Supplementary Material
Metabolomics data analysis improvement by use of filter diagonalization method

Hernán Joel Cervantes Felipe M. Kopel
Said R. Rabbani
Institute of Physics, USP
Rua do Matão, 1371 CEP 05508-090
Tel.: +51-11-30916692
email: hernanif.usp.br

August 16, 2019

1 Concise theory of filter diagonalization method

1.1 Introduction

The filter diagonalization method (FDM) was used initially in quantum mechanics by Neuhauser, (Neuhauser, 1990), in 1990. The FDM was originally employed for extracting high-energy eigenstates in any arbitrary range of energies, rejecting distant eigenstates. Later, Chen applied FDM to estimate the NMR spectrum from FID, Chen et al. (2000). The FDM models the Free induction decay (FID) as a sum of damped harmonic function, which is equivalent to say that the spectra of NMR contains only Lorentzian peaks. The FID is writing as, Chen et al. (2000) :

$$c_n = \sum_{k=1}^{K} d_k \, \mathrm{e}^{i\omega_k n\tau}, \; n = 0, 1, \cdots, N; \tag{1}$$

where c_n is nth point in the FID time series, in general a complex number; $d_k = a_k \, \mathrm{e}^{i\phi_k}$, a complex number which gives the peak area (a_k) and relative phase (ϕ_k); $\omega_k = 2\pi f_k + i\gamma_k$ contains the peak position (f_k) and line width (γ_k); τ is the sampling time interval, inverse of sample rate; N is the number of acquisition points of FID and K is the number of peaks present in the spectrum, usually an unknown parameter. As the samples in metabolomics are diluted , normally, in deuterated water, the liquid spectrum contains Lorentzian peaks, Goldman (1991).

1.2 Harmonic inversion problem

In FDM, the signal $c_n = c(t_n) = c(\tau n)$ is associated with a time autocorrelation function of a dissipation dynamical quantum system characterized by an effective non-Hermitian Hamiltonian operator $\hat{\mathbf{\Omega}}$ and some initial state $|\Phi_0)$,

$$c_n = (\Phi_0|\ \Phi_n) = (\Phi_0|\, \mathrm{e}^{in\tau\hat{\Omega}}\, |\Phi_0)\,, \tag{2}$$

where a complex symmetric inner product is used, $(a|b) = (b|a)$, without complex conjugate. Assuming that the operator $\hat{\mathbf{\Omega}}$ is diagonalizable, it is possible to write the operator in its spectral representation,

$$\hat{\mathbf{\Omega}} = \sum_k \Omega_k\, |\Upsilon_k)\, (\Upsilon_k|\,, \tag{3}$$

in equation (3), Ω_k are the eigenvalues and $|\Upsilon_k)$ its correspondent eigenvector, which satisfies the ortho-normality relationship

$$(\Upsilon_k|\ \Upsilon_{k'}) = \delta_{kk'}. \tag{4}$$

Projecting the equation (2) into the basis $|\Upsilon_k)\,(\Upsilon_{k'}|$, gives:

$$c_n = \sum_k (\Phi_0|\ \Upsilon_k)^2\, \mathrm{e}^{in\tau\Omega_k}, \tag{5}$$

and, comparing with equation (1), can be identified that $d_k^{1/2} = (\Phi_0|\ \Upsilon_k)$ and $\omega_k = \Omega_k$. It is possible to define the evolution operator $\hat{\boldsymbol{U}} \equiv \mathrm{e}^{i\tau\hat{\mathbf{\Omega}}}$. The representation of the evolution operator in the basis of eigenstates of the operator $\hat{\mathbf{\Omega}}$, gives

$$\hat{\boldsymbol{U}}\, |\Upsilon_k) = \mathrm{e}^{i\tau\Omega_k}\, |\Upsilon_k)\,. \tag{6}$$

Defining $u_k = \mathrm{e}^{i\tau\omega_k}$, the equation (6) is just an eigenvalue problem with u_k the eigenvalues and $|\Upsilon_k)$ the eigenvectors of the evolution operator $\hat{\boldsymbol{U}}$.

There is matrix representation of $\hat{\boldsymbol{U}}$ and $|\Phi_k)$, corresponding to the experimental time series, it is not necessary to know explicitly $\hat{\mathbf{\Omega}}$, $\hat{\boldsymbol{U}}$ and $|\Phi_0)$. As there are two parameters for each peak, it is possible to obtain up to $M = N/2$ Lorentzian lines for a time series of N complex points.

