

GUM2ALA – Uncertainty propagation algorithm for the Adaptive Linear Approximation according to the GUM

Tanja Dorst¹, Sascha Eichstädt², Tizian Schneider^{1,3}, Andreas Schütze^{1,3}

¹ ZeMA – Center for Mechatronics and Automation Technology gGmbH, Saarbruecken, Germany,

² Physikalisch-Technische Bundesanstalt, Braunschweig and Berlin, Germany,

³ Saarland University, Lab for Measurement Technology, Saarbruecken, Germany
E-mail (corresponding author): t.dorst@zema.de

Summary:

In machine learning, many feature extraction algorithms are available. To obtain reliable features from measured data, a propagation of measurement uncertainty is necessary for these algorithms. In this contribution, the Adaptive Linear Approximation (ALA) as one feature extraction algorithm is considered, and analytical formulas are developed for an uncertainty propagation in line with the Guide to the Expression of Uncertainty in Measurement (GUM). This extends the set of uncertainty-aware feature extraction methods, which already contains the discrete Fourier and Wavelet transform.

Keywords: measurement uncertainty, uncertainty propagation, feature extraction, adaptive linear approximation, machine learning

Motivation

The increasing demand for performance and efficiency in Industry 4.0 has led to the growing adoption of smart sensors, which allow data acquisition combined with internal signal processing, e.g., based on machine learning (ML). To analyze recorded data and develop data based evaluation models, a fully automated machine learning toolbox [1] can be used without knowledge of any physical process model and without expert knowledge. The best combination of five complementary methods for feature extraction (FE) and three for feature selection (FS) is calculated using a simple, but efficient classifier and k-fold cross validation for the training data set.

Whenever decisions are based on data, reliable data are important. Measurement uncertainties, sensor calibration and thus traceability to the SI units are the key principles in metrology. To be able to evaluate data quality and thus the quality of the ML results, it is necessary to consider the uncertainty information associated with the sensor data, ideally from calibration. However, calibrated sensors are seldomly used due to cost reasons or because removal and recalibration of the sensors may be difficult or even impossible. In these cases, information from the data sheets of the manufacturers could be used to obtain an indication of the data quality in form of a measurement uncertainty.

The available measurement uncertainty information associated with the raw sensor data then needs to be taken into account in the

subsequent data processing. However, the measurement uncertainty evaluation within the above-mentioned software toolbox has been neglected so far. To extend the toolbox for uncertainty analysis, the uncertainty associated with the raw sensor data must be propagated through the FE and FS methods. The reduction of the number of features is carried out by the FS methods. The classification afterwards is carried out in two steps: a further dimensionality reduction by Linear Discriminant Analysis (LDA) followed by the classification itself based on the Mahalanobis distance to the class centers. The selection and classification methods can use the measurement uncertainty information of the extracted features or methods of Bayesian statistics. It is worth noting that a Bayesian approach can also be implemented for FS and classification.

In this contribution, an algorithm for the uncertainty propagation in the Adaptive Linear Approximation (ALA) is developed in line with the *Guide to the Expression of Uncertainty in Measurement (GUM)* [2] and validated using a Monte-Carlo simulation corresponding to Supplement 2 to the GUM [3].

Results

ALA approximates a signal in linear segments of variable length. Mean and slope are extracted for each segment as features. The number of segments is calculated automatically, depending on the tradeoff between number of features and approximation error. The calculations for every

segment are the same, therefore they are shown here only for one segment.

Let $Y = (y_1, \dots, y_n) \in \mathbb{R}^{1 \times n}$ denote the real-valued time-domain values of a signal. The result of the ALA for Y is given by

$F = (\bar{y}_1, \dots, \bar{y}_{u_3+1}, b_1, \dots, b_{u_3+1}) \in \mathbb{R}^{1 \times 2(u_3+1)}$, where \bar{y}_k denotes the mean value and b_k the slope of the k -th segment, respectively. The index u_3 is the number of splits and therefore, $u_3 + 1$ the number of segments into which the signal is divided. The mean value and slope for the k -th segment are determined by

$$\bar{y}_k = f(y_i) = \frac{1}{v_{k+1} - v_k + 1} \sum_{i=v_k}^{v_{k+1}} y_i$$

and

$$b_k = h(y_i) = \frac{\sum_{i=v_k}^{v_{k+1}} (t_i - \bar{t}_k)(y_i - \bar{y}_k)}{\sum_{i=v_k}^{v_{k+1}} (t_i - \bar{t}_k)^2}.$$

