JAAS

Higher Order Generalized Singular Value Decompositionand Dynamic Time Warping of Spectra from Femtosecond Laser-Induced Breakdown Spectroscopy in the 424.4 nm Spectral Region for Classification of Gaseous UF6 Assay

Journal:	*Journal of Analytical Atomic Spectrometry*
Manuscript ID	JA-ART-02-2023-000062.R1
Article Type:	Technical Note
Date Submitted by the Author:	03-May-2023
Complete List of Authors:	Jordan, Benjamin; University of Tennessee Knoxville College of Engineering, Thomason, Michael; University of Tennessee Knoxville College of Engineering Readle, Jason; ORNL, Cooper, Jeffrey; Centrus Technical Solutions Hayward, Jason; University of Tennessee Knoxville College of Engineering Hall, Howard; University of Tennessee Knoxville College of Engineering

Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Classification of Gaseous UF_6 Assay By Femtosecond LIBS in the 424.4 nm Spectral Region Using Numerical HOGSVD-DTW Features

Benjamin S. Jordan,*[a] Michael G. Thomason,[b] Jason D. Readle,[b] Jeffrey R. Cooper,[c] Jason P. Hayward,[b] Howard L. Hall[b]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

This Technical Note presents experimental results using numerical features of fs-LIBS data to classify the assay value of gaseous UF_6 material. The data-driven feature vectors are computed by Higher Order Generalized Singular Value Decomposition (HOGSVD) and Dynamic Time Warp (DTW). The method achieves 96.97% accuracy in spectral classification testing with fs-LIBS samples obtained from UF_6 material with five known assay values ranging from 0.287% to 61.740%, with 100% accuracy for the four largest assay values ranging from 4.615% to 61.740%.

1 Introduction

A recent paper[1] reports the computation of assay values in samples of fs-LIBS data for gaseous UF_6 material by applying multivariate nonlinear spectral fitting (MVNLSF)[2]. MVNLSF uses predetermined reference isotopic shifts and fitness of multiple Lorentzian distributions. This Technical Note describes a different computational approach, specifically, the training of a classification algorithm for the fs-LIBS data using data-driven, numerical feature vectors for the samples.

We define a numerical feature vector for the fs-LIBS samples in section 2.3. This feature vector is computed for each sample, and the data for a known assay value is partitioned into training and test datasets. A training algorithm fits the parameters of a classification model to the training data vectors, thereby creating a reference model for the assay value. In classification experiments, a candidate feature vector is compared with each separate class model and is assigned for the best model match.

The classification algorithms available today include discriminant analysis, nearest neighbors, neural networks, and support vector machines, among others[3]. Instead of restricting the study to a single algorithm, we use a software tool *Classification-Learner*[3] which evaluates a suite of state-of-the-art methods and reports the best performance for the given data. For the fs-LIBS datasets in this Note, the best performance is obtained with a support vector machine.

The following section 2.1 discusses the established statistical method of *bootstrapping*[4,5] to address the sparsity of fs-LIBS samples. Sections 2.2 and 2.3 discuss the numerical feature vector for fs-LIBS data. Section 3 describes the training and test datasets for the UF_6 material and presents the classification results.

2 Methodology

2.1 Addressing Sparsity in LIBS Data by Bootstrapping

Each digitized fs-LIBS sample LS is a 1-D vector with $n = 339$ datapoints spanning wavelengths from 424.25 nm to 424.5 nm. The sparsity of LIBS samples relative to the number of discrete wavelength datapoints (here $n = 339$) is a challenge in post-acquisition processing of data[6]. Since we are using higher-order singular value methods which require that data matrices have full rank, our approach to the sparsity problem is the statistical method of *bootstrapping*[4,5,7,8].

Bootstrapping is a technique in sampling to create additional representative samples from a given dataset in a statistically controlled way[4,5]. Bootstrapping is not new to LIBS processes[9], although there is no indication that it has been used with LIBS spectra for the computations in this paper.