1.3 FDM in Fourier-type basis

The simplest basis that can be used to solve the FDM problem is the Krylov basis, generated by propagating the initial state, as:

$$|\Phi_n) = \hat{\boldsymbol{U}}^n\, |\Phi_0)\,, \tag{7}$$

resulting in the generalized eigenvalue problem:

$$\boldsymbol{U}^1 \boldsymbol{b}_k = u_k \boldsymbol{U}^0 \boldsymbol{b}_k. \tag{8}$$

$\boldsymbol{U}^1$ and $\boldsymbol{U}^0$ are matrices of dimension $M \times M$. For a typical NMR experiment, frequently, 32 768 (32 k), or even more complex points are registered, therefore to solve the equation (8) one needs a big computing platform, moreover the problem is highly ill-posed. In addition, the M ($M = N/2 = 16\,384$) is almost hundred times more than the Lorentzian peaks present in a NMR spectrum. This problem can be solved by using another basis $|\Psi_j)$ where the matrices $\boldsymbol{U}^1$ and $\boldsymbol{U}^0$ are block-diagonal, and the diagonalization can be carried out in multiple steps, using the divide to conquer strategy, dividing the whole spectrum into small spectral windows and analyze them separately. The basis that allows this block-diagonal matrices is the so called the Fourier type basis, defined as the linear combination of the Krylov basis, as, Mandelshtam and Taylor (1997),

$$|\Psi_j) = \sum_{n=1}^{M-1} \mathrm{e}^{in\tau\psi_j} \, |\Phi_n) \;\; j = 1, 2, \cdots, M, \tag{9}$$

where $\{\psi_j\} \subset \mathbb{R}$ are equidistant frequencies within the spectral window of interest. The transformation from the Krylov basis to Fourier basis is unitary and each $|\Psi_j)$ is localized in the frequency domain, i.e., only eigenvectors $|\Upsilon_k)$ of $\hat{\boldsymbol{\Omega}}$ with eigenvalues $\psi_j \simeq \Omega_k$ contribute significantly. This means that only a small subset, $K_{win} \ll M$, of frequencies ψ_j in the frequency domain are enough for the determination of whole spectrum. The frequency density is used to characterize the grid. In order to obtain a high resolution spectra the frequency density, $\rho(\psi_j)$, must be greater than the eigenfrequencies density, $\rho(\Omega_k)$, called the *local completeness condition*. The word of wisdom says to choose $\rho(\psi_j)$ between 1.1 and 1.2 times $\rho(\Omega_k)$. The $\rho(\Omega_k)$ can be determined by the Nyquist theorem $\rho(\Omega_k) = N\tau/2\pi$. For a equidistant grid within the frequency interval (ψ_a, ψ_b) and J frequencies,

$$\rho(\psi_j) = \frac{2\pi}{N\tau} \frac{J}{\psi_b - \psi_a} = \frac{2\pi J}{N\tau(\psi_b - \psi_a)}. \tag{10}$$

The number of frequencies within $\{\psi_j\}$, preserving $\rho(\psi_j)/\rho(\Omega_k) \approx 1.1 - 1.2$ condition, equation (10) and Nyquist density can be determined.

The matrices projected in the Fourier basis are quasi-diagonal, near off-diagonal frequencies are not zero but the off-diagonal elements far from the diagonal are practically zero. Then, these matrices can be diagonalized in a block wise scheme, with block dimension $L \ll J$ splitting the entire spectrum into W subintervals with L frequencies ψ_j, Magon (2007). One should remember that peaks near interval boundaries are less accurate, because they use fewer eigenfrequencies. This problem can be overcome by implementing superimposed windows weighted by a function to penalize the peaks at the limits of each interval. (Chen et al., 2000, page 58), another scheme is to select the central 50% of overlapping intervals, excluding the initial and final 25% of each interval, (Magon, 2007, page 55). In this work the last scheme was used.

Summarizing, the FDM algorithm in the Fourier base consists in choosing the spectral window of interest within the Nyquist interval, the number of intervals

