Using ANOVA-PCA to facilitate screening of large 2D-IR datasets

Contact *neil.hunt@strath.ac.uk*

R. Fritzsch

Department of Physics, University of Strathclyde, SUPA
107 Rottenrow East, Glasgow G4 0NG, UK

P. M. Donaldson

Central Laser Facility, STFC Rutherford Appleton Laboratory
Harwell Campus, Didcot, UK

G. M. Greetham

Central Laser Facility, STFC Rutherford Appleton Laboratory
Harwell Campus, Didcot, UK

M. Towrie

Central Laser Facility, STFC Rutherford Appleton Laboratory
Harwell Campus, Didcot, UK

A. W. Parker

Central Laser Facility, STFC Rutherford Appleton Laboratory
Harwell Campus, Didcot, UK

M. J. Baker

Department of Pure and Applied Chemistry, University of Strathclyde, WestCHEM, 99 GeorgeStreet, Glasgow G1 1RD, UK

N. T. Hunt

Department of Physics, University of Strathclyde, SUPA
107 Rottenrow East, Glasgow G4 0NG, UK

Introduction

Time-resolved, two-dimensional infrared (2D-IR) spectroscopy is a powerful technique that has provided insights into the structure and function of a range of proteins[1–5], enzymes[6,7] and DNA[8–10] by studying the dynamics of the vibrational energy landscape of the sample.

Modern 2D-IR spectrometers, such as the LIFEtime instrument[11] at the CLF, are capable of acquiring individual spectra within seconds and can obtain time-resolved 2D-IR measurements within a few minutes. This means that the expanded information content of 2D-IR spectroscopy is now available from a measurement that takes the same time as a conventional infrared absorption spectrum, paving the way for larger 2D-IR studies across a broader range of analytes and samples. In parallel with development in 2D-IR instrumentation however, there is a growing demand for high throughput-data analysis tools that are able to summarize and reduce the spectroscopic output relevant to a particular research question.

Many research studies investigate the effects of one or more systematically controlled factors within their experiment: How does a specific target molecule like a protein interact with a range of drug candidates? How does a specific drug candidate interact with a range of proteins? The outcome of these experiments is dictated by the particular choice made for each factor (which drug candidate and protein is being used) and the variance between the results usually builds the foundation for any further interpretation: Drug candidates 1, 2 and 3 give similar results when added to protein 2 and therefore interact in a similar fashion with the protein. It is possible to use this specific, experimental design to separate spectral features within a set of 2D-IR spectra according to the previously well-defined factors using ANOVA-PCA[12] (Analysis of variance combined with principal component analysis). This approach allows us to study the effect of each factor on the 2D-IR spectrum individually and gives an insight into whether certain combinations of factors return common spectral features.

The combination of ANOVA-PCA and 2D-IR has recently been demonstrated in a study featuring a dataset of 2016 2D-IR spectra[9] taken from 12 different oligomer DNA sequences in the presence/absence of DNA minor-groove binding molecule Hoechst 33258. It was shown that ANOVA-PCA enables fast differentiation between the ligand interactions of different DNA sequences. The following description is designed to illustrate how the ANOVA-PCA algorithm can be applied to a generalized 2D-IR research problem.

Experimental design, data formatting and preprocessing

Suppose there are two factors A and B, for example protein and drug, which determine the outcome of a 2D-IR experiment (see Figure 1). Each of these will contain a certain number of controlled parameters (*i* parameters, a_i, in A; *j* parameters, b_j, in B). The number of 2D-IR spectra measured has to be exactly the same for each combination of the parameters in each factor A and B in order to use ANOVA-PCA. This ensures an equal weighting of each combination of A and B for the analysis. If time-resolved 2D-IR spectra were recorded, the outcome of each measurement will also depend upon the waiting time variable and an additional, third factor, T, with *k* time points t_k, has to be considered.