Each digitized LIBS sample LS is a 1×339 vector of datapoints ordered left-to-right by increasing wavelength. The bootstrapping of LS has a requirement similar to the bootstrapping of time-series data sequenced by increasing time stamp[7,8], namely, the ordering of the original datapoints in the sample spectrum LS must be retained for the increasing wavelengths. A *bootstrap sample* of a LIBS spectrum LS is a sequence $BLS = d_1 d_2 \cdots d_{339}$ of $n = 339$ datapoints generated probabilistically by

- stepping the datapoint index i from 1 to $n-1$
- for each i, keeping the datapoint value d_i from LS unchanged

* Corresponding author.
[a] University of Tennessee, Knoxville, TN 37996, USA. Email: bjordan1@vols.utk.edu
[b] University of Tennessee, Knoxville, TN 37996, USA.
[c] Centrus Technical Solutions, 400 Centrifuge Way, Oak Ridge, TN 37830, USA.

with probability 0.5 or replacing d_i by d_{i+1} from *LS* with probability 0.5.

Digital random number generation is used to create bootstrap samples for a given LIBS spectrum *LS*. The probability of regenerating *LS* unchanged or of repeating the same bootstrapped sample underflows to 0. These bootstrap samples can be said to represent small, random jitter in positions locally in sampling the datapoints in *LS* while maintaining the association with increasing wavelength. On average, a third of the datapoint pairs in *LS* are unchanged, a third are duplicated, and a third are skipped. Each bootstrap sample *BLS* preserves the sample length n, has one expected occurrence of each individual datapoint in *LS* (other than the endpoints), and preserves the left-to-right ordering of the datapoints that occur.

The practice in this research has been to have 6 times as many bootstrap samples as n, i.e., 6 times as many rows as columns in a bootstrapped data matrix for a given LIBS spectrum *LS* with $n = 339$ columns.

2.2 SVD extended to HOGSVD

Numerical techniques used in the analysis of LIBS-type spectral data include Partial Least Squares (PLS) regression, Principal Component Analysis (PCA), and Singular Value Decomposition (SVD) [10–13]. This Note uses a version of SVD [14] generalized to multiple matrices. For reference, recall that the SVD of a $m \times n$ matrix A, $m \geq n$, is its numerical factorization into three matrices

$$A = USV'$$

where the columns of $m \times n$ matrix U are the left singular vectors, the diagonal entry s_j in $n \times n$ diagonal matrix S is the jth singular value, and the columns of $n \times n$ matrix V are the right singular vectors. The inverse of V in SVD is $V^{-1} = V'$. The *energy of A oriented in the column-vector direction* $(V(j,:)^{-1})'$ is

$$E_A((V(j,:)^{-1})') = ||AV(:,j)||^2 = s_j^2 > 0,$$

the squared singular value. The projection direction computed by SVD for the largest singular value is optimal in the sense that no other direction achieves a greater oriented energy for matrix A [15].

Higher Order Generalized SVD (HOGSVD) [16–18] is akin to SVD but extended to multiple matrices which are full rank and have the same number of columns (they can have different numbers of rows). For our purposes, HOGSVD is a simultaneous singular value-based decomposition of multiple matrices which can reveal the numerical similarities (quantified in oriented energy ratios) that all the matrices have in common as well as pairwise dissimilarities that distinguish one matrix from another.

Let $A_1, A_2, \ldots, A_N$ for $N > 1$ be N matrices of respective sizes $m_i \times n$, $m_i \geq n$, for $i = 1, 2, ..., N$ with all matrices being full rank n. HOGSVD is the simultaneous decomposition of these N matrices in the form

$$\begin{aligned} A_1 &= U_1 S_1 X' \\ A_2 &= U_2 S_2 X' \\ &\vdots \\ A_N &= U_N S_N X' \end{aligned}$$

where for matrix A_i the columns of $m_i \times n$ matrix U_i are the left singular vectors and the diagonal entry s_{ik} in $n \times n$ diagonal matrix S_i is the kth singular value. The same $n \times n$ matrix X occurs in the decomposition of every matrix, and its columns are the right singular vectors for all N matrices. The matrix X has a well-defined inverse X^{-1} and the projection direction vectors for oriented energies are the column vectors of $(X^{-1})'$. The row-k diagonal entry s_{ik} in the diagonal matrix S_i is the singular value of matrix A_i in the direction $(X(k,:)^{-1})'$ and the oriented energy of A_i in that direction is s_{ik}^2. The *generalized singular value* for matrix A_i over matrix A_j in that direction is the ratio

$$gs_{ijk} = \frac{s_{ik}}{s_{jk}}$$

and the corresponding oriented energy ratio is gs_{ijk}^2.

