

PAPER • OPEN ACCESS

Autoregressive modelling of chromatographic signals from urine samples for prostate cancer diagnosis

To cite this article: A J Soto-Vergel *et al* 2021 *J. Phys.: Conf. Ser.* **1938** 012011

View the [article online](#) for updates and enhancements.

You may also like

- [Defluorination/Fluorination Reaction of BiF₃ Nanoparticles Prepared Via Sol-Gel Method for Fluoride Shuttle Battery](#)
Ken-ichi Okazaki, Hisao Kiuchi, Koji Nakanishi *et al.*
- [Resolving of Overlapping Asymmetrical Chromatographic Peaks by Using Wavelet-Transform and Gram-Charlier Peak Model](#)
A V Bochkarev
- [Light scattering detector based on light-emitting diodes-Solar cells for a flow analysis of Warfarin in pure form and pharmaceutical formulations](#)
Naghham S Turkey and Jalal N Jeber



The Electrochemical Society
Advancing solid state & electrochemical science & technology

243rd ECS Meeting with SOFC-XVIII

More than 50 symposia are available!

Present your research and accelerate science

Boston, MA • May 28 – June 2, 2023

[Learn more and submit!](#)

Autoregressive modelling of chromatographic signals from urine samples for prostate cancer diagnosis

A J Soto-Vergel¹, B Medina-Delgado¹, and W Palacios-Alvarado¹

¹ Universidad Francisco de Paula Santander, San José de Cúcuta, Colombia

E-mail: angelojosephsv@ufps.edu.co

Abstract. This article evaluates autoregressive modeling as a feature extraction method in a database of chromatographic signals from urine samples for non-invasive diagnostic support of prostate cancer in response to the research question: Can chromatographic signals from urine be characterized and used as a non-invasive method for cancer diagnosis? For this purpose, a database of 18 patients, 9 diagnosed with prostate cancer and 9 control patients, is consolidated, statistical methods are implemented to generate autoregressive coefficients from the data signals, and finally, the principal component analysis technique is applied for cross-class classification. As a result, a correct classification was obtained in the total number of samples validating the autoregressive modelling as a feature extraction method in contrast to the conventional methodology usually followed in chromatographic signal processing.

1. Introduction

Chromatography is considered a standard in chemical signal analysis and its popularity has allowed the development and application of algorithms that seek to classify, in some way, chromatographic signals regardless of the area of application [1,2]. These classification algorithms must have, at least, two main structural blocks: one in charge of performing the identification and extraction of features and another one that allows the recognition to discriminate between possible categories; thus, in the literature, several methods that fulfil the classification function in each chromatographic data set have been identified. Examples of the most used algorithms are principal component analysis and genetic algorithms that in combination with support vector machines allow delimiting representative regions of the data that provide the basis for classification [3,4].

However, although the results of the algorithms are quite accurate and manage to flag the different features between samples, they require a high level of pre-processing to work properly [5]. This preprocessing necessary for feature extraction usually requires normalization and dead-time removal, baseline correction and time-domain peak alignment prior to data processing, and when chromatographic signals are involved, these preprocessing steps are even more relevant [6]. Another limitation of these feature extraction algorithms is that a new learning sample requires recalculation of the entire set of principal components of the system and therefore pre-processing of the entire data set since peak alignment must consider each signal, all of which requires significant time due to the complexity of the data signals.

On the other hand, scientific evidence for the use of chromatograms and classification algorithms applied to the study and possible diagnosis of prostate cancer is scarce. Some research in this direction performs an analysis of the chemical components present in the chromatogram obtained from urinary exosomes but does not implement any type of computer analysis to support the extraction of features



that may not be evident since chromatography generates complex signals in its output [7]; likewise [6], describes robust preprocessing and classification based on signal energy analysis concluding significant differences between prostate cancer samples and control samples, however, the feature extraction method involves different processing methods for its success.

Therefore, it is important to explore other feature extraction methods that are concise, reliable and do not involve recalculating features based on the entire dataset for the analysis of this type of signals and their classification. One such method is autoregressive models that have proven useful in the processing of audio signals [8], signals from chemical sensors [9] and, in general, any signal considered as a time series [10]. However, no results have been reported in the literature on the application of autoregressive modelling to extract features from chromatographic signals from urine samples for characterization and classification, and that these results can be used in a possible non-invasive diagnosis of prostate cancer.