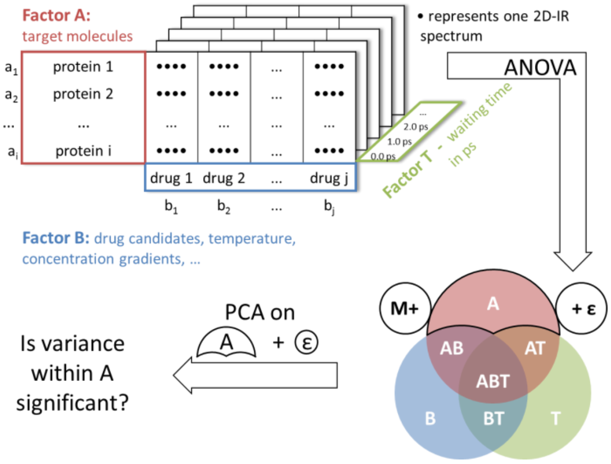
Factor A:
target molecules
• represents one 2D-IR spectrum
a1
a2
...
ai
protein 1
protein 2
...
protein i
drug 1
drug 2
...
drug j
b1
b2
...
bj
ANOVA
0.0 ps
1.0 ps
2.0 ps
...
Factor T - waiting time in ps
Factor B: drug candidates, temperature, concentration gradients, ...
M+
A
+ ε
AB
AT
ABT
B
BT
T
PCA on
A
+ ε
Is variance within A significant?

Figure 1 Schematic representation of the ANOVA−PCA method. A 2D-IR study is performed by introducing three main sources of variance into the dataset, factors A, B and T. The ANOVA step generates subsets of the data with variances attributable to each factor, effects due to combinations of these factors and a residual variance, ε (e.g., noise). Individual subsets are compared to residuals, ε, and analyzed using PCA to test for significance.

A 2D-IR experiment will return two-dimensional spectra, which need to be reformatted in order to be used as an input for ANOVA-PCA. An individual spectrum can be concatenated into vector-form so that each pixel of the 2D spectrum is treated as an independent variable in the analysis. The complete dataset can then be represented by a large *m*-by-*n* matrix $\boldsymbol{X}$ with *m* rows as spectra x_m, and *n* columns as individual 2D-IR pixels (Figure 2(a)). Each spectrum should be clearly identifiable by a set of labels (*e.g.* a, b, t) that correspond to the parameters used in each factor to generate the spectrum.

Preprocessing techniques,[13] such as PCA noise removal, Savitzky-Golay-Smoothing and normalization methods can be applied to improve the significance of the analysis results. It should be noted however that normalization methods across different waiting times will remove the kinetic information and should therefore be avoided.

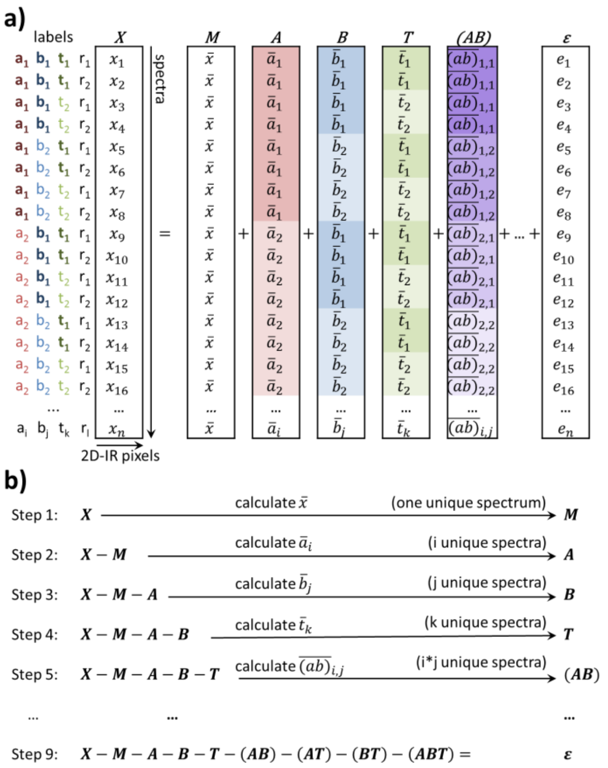
a)
labels
X
M
A
B
T
(AB)
ε
spectra
2D-IR pixels
b)
Step 1: $\boldsymbol{X} \xrightarrow{\text{calculate } \bar{x} \quad \text{(one unique spectrum)}} \boldsymbol{M}$
Step 2: $\boldsymbol{X}-\boldsymbol{M} \xrightarrow{\text{calculate } \bar{a}_i \quad \text{(i unique spectra)}} \boldsymbol{A}$
Step 3: $\boldsymbol{X}-\boldsymbol{M}-\boldsymbol{A} \xrightarrow{\text{calculate } \bar{b}_j \quad \text{(j unique spectra)}} \boldsymbol{B}$
Step 4: $\boldsymbol{X}-\boldsymbol{M}-\boldsymbol{A}-\boldsymbol{B} \xrightarrow{\text{calculate } \bar{t}_k \quad \text{(k unique spectra)}} \boldsymbol{T}$
Step 5: $\boldsymbol{X}-\boldsymbol{M}-\boldsymbol{A}-\boldsymbol{B}-\boldsymbol{T} \xrightarrow{\text{calculate } \overline{(ab)}_{i,j} \quad \text{(i*j unique spectra)}} (\boldsymbol{AB})$
Step 9: $\boldsymbol{X}-\boldsymbol{M}-\boldsymbol{A}-\boldsymbol{B}-\boldsymbol{T}-(\boldsymbol{AB})-(\boldsymbol{AT})-(\boldsymbol{BT})-(\boldsymbol{ABT}) = \boldsymbol{\varepsilon}$