An individual generalized singular value is computed for a specific pair of matrices; however, information about all N matrices simultaneously is available in the form of n eigenvalues of a secondary matrix computed in the HOGSVD algorithm [17,18]. These eigenvalues $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ are real numbers (not complex), equal to or greater than 1, and related to the generalized singular values as follows:

- $\lambda_k = 1$ if and only if gs_{ijk} is 1 for all pairs of matrices A_i, A_j
- $\lambda_k > 1$ if and only if gs_{ijk} is not 1 for at least one pair of matrices A_i, A_j.

The value $\lambda_k = 1$ signifies that all N matrices have the *same oriented energy* (in a certain projection direction), whereas increasing values of $\lambda_k > 1$ signify an *increasing dissimilarity* among pairs of matrices with respect to oriented energies.

Since an eigenvalue λ_k can be very large, it is convenient to look instead at $1/\lambda_k$ which is in the bounded range 0 to 1. The distribution of the n values in $1/\Lambda$ is ensemble information about similarity of the N matrices with respect to their oriented energy ratios. In comparative terms, the more skewed the distribution of $1/\Lambda$ is towards 1, the more similar are the N matrices in their oriented energy ratios; the more skewed $1/\Lambda$ is towards 0, the less similar are the N matrices. Statistical characteristics of these distributions are used as numerical features for comparing one set of matrices with another.

2.3 HOGSVD-DTW Features for fs-LIBS Spectra

The feature vector of each fs-LIBS spectrum *LS* consists of (i) the error in an optimal alignment computed by Dynamic Time Warp (DTW) and (ii) statistical measurements of the distribution of $1/\Lambda$ for the n eigenvalues $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ computed by HOGSVD.

DTW[3,19] computes an alignment of two 1-D vectors optimally to minimize an error measure and has been used in diverse applications in spectroscopy[20,21]. In this Note, the DTW error is one numerical feature—essentially a baseline comparison of a candidate spectrum with a training data reference for a known assay value.

The remaining features are statistical information about the distribution of $1/\Lambda$ for the n HOGSVD eigenvalues Λ discussed in section 2.2. Specifically:

1. The *median*[22] partitions the samples ordered by magnitude into two equal subsets and is an informative measure of central tendency in asymmetrical distributions.

2. The *skewness*[22] is a measure of the asymmetry of the dataset values around the sample mean. The numerical value of skewness (or the *skewness coefficient*) quantifies the asymmetry in terms of the sample mean, median, and standard deviation.

   If skewness is negative, the values of $1/\Lambda$ spread out more to the left of the mean (towards value 0) than to the right. If skewness is positive, the data spreads out more to the right (towards value 1).

The HOGSVD-based values that complete the numerical features are the skewness and median of $1/\Lambda$.

3 Results and Discussion for UF_6 Material

3.1 fs-LIBS Training and Test Data

The fs-LIBS datasets are obtained for UF_6 material with five known assay values: 0.287%, 4.615%, 18.525%, 38.103%, 61.740%. These values are the five disjoint classes of spectra.

Five datasets of 99 LIBS samples each are used for these five disjoint classes. Each of the class datasets is partitioned randomly into training data (66 samples) and test data (33 samples). Each 1×339 LIBS sample LS within a class is represented by its 2035×339 matrix consisting of LS itself and its bootstrapped samples (see section 2.1).

The total LIBS data used for training is 66 samples for each class. In the training phase for class i, 33 samples are used as a class reference set R_i and the additional 33 samples $C_i(j)$ for $j = 1,2,...,33$ are used to obtain 33 representative feature vectors for within-class and across-class variations. The 33 training feature vectors for class i are obtained as follows. For each of the 33 samples $C_i(j)$ of class i and for each class k, the subvector FV_k is computed:

1. The DTW error is computed between the 1×339 sample mean vector of R_k and the 1×339 sample mean vector of $C_i(j)$.

2. The statistical feature values are computed for the eigenvalues of HOGSVD$(R_k, C_i(j))$.

The 1×15 feature vector of class i processed by *Classification-Learner* is the concatenation of these five training class-subvectors $[\mathrm{FV}_1\ \mathrm{FV}_2 \ldots\ \mathrm{FV}_5]$. These 33 feature vectors constitute the training dataset for class i for the numerical classification algorithms.