For the propagation of uncertainties according to GUM, the sensitivities of the mapping $Y \mapsto F$ are calculated as

$$c_{k,j} = \frac{\partial \bar{y}_k}{\partial y_j} = \frac{1}{v_{k+1} - v_k + 1}$$

and

$$d_{k,j} = \frac{\partial b_k}{\partial y_j} = \frac{t_i - \bar{t}_k}{\sum_{i=v_k}^{v_{k+1}} (t_i - \bar{t}_k)^2},$$

for $j = v_k, \dots, v_{k+1}$, where v_k denotes the index of the input quantities at the beginning of a segment and v_{k+1} at the end, respectively. These coefficients can be stored in a sensitivity matrix

$$\mathbf{J}_{\bar{y},b}^m = \begin{pmatrix} \mathbf{C} \\ \mathbf{D} \end{pmatrix} \in \mathbb{R}^{2(u_3+1) \times n},$$

where $\mathbf{C} \in \mathbb{R}^{(u_3+1) \times n}$ denotes the upper submatrix with sensitivity coefficients for the mean values, and \mathbf{D} the lower submatrix with sensitivity coefficients for the slopes. The covariance matrix $\mathbf{U} \in \mathbb{R}^{n \times n}$ of the input quantities leads to the following expression for the covariance matrix $\mathbf{U}_F \in \mathbb{R}^{n \times n}$ associated with F :

$$\mathbf{U}_F = \mathbf{J}_{\bar{y},b}^m \cdot \mathbf{U} \cdot \mathbf{J}_{\bar{y},b}^{mT} = \begin{pmatrix} \mathbf{C} \mathbf{U}_y \mathbf{C}^T & \mathbf{C} \mathbf{U}_y \mathbf{D}^T \\ (\mathbf{C} \mathbf{U}_y \mathbf{D}^T)^T & \mathbf{D} \mathbf{U}_y \mathbf{D}^T \end{pmatrix}$$

The block structure of the covariance matrix \mathbf{U}_F can be used to deal with computer memory issues. Since \mathbf{U}_F is symmetric, only three blocks need to be stored, see also [4].

Fig. 1 shows the features for one signal and the validation of the features with a Monte Carlo simulation with 1.000.000 trials. The comparison of the GUM2ALA to the Monte Carlo results shows that the features and their associated uncertainties are the same for both approaches. However, the application of a Monte Carlo

simulation is much more time-consuming. Furthermore, a straightforward implementation of the Monte Carlo simulation is in most cases not feasible for standard computers due to the amount of computer memory required.

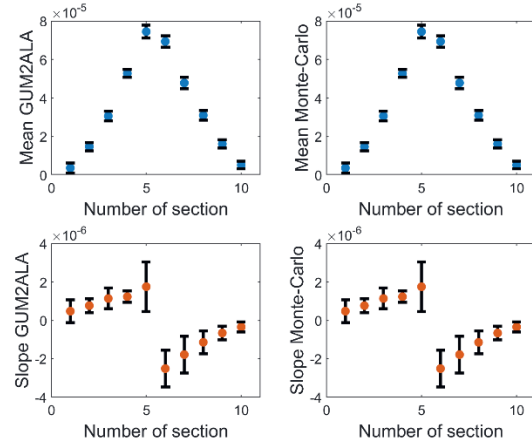


Fig. 1: Mean (top) and slope (bottom) as features for one cycle divided into 10 sections, calculated by the GUM2ALA algorithm (left) and validated by a Monte-Carlo simulation with 10^6 runs (right).

Outlook

After extending the approach to all FE methods in the toolbox, the feature uncertainty will be considered in the FS and classification steps for improving the stability and performance of the automated ML toolbox.

Acknowledgement

Part of this work has been developed within the Joint Research project 17IND12 Met4FoF of the European Metrology Programme for Innovation and Research (EMPIR). The EMPIR is jointly funded by the EMPIR participating countries within EURAMET and the European Union.

References

- [1] T. Schneider, N. Helwig, and A. Schütze, "Industrial condition monitoring with smart sensors using automated feature extraction and selection," *Meas. Sci. Technol.*, vol. 29, no. 9, 2018, doi: 10.1088/1361-6501/aad1d4.
- [2] Joint Committee for Guides in Metrology, "JCGM 100: Evaluation of measurement data — Guide to the expression of uncertainty in measurement," *JCGM100:2008*, 2008.
- [3] Joint Committee for Guides in Metrology, "JCGM 102: Evaluation of measurement data — Supplement 2 to the 'Guide to the expression of uncertainty in measurement' — Extension to any number of output quantities," *JCGM102:2011*, 2011.
- [4] S. Eichstädt and V. Wilkens, "GUM2DFT — a software tool for uncertainty evaluation of transient signals in the frequency domain," *Meas. Sci. Technol.*, vol. 27, no. 5, 2016, doi: 10.1088/0957-0233/27/5/055001.