Figure 1. (a) Raw data formatted as an *m*-by-*n* matrix ***X*** and its decomposition into factor matrices. Each row represents one 2D-IR spectrum concatenated to form a vector. Blocks of color indicate averaged spectra in each factor matrix. (b) Schematic structure of the step-wise subtraction to calculate factor matrices. The algorithm allocates variance from the raw data based on their origin into several new matrices with the same dimensions by calculating averages. The residual matrix ideally just contains the variance between repeats (r).

The ANOVA-PCA algorithm: ANOVA

The algorithm is based on a publication by Harrington *et al.*[12] and the first step is a multivariate version of a three-way ANOVA that aims to separate the variance within the raw data matrix ***X*** according to the three pre-defined factors A, B and T. For each factor, a separate matrix with the same dimensions as ***X*** is sequentially generated and subtracted from ***X*** to obtain a residual matrix, $\boldsymbol{\varepsilon}$. An ANOVA-PCA with three factors decomposes the raw data ***X*** according to the following equation:

$$\begin{aligned}\boldsymbol{X} = \quad & \boldsymbol{M}+\boldsymbol{A}+\ \boldsymbol{B}+\boldsymbol{T} \\ & +(\boldsymbol{AB})+(\boldsymbol{AT})+(\boldsymbol{BT})+(\boldsymbol{ABT}) \\ & +\boldsymbol{\varepsilon}. \end{aligned} \qquad (1)$$

Matrices ***A***, ***B*** and ***T***, contain 2D-IR responses that are only due to the parameter change in one factor. Additional matrices ***(AB)***, ***(AT)***, ***(BT)***, and ***(ABT)***, are generated for combinations of any of the three factors. These so-called interactions describe any variance that occurs in a specific combination of the main factors. While matrix ***A*** for example would contain the 2D-IR response of each protein averaged over all drug candidates, ***(AB)*** would only include variations from this average response due to specific drug-protein combinations. Matrix ***(AT)*** in this example would contain protein specific deviations from the average kinetics found in ***T***.

The matrices for equation 1 are calculated in a stepwise manner from the raw data ***X*** as shown in Figure 2(b). In the first step, a global average spectrum $\bar{x}$ for all spectra in ***X*** is calculated and a new matrix ***M*** is generated. Matrix ***M*** still has the same dimensions as ***X*** but as shown in Figure 2(b), only contains repeatedly $\bar{x}$ in every row. Subtraction $\boldsymbol{X}-\boldsymbol{M}$ generates a residual matrix which is used in subsequent steps. In step 2, matrix ***A*** for factor A is being generated. All rows of $\boldsymbol{X}-\boldsymbol{M}$ that belong to the same parameter a_i are averaged to give spectra $\bar{a}_i$. These i average spectra $\bar{a}_i$ are repeated according to their dedicated rows to form matrix ***A***. The subtraction $\boldsymbol{X}-\boldsymbol{M}-\boldsymbol{A}$ then results in a new residual matrix containing reduced variance. In the next step, $\boldsymbol{X}-\boldsymbol{M}-\boldsymbol{A}$ is used to calculate j average spectra $\bar{b}_j$, thus forming matrix ***B*** accounting for variance due to factor B. $\boldsymbol{X}-\boldsymbol{M}-\boldsymbol{A}-\boldsymbol{B}$ is then used to create matrix ***T*** in which all spectra with a common waiting time (t_k) are averaged. This is subtracted in turn to account for the delay time factor T. This step-wise subtraction is repeated for all possible interactions between the three factors, so that ***(AB)*** will consist of $i \cdot j$ unique spectra $\overline{(ab)}_{i,j}$, ***(AT)*** will consist of $i \cdot k$ unique spectra $\overline{(at)}_{i,k}$ and so on. The last subtraction will result in residual matrix $\boldsymbol{\varepsilon}$, which theoretically contains only variance between repeats and instrumental noise.