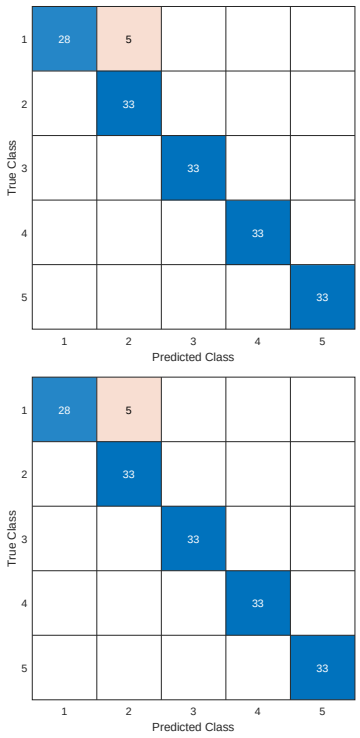

Fig. 1 Confusion Matrices for LIBS Datasets For HOGSVD-DTW 1×15 Feature Vectors. Top is Training Data (165 samples); Bottom is Test Data (165 samples).

The validation option for training is 5-fold cross-validation which protects against overfitting the class models to the training data[3]. A support vector machine has the best classification performance for these fs-LIBS datasets.

The training phase for each of the five classes yields a classification model for the class. In the subsequent classification experiments, a feature vector is computed for a candidate spectrum, say spectrum LS_T, and LS_T is assigned to the best-matching class model. The feature vector of the candidate LS_T is obtained as follows. The subvector is computed for each class k:

1. The DTW error is computed between the sample mean vector of R_k and the sample mean vector of LS_T.

2. The statistical feature values are computed for the eigenvalues of HOGSVD(R_k, LS_T).

The 1×15 feature vector of candidate LS_T is the concatenation of these five class-subvectors.

3.2 Classification Results: The Confusion Matrices

The classification results for the support vector machine are summarized in the 5×5 classification confusion matrices in which diagonal counts are correct classifications and off-diagonal are misclassifications. Fig. 1 shows the confusion matrices for the five assay classes labelled 1-5 respectively from the lowest 0.287% to the highest 61.740%. The top matrix is the training data itself (165 samples) classified by the models built from it. The bottom matrix is the separate test data (165 samples) classified by the models.

Observations about Fig. 1:

1. The classification rates for the five classes are 96.97% for training data and 96.97% for test data.

2. The classification rate is 100% in both confusion matrices for the four largest assay values 4.615%, 18.525%, 38.103%, 61.740%.

3. All misclassifications are the lowest assay value 0.287% misclassified as the next higher 4.615%. There are no misclassifications for any other values in these samples.

   The differences between the spectra collected for the two lowest assay samples (0.287% and 4.615%) are more subtle than between the higher assay samples. The spectral emission lines associated with ^{235}U are not prominent due to the low assay. Further, the 0.287% and 4.615% samples are closer in ^{235}U content than any of the other samples. Hence, a misclassification is more likely to occur between these two samples than others in this research.

4 Conclusions

This Note presents a data-driven classification method with good experimental performance on fs-LIBS spectra obtained from UF_6 with five known assay values ranging from 0.287% to 61.740%. Using the 1×15 HOGSVD-DTW feature vector and a support vector machine classification model, the five-class classification rate is close to 97% and is 100% for the four highest assay values in these datasets.

This research is relevant to future work using fs-LIBS data in the estimation of the assay value of a candidate sample that falls between the trained class values. In the ongoing development of this estimation algorithm, the known classes serve as fixed-reference landmarks and are incorporated into an expanded HOGSVD-based computation.

Author Contributions

BSJ: Conceptualization, Methodology, Investigation, Writing-Original, Visualization. MGT: Software, Methodology, Writing – Review and Editing. JDR: Conceptualization, Validation, Writing – Review and Editing. JRC: Conceptualization. JPH: Resources, Writing – Review and Editing, Supervision, Project Administration, Funding Acquisition. HLH: Resources, Project Administration, Funding Acquisition.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

Special thanks are owed to both Dr. Leigh Martin and Dr. Glenn Fugate of Oak Ridge National Laboratory for their assistance in sourcing the UF_6 sample material used in this research.