Now, in chromatography, varying conditions in sample collection and data generation lead to time shifts and peak misalignment, which is one of the challenges that has been most worked on in the processing of this type of signal [11,12]. Therefore, this work proposes autoregressive modelling of chromatogram data, which makes use of regression coefficients to represent a time series of data, whereby a significant advantage of this method is its independence of the time dimension of the original sample [13], providing a filter to the noise component and virtually compressing the data while preserving the important features of the signal and thus providing a feature extraction method that does not require the conventional pre-processing that usually ends, in the case of chromatographic signals, with the alignment of the signal peaks..

2. Methodology

Methodology consists of three stages: first, the consolidation of the database comprising the collection of urine samples from patients diagnosed with prostate cancer and control patients, as well as the chemical preparation of each sample; this is followed by the generation of the chromatogram because of the digitization of the samples in text files with the intensities in millivolts; each patient has an associated chromatogram, which is constructed iteratively until the appropriate resolution is obtained [14,15].

Second stage is autoregressive modelling which involves the selection of the model order and the generation of the model coefficients; the model order must be chosen in such a way that it preserves the trade-off between goodness-of-fit and a high data compression ratio, for this, use is made of the Akaike information criterion which is one of the most ubiquitous tools in statistical modelling [16]; its interpretation consists of evaluating models in a given range of orders by choosing the one that minimizes the information criterion. Akaike is denoted by AIC and is calculated by Equation (1), whose parts denote the trade-off between model complexity and goodness-of-fit, i.e., minimizing the variance of the process by characterizing the number of parameters to be estimated in the model.

$$AIC(k) = \log(\rho_k) + 2 \frac{k+1}{N}, \quad (1)$$

where: ρ_k is the variance of the noise of the order model k , k is the order of the model, N is the sample size of the signal.

For the generation of the model coefficients, the Yule-Walker method is implemented which returns a polynomial corresponding to the estimation of the parametric signal model of the input data using autocorrelation and minimizing the prediction error by least squares. Equation (2) represents the matrix form of the Yule-Walker method which is solved by the Levinson-Durbin recursion [17,18]; where the outputs consist of the coefficients, the estimated variance of the white noise process and the reflection coefficients. These results can be used to estimate the optimal order by using information criteria, which for this case is Akaike.

$$\begin{bmatrix} r_1 & r_2^* & \cdots & r_n^* \\ r_2 & r_1^* & \cdots & r_{n-1}^* \\ \vdots & \vdots & \ddots & \vdots \\ r_n & \cdots & r_2 & r_1 \end{bmatrix} \begin{bmatrix} a_2 \\ a_3 \\ \vdots \\ a_{n+1} \end{bmatrix} = \begin{bmatrix} -r_2 \\ -r_3 \\ \vdots \\ -r_{n+1} \end{bmatrix}, \quad (2)$$

where, r_n are the autocorrelation terms a_n are the scaling factors of the coefficients of the autoregressive model.

Finally, the third stage, classification, validates the autoregressive modelling method as a feature extractor of chromatographic signals by means of principal component analysis. The latter analysis is commonly used to validate whether the chosen features of a signal, in this case the coefficients of the autoregressive signal model, are correct for classification [19,20].

3. Results

3.1. Database consolidation

To carry out the experiments, we have a database of 18 chromatographic signals corresponding to urine samples, 9 from patients with prostate cancer and 9 from control patients, which were taken in the quality control laboratory of the University of Pamplona, Colombia; the number of samples corresponds to the unusual nature of this type of examination and the existence of the laboratory equipment. The signals are exported by the measuring equipment in text files in .txt format where the x-axis is the time in minutes and the y-axis is the intensity of the sample in millivolts. Figure 1 presents graphically the chromatogram of a prostate cancer patient and a control patient where some indications of their differences can be seen.