It should be noted that the result of the stepwise subtraction is only independent of the order of the subtracted factors if the number of spectra analyzed is exactly the same for each parameter of each factor. Each parameter combination will then have the same weighting during the analysis and matrix ***X*** can be described as a balanced dataset. This implies that if any of the measured spectra were to be excluded, for example because of unusual high scatter artifacts, an additional number of spectra will have to be excluded accordingly to maintain a balanced dataset.

The ANOVA-PCA algorithm: PCA

Once the matrices for factors and interactions are assembled, they can be tested for significance via PCA. The residual matrix $\boldsymbol{\varepsilon}$ represents any variance unexplainable by any of the factors and can therefore act as a reference for noise within the dataset. Any of the matrices obtained from the stepwise subtraction can be added back to the residuals, $\boldsymbol{\varepsilon}$**,** and investigated via PCA. The PCA will transform this subset from possibly correlated variables (2D-IR pixels) into orthogonal principal components. The first few principal components describe the majority of the variance and if there is a significant, systematic 2D-IR response due to the parameter change of a factor, the first few principal components will retrieve this change. Otherwise, the noise from residual matrix $\boldsymbol{\varepsilon}$, will dominate any systematic change and the first principal component will resemble random noise.

Conclusion

With the decomposition into subsets it is possible to study the 2D-IR dataset step-by-step by examining the effect of individual factors and interactions. It is also possible to subsequently add more factors and interactions to the subset to gradually increase the complexity of the data analyzed and understand the origin of subtle differences. A detailed application of this to a proof-of-concept DNA-ligand 2D-IR screening experiment is described in a recent publication by Fritzsch *et al.*[9]

Acknowledgements

Funding is gratefully acknowledged from STFC for programme access to the Central Laser Facility systems (ST/M000125/1) and from BBSRC (BB/L014335/1) for the LIFEtime spectrometer.

References

1 Z. Ganim, H. S. Chung, A. W. Smith, L. P. Deflores, K. C. Jones and A. Tokmakoff, *Acc. Chem. Res.*, 2008, **41**, 432–441.

2 K. Adamczyk, M. Candelaresi, K. Robb, A. Gumiero, M. a Walsh, A. W. Parker, P. a Hoskisson, N. P. Tucker and N. T. Hunt, *Meas. Sci. Technol.*, 2012, **23**, 062001.

3 T. Elsaesser, *Chem. Rev.*, 2017, **117**, 10621–10622.

4 N. T. Hunt, *Chem. Soc. Rev.*, 2009, **38**, 1837–1848.

5 P. J. M. Johnson, K. L. Koziol and P. Hamm, *J. Phys. Chem. Lett.*, 2017, **8**, 2280–2284.

6 M. C. Thielges, J. K. Chung and M. D. Fayer, *J. Am. Chem. Soc.*, 2011, **133**, 3995–4004.

7 P. Pagano, Q. Guo, A. Kohen and C. M. Cheatum, *J. Phys. Chem. Lett.*, 2016, **7**, 2507–2511.

8 G. Hithell, P. M. Donaldson, G. M. Greetham, M. Towrie, A. W. Parker, G. A. Burley and N. T. Hunt, *Chem. Phys.*, 2018, in Press.

9 R. Fritzsch, P. M. Donaldson, G. M. Greetham, A. W. Parker, M. J. Baker and N. T. Hunt, *Anal. Chem.*, 2018, **90**, 2732–2740.

10 P. J. Sanstead, P. Stevenson and A. Tokmakoff, *J. Am. Chem. Soc.*, 2016, **138**, 11792–11801.

11 P. M. Donaldson, G. M. Greetham, D. J. Shaw, A. W. Parker and M. Towrie, *J Phys Chem A*, 2018, **122**, 780–787.

12 P. D. B. Harrington, N. E. Vieira, J. Espinoza, J. Kae, R. Romero and A. L. Yergey, *Anal. Chim. Acta*, 2005, **544**, 118–127.

13 B. R. Smith, M. J. Baker and D. S. Palmer, *Chemom. Intell. Lab. Syst.*, 2018, **172**, 33–42.