Nuclear Science and Security Consortium

This material is based upon work supported by the Department of Energy National Nuclear Security Administration through the Nuclear Science and Security Consortium under Award Number DE-NA0003180. This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or limited, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

References

1 B. S. Jordan, J. D. Readle, J. R. Cooper, M. Cook, J. P. Hayward, J. D. Hartman, D. Forrest and H. L. Hall, *Ann. of Nuc. Energy*, 2023, **181**, 109523.

2 J. J. Song, G. C. Y. Chan, X. L. Mao, J. D. Woodward, R. W. Smithwick, T. G. Schaaff, A. C. Stowe, C. D. Harris, R. E. Zheng, V. Zorba and R. E. Russo, *Spectrochim. Acta Part B-Atom. Spect.*, 2018, **150**, 67–76.

3 `https://www.mathworks.com`.

4 B. Efron, *The Ann. of Statis.*, 1979, **7**, 1–26.

5 B. Efron, *The Jackknife, the Bootstrap, and Other Resampling Plans Vol. 38*, Society of Industrial and Applied Mathematics CBMS-NSF Monographs, 1982.

6 E. Kepes, J. Vrabel, P. Porizka and J. Kaiser, *J. Anal. At. Spectrom.*, 2021, **36**, 1410–1421.

7 H. Wolfgang, J. Horowitz and J.-P. Kreiss, *Bootstrap methods for time series (paper 373)*, Coll. Res. Cntr., Humboldt Univ. Berlin, 2001.

8 P. Buhlman, *Stat. Sci.*, 2002, 52–72.

9 E. Képeš, P. Porízka and J. Kaiser, *J. Anal. At. Spectrom.*, 2019, **34**, 2411–2419.

10 X. L. Mao, G. C. Y. Chan, V. Zorba and R. E. Russo, *Spectrochim. Acta Part B-Atom. Spect.*, 2016, **122**, 75–84.

11 A. S. Mangsor, Z. H. Rizvi, K. Chaudhary and M. S. Aziz, *in International Laser Technology and Optics Symposium, Journal of Physics Conference Series*, Iop Publishing Ltd, Bristol, 2018.

12 H. Y. Kong, L. X. Sun, J. T. Hu, Y. Xin and Z. B. Cong, *Plasma Sci. & Tech.*, 2015, **17**, 964–970.

13 S. Romppanen, H. Hakkanen and S. Kaski, *Spectrochim. Acta Part B- Atom. Spect.*, 2017, **134**, 69–74.

14 G. H. Golub and C. F. VanLoan., *Matrix Computations (Fourth Ed.)*, Johns Hopkins Univ. Press, Baltimore MD, 2013.

15 B. DeMoor, J. Staar and J. Vandewalle, *in SVD and Signal Processing: Algorithms, Applications, and Architectures (Elsevier Sci. Pub.)*, 1988, 209–232.

16 L. DeLathauwer, B. DeMoor and J. Vandewalle, *SIAM J. Mat. Anal. and Appls.*, 2000, 1253–1278.

17 S. P. Ponnapalli, M. A. Saunders, C. F. VanLoan and O. Alter, *PLoS ONE*, 2011, **6**, e28072.

18 C. F. VanLoan, *Four Talks on Tensor Computations: Part 3. More Decompositions and Iterations,* SCAN Seminar, Cornell Univ., 2014.

19 C. Myers and L. Rabiner, *Bell Sys. Tech. J.*, 1981, **60**, 1389–1409.

20 H. J. Ramaker, E. N. M. van Sprang, J. A. Westerhuis and A. K. Smilde, *Anal. Chimica Acta*, 2003, **498**, 133–153.

21 C. M. Zou, H. M. Zhu, J. R. Shen, Y. He, J. E. Su, X. Q. Fan, H. M. Lu, Z. M. Zhang and Y. Chen, *Anal. Meth.*, 2019, **11**, 4481–4493.

22 J. Neter, W. Wasserman and G. A. Whitmore, *Applied Statistics (Second Ed.)*, Allyn and Bacon, Boston, 1978.