Journal of Pharmaceutical and Biomedical Analysis 152 (2018) 120–127

Contents lists available at ScienceDirect

Journal of Pharmaceutical and Biomedical Analysis

journal homepage: www.elsevier.com/locate/jpba

Wavenumber selection method to determine the concentration of cocaine and adulterants in cocaine samples

A. Kahmann[a], M.J. Anzanello[a,d,*], F.S. Fogliatto[a], M.C.A. Marcelo[b], M.F. Ferrão[b], R.S. Ortiz[c,d], K.C. Mariotti[c,d]

[a] *Departamento de Engenharia de Produção e Transportes – Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil*
[b] *Programa de Pós-Graduação em Química, Instituto de Química – Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil*
[c] *Superintendência da Polícia Federal, Porto Alegre/RS, Porto Alegre, RS, Brazil*
[d] *Instituto Nacional de Ciência e Tecnologia Forense (INCT Forense), Brazil*

ARTICLE INFO

Article history:
Received 9 October 2017
Received in revised form 23 January 2018
Accepted 24 January 2018

Keywords:
Wavenumber selection
Prediction
FTIR
Cocaine
Adulterants

ABSTRACT

Street cocaine is typically altered with several compounds that increase its harmful health-related side effects, most notably depression, convulsions, and severe damages to the cardiovascular system, lungs, and brain. Thus, determining the concentration of cocaine and adulterants in seized drug samples is important from both health and forensic perspectives. Although FTIR has been widely used to identify the fingerprint and concentration of chemical compounds, spectroscopy datasets are usually comprised of thousands of highly correlated wavenumbers which, when used as predictors in regression models, tend to undermine the predictive performance of multivariate techniques. In this paper, we propose an FTIR wavenumber selection method aimed at identifying FTIR spectra intervals that best predict the concentration of cocaine and adulterants (e.g. caffeine, phenacetin, levamisole, and lidocaine) in cocaine samples. For that matter, the Mutual Information measure is integrated into a Quadratic Programming problem with the objective of minimizing the probability of retaining redundant wavenumbers, while maximizing the relationship between retained wavenumbers and compounds' concentrations. Optimization outputs guide the order of inclusion of wavenumbers in a predictive model, using a forward-based wavenumber selection method. After the inclusion of each wavenumber, parameters of three alternative regression models are estimated, and each model's prediction error is assessed through the Mean Average Error (MAE) measure; the recommended subset of retained wavenumbers is the one that minimizes the prediction error with maximum parsimony. Using our propositions in a dataset of 115 cocaine samples we obtained a best prediction model with average MAE of 0.0502 while retaining only 2.29% of the original wavenumbers, increasing the predictive precision by 0.0359 when compared to a model using the complete set of wavenumbers as predictors.

1. Introduction

Cocaine is one of the most consumed drugs around the world. In spite of the powerful anesthetic properties of cocaine, its continuous and abusive intake may lead to depression and a large variety of harmful effects to the cardiovascular system, lungs and brain [1–3]. In addition, several adulterants are typically mixed with the drug, most notably caffeine (stimulant), phenacetin (analgesic), levamisole (anthelmintic), and lidocaine (local anesthetic) [2,4,5]. Thus, determining the concentration of cocaine and its adulterants in seized samples is not only valuable from a clinical perspective, but also provides relevant information to investigative forces towards interrupting drug trafficking [6].

Fourier Transformed Infrared (FTIR) spectroscopy is a fast and non-destructive technique that enables identification and quantification of several chemical compounds [7]. FTIR datasets are comprised of a large number of highly correlated and noisy variables known as wavenumbers. When used as input data in multivariate prediction techniques, such type of data tends to compromise the performance of predictive models. To overcome such drawback, wavenumber selection techniques have become key in reducing the influence of ill-conditioned FTIR data on multivariate techniques [8,9]. Some of such techniques have relied on slicing the spectra into equidistant intervals (e.g. Interval Partial Least Squares

---

* Corresponding author at: Departamento de Engenharia de Produção e Transportes – Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil.
E-mail address: anzanello@producao.ufrgs.br (M.J. Anzanello).

https://doi.org/10.1016/j.jpba.2018.01.050
0731-7085/

– iPLS) in order to find the most informative region [10–12], while others have integrated wavenumber importance indices to backward/forward feature selection [13–15]. The analysis and selection of the most relevant regions of the FTIR spectra generates models that are simpler and easier to interpret by highlighting not only relationships between wavenumbers, but of wavenumbers and the investigated property [7,16]. Finally, model complexity and computational cost is reduced through wavenumber selection [17].

This paper proposes a new method for wavenumber selection aimed at predicting the concentration of cocaine and adulterants in cocaine samples. We propose the use of Quadratic Programming (QP) to simultaneously minimize the probability of retaining redundant wavenumbers, and to maximize the relationship between retained wavenumbers and the response variable (compound's concentration). Optimization gives rise to an importance index that quantifies the predictive potential of each wavenumber. Using the index, wavenumbers are inserted into three regression techniques for predicting the concentration of cocaine and adulterants; namely: Multiple Linear Regression, Principal Components Regression, and Partial Least Squares Regression. The proposed framework was applied to a FTIR dataset consisting of 115 cocaine samples described by 662 wavenumbers, substantially reducing the number of wavenumbers to be inserted into predictive models.

