbf{X}_{all} = [\mathbf{X}; \mathbf{X}_{best}]$ ;
5. Repeat steps 2-4 until models converge.

### 3 Results

The described algorithm was tested on artificial data-sets. Models of specified order and terms number were generated, as well as, appropriate data-sets.

An artificial data-set  $\mathbf{X}$  is created in accordance with Gaussianian distribution  $n(0, \sigma_x^2)$  in all tests below  $\sigma_x^2 = 10$ . An initial model  $M_{init}$  was generated and evaluated on data-set  $\mathbf{X}$ , in order to obtain dependent variable realizations  $\mathbf{y} = M_{init}(\mathbf{X})$ . An input data-set is forged as follows: a superposition of “real” data and outliers  $\mathbf{X}_{fit} = \mathbf{X} + n(0, 7\sigma_x^2)$  is made for input data-set and a superposition of “real” output, systematic error  $\varsigma$ , and outliers of  $\mathbf{y}_{fit} = \mathbf{y} + n(0, 1) + n(0, 3\sigma_y^2)$  for output variable is made. When  $M_{fit}$  is build by algorithm it’s tested on “clear” from outliers data-set  $\mathbf{X}_{test} \sim n(0, \sigma_x^2)$ , and average of square of residuals  $RS = 1/|\mathbf{X}_{test}| \cdot \sum_i (M_{fit}(\mathbf{x}_i) - M_{init}(\mathbf{x}_i))^2; \forall \mathbf{x}_i \in \mathbf{X}_{test}$  is recorded as performance measure of the fit model.

General performance of algorithm was estimated by comparison with basic PNN algorithm robust to outliers in dependent variables only [2]. Initial models were of 2-nd order and of 3 terms from 4 variables and a constant term. Thought, it could happen that some models did not use all available variables (for example:  $f(\mathbf{x}) = x_3^2 + x_4^2 + x_3 x_4 + 1$ , uses only 2 variables out of all). The algorithms were limited to search models up to second order and consisting from up to 6 terms. In all experiments 100 data-points with total of 25 outliers (15 in  $\mathbf{X}$  and 10 in  $\mathbf{y}$ ) were used and 200 receptions were made. The results are summarized in the Tab. 1.

Tab. 1: General performance / Prediction accuracy

Method used	RS, best 80%		RS, worst 20%		Models w. $RS > 10^3$
	mean	std	mean	std	
RPNN	0.035	0.038	128 433.338	291 869.810	8.5%
PNN	3 017 871.091	3 523 001.124	35 533 726.551	27 921 507.766	97.5%

To measure model selection features of the algorithm we increased maximum allowed length of model from 6 to 12. And run algorithm with in two versions: first as described above and second is with plain RSSw criteria instead of AICr. The results are summarized in the Tab. 2.

Tab. 2: Model selection performance

Criteria	RS, best 80%		RS, worst 20%		Models w. $RS > 10^3$	Terms found (average)
	mean	std	mean	std		
AICr	0.036	0.039	70 819.271	202 318.240	6.5%	5.745
RSSw	0.076	0.069	92 588.370	315 119.183	6.5%	10.610

Sensibility to outlier quantity was tested by running algorithm on models with 10 outliers in dependent and  $L = \{0, 5, 10, 15, 20, 25, 30\}$  outliers in explanatory variables. Most illustrative results are summarized in Tab. 3 and on Fig. 1.

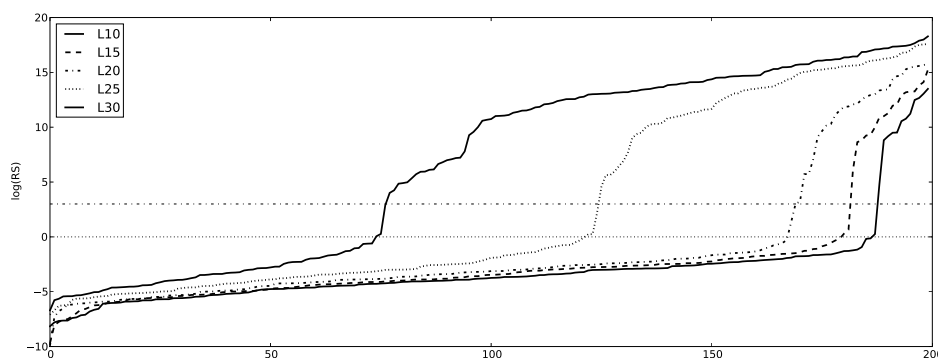


Fig. 1: Robustness to outlier quantity. Sorted residuals (RS) are displayed on log-scale. The algorithm was tuned to cope with up to 30% of outliers.

Tab. 3: Robustness to outlier quantity

Outliers Y+X	RS, best 80%		RS, worst 20%		Models w. $RS > 10^3$
	mean	std	mean	std	
O10+L0	0.020	0.021	0.194	0.239	0.5%
O10+L10	0.026	0.028	32 012.983	95 411.709	5.5%
O10+L20	0.051	0.059	972 748.470	1 816 260.544	13.5%
O10+L30	304 847.921	594 829.686	18 354 288.238	15 697 472.667	55.0%

## 4 Discussion

As it is visible from Fig. 1 fit models are quite accurate until certain threshold, but after it is a great “jump” of RS criteria to values of  $10^6$  and higher. This happens when the algorithm is failed to find one or more terms of actual model. When this is multiplied by the order of model it causes such high errors. In other cases fit model can have inaccurate parameters and/or excessive terms with coefficients close to zero, but, according to our experience, this would provoke errors up to  $10^3$ . After all proposed enhanced RPNN preserves good accuracy of the automatic structure synthesis of its predecessor and offers robustness to outliers in both explanatory and dependent variables.

## Acknowledgements

The authors acknowledge the support by the European Union FP6 grant #034632 (PERPLEXUS) and by the Grant ICOBI of Foundation “Nanoscience at the limits of Nanoelectronics”.

## References

- [1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716-723, 1974.
- [2] T. I. Aksenova, V. Volkovich, and A.E.P. Villa. Robust structural modeling and outlier detection with gmdh-type polynomial neural networks. *LNCS*, pages 881–886, 2005.
- [3] K.P. Burnham and D.R. Anderson. Multimodel inference: Understanding aic and bic in model selection. *Sociological Methods Research*, 33(2):261–304, November 2004.
- [4] D. L. Donoho and P. J. Huber. The notion of breakdown point. *Festschr. for Erich L. Lehmann*, pages 157–184, 1983.
- [5] F. R. Hampel. A general qualitative definition of robustness. *Ann. Math. Stat.*, 42:1887–1896, 1971.
- [6] A. G. Ivakhnenko. Polynomial theory of complex systems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-1(4):364–378, 1971.
- [7] E. Ronchetti. Robustness aspects of model choice. *Statistica Sinica*, 7:327–338, 1997.
- [8] P. Rousseeuw and V. Yohai. Robust regression by means of s-estimators. *Robust and nonlinear time series analysis*, SMC-1(26):256–272, 1983.
- [9] P.J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79:871–880, 1984.
- [10] V. Yohai. High breakdown-point and high efficiency robust estimates for regression. *Ann. Stat.*, 15:642–656, 1987.