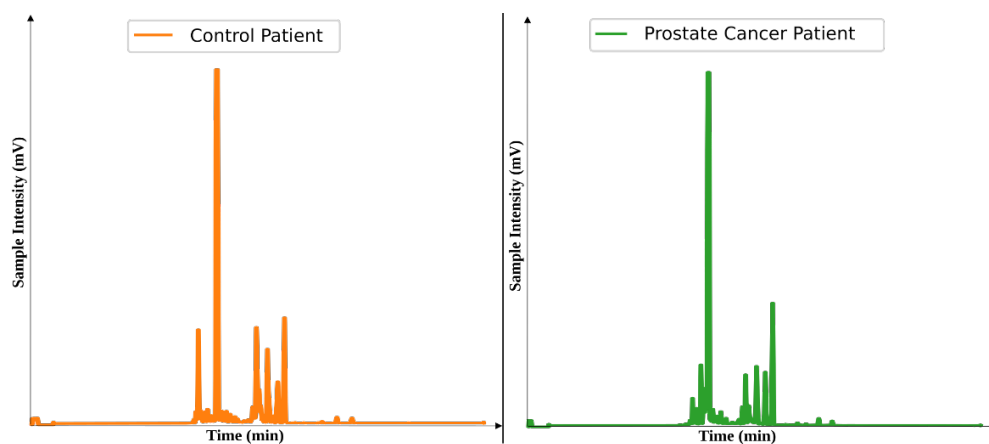


Figure 1. Representation of chromatographic signals.

3.2. Autoregressive modeling

Figure 2 shows the relationship between the information criterion and the order of the autoregressive model for both control and prostate cancer patient signals; this information criterion was calculated with the Akaike method using Equation (1) from which the order of the model is chosen in 20 to represent each sample. It is important to point out that Figure 2 is for a sample of each type of patient, and it is not considered relevant to present the results for the other samples since the curve is very similar in all other cases.

As for the feature matrices, these are formed with the coefficients of the autoregressive model because of applying the Yule-Walker method to each signal, these matrices are the ones that will be used as input for the classification analysis. Figure 3 shows an example of the power spectral density of a real chromatogram and one generated with the model, showing the shared similarity in their spectra being characteristic of the information contained in the data.

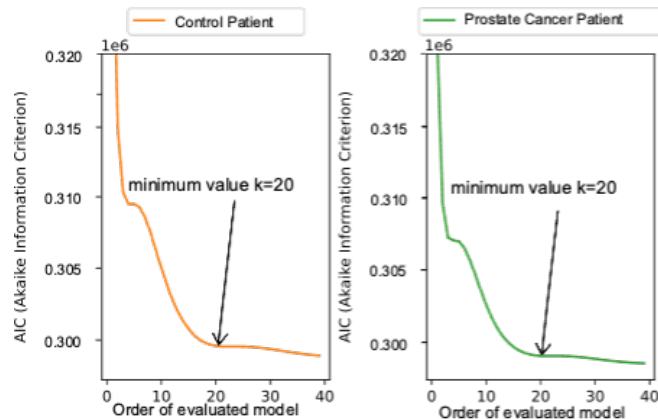


Figure 2. Akaike criterion as a function of the order of the autoregressive model.

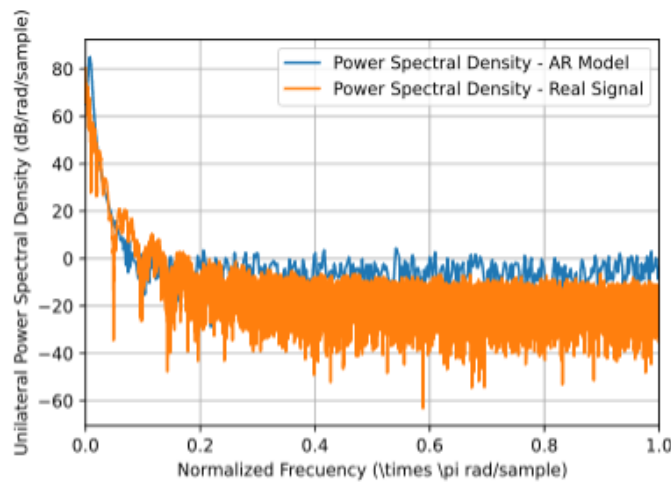


Figure 3. Power spectrum of the real and modeled signal.

3.3. Classification

Figure 4 shows the result of applying the principal component algorithm using a sedimentation plot whose eigenvalues were calculated from the coefficients of the autoregressive model, in which a single characteristic curve is evident, whose classification does not show overlap between the first two components; that is, principal components 1 and 2 are the ones that contain the most information of the characteristic chromatograms when applying the parametric Yule-Walker method.