2. Materials and method

2.1. Samples, sample preparation, and instrumentation

Fifty-eight samples of cocaine (crack, freebase, and salt cocaine) seized by the Brazilian Federal Police (BFP) in the state of Rio Grande do Sul between 2013 and 2015 were used in this study. In addition, 57 solid samples were prepared mixing cocaine and its adulterants. All samples were homogenized using an agate mortar prior to analysis; the cocaine standard was provided by the BFP. Lidocaine (Delaware, Brazil), levamisole (Sigma-Aldrich), caffeine (Acrosorganics 98.5%, NJ, USA) and phenacetin (Delaware, Brazil) were used as adulterants in the prepared mixtures, which had two to five components each.

FTIR was the analytical technique chosen in this study due to the several advantages it presents when compared to other classical techniques; e.g. GC/MS and LC–MS/MS. FTIR (i) does not require time-consuming procedures for sample preparation (e.g., samples assessed via GC and LC need to be liquid), avoiding the waste of laboratory materials for sample dilution; (ii) yields fast results when compared to other classical techniques; and (iii) requires equipment substantially cheaper than the ones needed by GC and LC. The main disadvantage of FTIR compared to other analytical techniques relies on its sensitivity and specificity, and the need for chemometrics tools to support the assessment of obtained data.

FTIR spectra were acquired through a Nicolet 380 FTIR Spectrometer (Nicolet Instrument Co., Madison, USA) equipped with DTGS (deuterated triglycine sulphate) detector and a smart orbit single reflection diamond ATR sampling accessory. Thirty-two scans were performed with resolution of 4 cm^{-1}. The fingerprint spectral region ranging from 550 to 1800 cm^{-1} was selected for the multivariate analysis; spectra data was not preprocessed.

A liquid chromatography system with diode array detector (Agilent Technologies, USA) was used as reference method for determining the concentration of cocaine and its adulterants in the seized samples. The separation was carried out in isocratic mode, C18 column (Zorbax EclipsePlus, 4.6° 250 mm; Agilent, USA) at 30 °C with the full UV–vis monitored, a bandwidth of 4 nm and 0.5 nm resolution. For the mobile phase a 1:1 solution of acetonitrile (Panreac, Spain) and water purified in a Milli-Q system (Millipore) was used, with amonium acetate (F. Maia, Brazil) as buffer (pH 8.3) and a flow of 1 mL min^{-1}. Compound concentrations were assessed in *mg*/*g* through external calibration. For more details on methodology and figures of merit, see [18].

2.2. Data analysis and optimization techniques

We now present the fundamentals of the data analysis and optimization techniques used in the proposed method; namely: Mutual Information, Quadratic Programming, Multiple Linear Regression, Principal Component Regression, and Partial Least Squares Regression.

Mutual Information quantifies the amount of information shared between two wavenumbers using a symmetric and non-negative dependency measure. The Mutual Information between wavenumbers W and Y is given by:

$$\mathrm{I}(W;Y)=\sum_{y\varepsilon Y}\sum_{w\varepsilon W}p(x,w)\log\left(\frac{p(w,y)}{p(w)p(y)}\right) \quad (1)$$

where $p(w{,}y)$ represents the joint probability distribution function of W and Y, and $p(w)$ and $p(y)$ are the marginal probability distribution functions of W and Y, respectively [19]. Distributions may be either estimated by the approximation of a prior determined density function or the density function discretization [20].

A Multiple Linear Regression (MLR) equation provides a mathematical description of the relationship between a dependent variable (e.g. compound concentration) and one or more independent variables (e.g. wavenumbers). By assumption, independent variables should not display multicollinearity, which is usually the case when wavenumbers from an FTIR analysis are used as inputs. To overcome that the wavenumber dataset may be treated using dimensionality reduction techniques [21,22].

Principal Component Analysis (PCA) is a data reduction technique that replaces the original correlated variables (such as FTIR wavenumbers) by new uncorrelated variables known as Principal Components (PCs). Dimensionality reduction occurs when only a few components that account for most of the variability in the data replace the original set of variables; that enables, for example, the handling of high-dimensional datasets as the one resulting from spectroscopy analyses [23,24].

Merging concepts from MLR and PCA, Principal Components Regression (PCR) builds a regression equation using PC scores instead of observations from the original variables used to obtain the PCs. PCR is an alternative to MLR when independent variables present multicollinearity [7,25]; drawbacks include the potential information loss resulting from selecting a limited number of PCs, and interpretability issues arising from using transformed rather than original variables in the regression [22].

Finally, Partial Least Squares Regression (PLSR) seeks a regression model describing the relationship between two sets of PCA-reduced variables. The first set corresponds to dependent variables, while the second set corresponds to independent variables. In PLSR regression coefficients are determined by maximizing the covariance between the two sets of reduced variables, which is not the same criterion used in PCR. Thus, modeling a single response problem through PLSR and PCR, which is our case in this paper, should lead to different results. In a dataset containing N wavenumbers, PLSR regression generates A orthogonal linear combinations (where $A<N$) of the original wavenumbers [26]. Data reduction through PCA enables PLSR to handle data with strong collinearity, high levels of noise and substantially fewer samples than wavenumber. Although being more popular than PCR there is no obvious advantage of PLSR over PCR [27], justifying the comparison of methods proposed here.