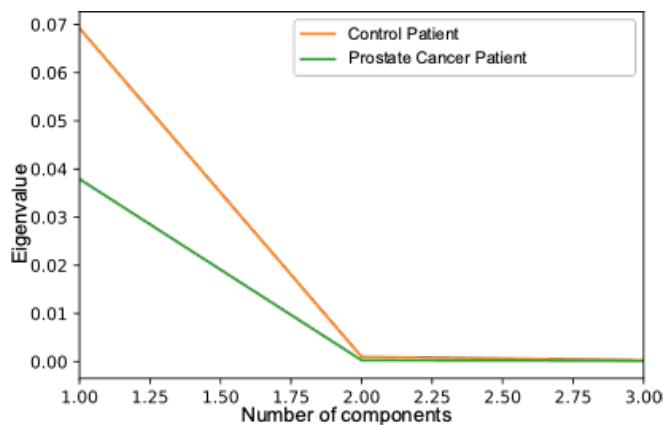


Figure 4. Principal component sedimentation plot - autoregressive modeling analysis.

Figure 5 shows the analysis of the data in a score plot, with principal components 1 and 2 as axes. This analysis shows that control and cancer patients have differences in their chromatograms and that the autoregressive modelling method is appropriate for feature extraction. Importantly, these results allow a clear classification between the groups which was not possible using conventional digital signal processing for this same dataset according to the results reported by [6].

Finally, a test of the time it takes to perform the conventional processing for this type of signals in comparison with the proposed methodology based on autoregressive models is carried out. As a result, the time normalization, dead time elimination, baseline correction, denoising and peak alignment techniques take an average of 56.1 seconds to run, while the autoregressive modelling takes 1.7 seconds to run for all samples. It is important to mention that Google Collaboratory has the following technical specifications: between 0.77 GB and 12 GB of RAM, 6 Intel Xeon 2.20 GHz distributed CPUs; in that context, the computational cost of the model is estimated taking as a metric the execution time comparing the proposed approach using the autoregressive model and the conventional approach that follows the time normalization, dead time removal, baseline correction, denoising and peak alignment that was reported by [6].

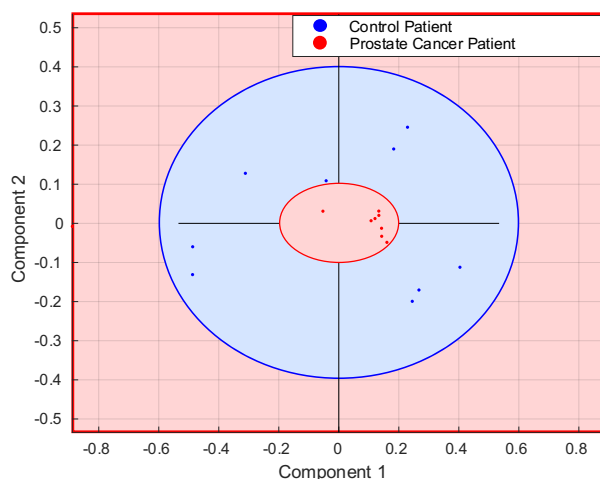


Figure 5. Score's chart - principal component correlation.

4. Conclusions

This paper implements autoregressive modeling as a reliable and computationally more efficient alternative for the feature extraction phase of chromatographic signals from the urine of prostate cancer patients and control patients compared to a system requiring conventional signal processing. For this purpose, it was first necessary to consolidate the database experimentally, followed by the implementation of the Yule-Walker method for parametric modeling and finally the use of a classification algorithm to validate the feature extraction using the coefficients of an autoregressive model.

The autoregressive modeling was validated as a suitable feature extractor for the chromatographic signals analyzed, presenting an average computation time of 1.7 seconds, which is 33 times faster than the conventional processing used for this type of analysis. The above is evidenced in the sedimentation graphs and scores that present a unique behavior of the principal components corresponding to control patients and prostate cancer patients, specifying that the use of the coefficients of an autoregressive model by means of the Yule-Walker method is a differentiating factor.

In the score plot, three classification regions can be discriminated by means of principal component analysis. This type of analysis is usually the input of methods that delimit these regions by means of different statistical techniques such as support vector machines or neural networks; however, the small number of samples does not allow a correct implementation of these methods and for this case where obtaining the data signals requires such an elaborate and costly process makes it even more difficult to apply them soon.

Finally, the results show that it is possible to apply autoregressive modeling as a pattern extraction technique to find characteristics that differentiate and accentuate the classification of the chromatograms of prostate cancer patients and control patients. Therefore, it can be affirmed that the methodology based on the autoregressive modeling theory to characterize chromatographic signals and perform their subsequent simulation or classification is valid and that these models were able to adequately represent the characteristics in real signal frequency.

References

- [1] Serrano N, Cetó X, Núñez O, Aragón M, Gámez A, Ariño C, Díaz J 2018 Characterization and classification of Spanish paprika (*Capsicum annum* L) by liquid chromatography coupled to electrochemical detection with screen-printed carbon-based nanomaterials electrodes *Talanta* **189(1)** 296
- [2] Merchak N, Rizk T, Silvestre V, Remaud G, Bejjani J, Akoka S 2018 Olive oil characterization and classification by ¹³C NMR with a polarization transfer technique: a comparison with gas chromatography and ¹H NMR *Food Chem.* **245(1)** 717
- [3] Tao Y, Huang S, Yan J, Cai B 2019 Determination of major components from *Radix Achyranthes bidentate* using ultra high-performance liquid chromatography with triple quadrupole tandem mass spectrometry and an evaluation of their anti-osteoporosis effect in vitro *J. Sep. Sci.* **42(13)** 2214
- [4] Zhang J, Zheng C, Xia Y, Wang B, Chen P 2017 Optimization enhanced genetic algorithm-support vector regression for the prediction of compound retention indices in gas chromatography *Neurocomputing* **240(1)** 183
- [5] Allen-Zhu Z, Li Y, Liang Y 2019 Learning and generalization in overparameterized neural networks, going beyond two layers *33rd Conference on Neural Information Processing Systems* (Canada: NeurIPS)
- [6] Soto-Vergel A, Mendoza L, Delgado B 2019 Analysis of energy and major components in chromatographic signals for the diagnosis of prostate cancer *Respuestas* **24(1)** 76
- [7] Yang J, Lee J, Byeon S, Rha K, Moon M 2017 Size dependent lipidomic analysis of urinary exosomes from patients with prostate cancer by flow field-flow fractionation and nanoflow liquid chromatography-tandem mass spectrometry *Analytical Chemistry* **89(4)** 2488
- [8] Rohrbach D, Mamou J 2018 Autoregressive signal processing applied to high-frequency acoustic microscopy of soft tissues *IEEE Transaction on Ultrasonic, Ferroelectric, Frequency Control* **65(11)** 2054
- [9] Büyükkabasakal K, Acikbas S, Deniz A, Acikbas Y, Capan R, Erdogan M 2019 Chemical sensor properties and mathematical modeling of graphene oxide langmuir-blodgett thin films *IEEE Sens. J.* **19(20)** 9097
- [10] Chen G, Gan M, Chen G 2018 Generalized exponential autoregressive models for nonlinear time series: stationarity, estimation, and applications *J. Inf. Sci.* **438(1)** 46
- [11] Kumar K 2018 Optimizing the process of reference selection for correlation optimised warping (COW) and interval correlation shifting (icoshift) analysis: automating the chromatographic alignment procedure *Anal. Methods* **10(2)** 190
- [12] Ottensmann M, Stoffel M, Nichols H, Hoffman J 2018 GCalignR: An R package for aligning gas-chromatography data for ecological and evolutionary studies *PloS one* **13(6)** 1
- [13] Wu H, Zhou Y, Yang C, Zhu H, Hao D, Ren S 2020 A method of prediction for transformer malfunction based on oil chromatography *5th International Conference on Automation, Control and Robotics Engineering* (China: IEEE) p 444
- [14] Patiño-Domínguez B 2016 *Determinación de Parámetros Operacionales Necesarios en el Empaquetado de Columnas de Cromatografía* (España: Universidad da Coruña)
- [15] Cazes J 2010 *Encyclopedia of Chromatography* (New York: CRC Press)
- [16] Cavanaugh J, Neath A 2019 The Akaike information criterion: background, derivation, properties, application, interpretation, and refinements *Wiley Interdiscip. Rev. Comput. Stat.* **11(3)** e1460
- [17] Gupta V, Mittal M 2019 R-Peak detection in ECG signal using Yule-Walker and principal component analysis *IETE Journal of Research* 1
- [18] Libal U, Johansson K 2019 Yule-Walker equations using higher order statistics for nonlinear autoregressive model *Signal Processing Symposium* (Polonia: IEEE)
- [19] Franses P, Janssens E 2019 Spurious principal components *Appl. Econ. Lett* **26(1)** 37
- [20] Li Z, et al. 2020 Optimization of SEMG classification model based on correlation analysis and feature Selection *5th International Conference on Advanced Robotics and Mechatronics* (China: IEEE) p 402