W, keeping constant the number of frequencies within each interval (L), and select the density $\rho(\psi_j)$. These three parameters must be chosen so that $L \gg 1$, the density ratio $\rho(\psi_j)/\rho(\Omega) \approx 1.1 - 1.2$. A good value for L is $L \approx 10 - 100$, if $L \not\gg 1$ then $W = 1$ should be chosen, and making L as high as possible. Once these parameters have been chosen, the generalized eigenvalue equation

$$\breve{U}^1 \breve{B}_k = \mu_k \breve{U}^0 \breve{B}_k \tag{11}$$

must be solved. In equation (11), $\breve{U}^p$, $p = 0, 1$, is the projection of the evolution operator $\hat{\boldsymbol{U}}^p$ in the Fourier basis for each interval whose elements are given by:

$$\begin{array}{rl} \breve{U}^p_{jk} = & \frac{1}{z_j - z_k}\left\{ z_j \left[G^p_k - z_j^{-M} G^{M+p}_k \right] - z_k \left[G^p_j - z_k^{-M} G^{M+p}_j \right] \right\}; \\ & j,k = 0, 1, \cdots, L-1;\ j \neq k;\ p = 0, 1 \\ \breve{U}^p_{jj} = & \sum_{n=0}^{M-1} \left[(n+1)\, z_j^{-n} c_{n+p} + (M-n-1)\, z_j^{-M} z_j^{-n} c_{n+M+p} \right]; \\ & j = 0, 1, \cdots, L-1;\ p = 0, 1 \\ G^k_j = & \sum_{n=0}^{M-1} z_j^{-n} c_{n+k};\ j = 0, 1, \cdots, L-1, \end{array} \tag{12}$$

where $z_j = \mathrm{e}^{-i\tau\psi_j}$, $M = [\![N/2]\!]$, c_j, $j = 0, 1 \cdots, N-1$, are the recorded time series data.

The generalized eigenvalue problem, equation (11) with parameters given by equation (12), can be solved, using the generalized eigenvalue problem solver from LAPACK library, Anderson et al. (1999), as suggested in Chen et al. (2000); Magon (2007). The LAPACK's routine *zggev* returns the eigenvalues as two vectors α and β, with $\mu_k = \alpha_k/\beta_k$ if $\beta_k \neq 0$. Since the data is noisy and, the matrix dimensions are not large enough, the matrices can be quasi-singular. To avoid the inclusion of wrong peaks, as each eigenvalue, eigenvector pair correspond to only one peak, the so-called SVD regularization, Neumaier (1998), is used to exclude the eigenvalues smaller than a previously defined threshold.

The routine *zggev* also returns the so-called right eigenvectors that meet the ortho-normality relationship

$$\left(\breve{B}_k \right)^{\mathrm{T}} \breve{U}^0 \breve{B}_{k'} = \delta_{kk'}, \tag{13}$$

where $()^{\mathrm{T}}$ means matrix transpose.

Once the eigenvalues and eigenvector are obtained, by solving the eigenvalue problem subject to the ortho-normality condition and equation (13), the Lorentzian peaks parameters can be calculated with:

$$\begin{array}{l} \Omega_k = \frac{i}{\tau} \ln \mu_k \\ d_k = \left[\left(\breve{B}_k \right)^T \breve{C} \right]^2, \quad \text{where, } \breve{C} = \sum_{n=0}^{M-1} z_j^{-n} c_n. \end{array} \tag{14}$$

In general, μ_k is a complex number, therefore the logarithm in equation (14), pertain to the principal branch of the complex logarithm whose imaginary part lies in the interval $(-\pi, \pi]$, Geroldinger and Halter-Koch (2006), defined by:

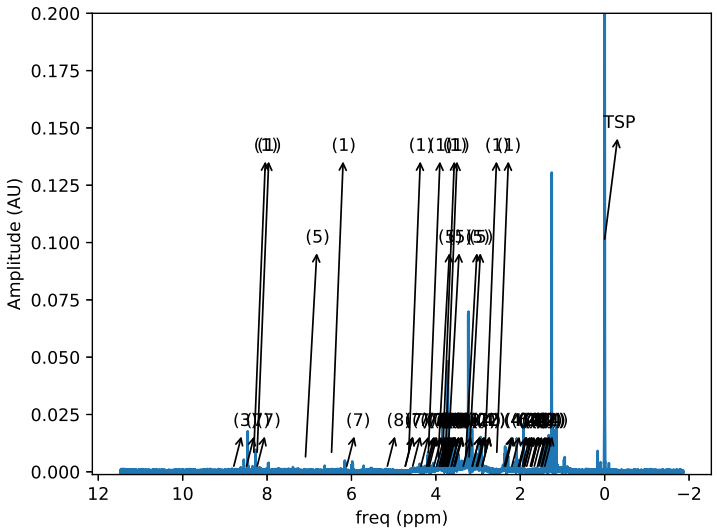

Figure 1: Full NMR spectrum for the sample *CTL-ctl_19-02-16*. The spectrum was processed with FFT. Some identified compounds are present, see table (1) for legend.

$$\ln z = \ln |z| + i \arg z. \tag{15}$$

2 NMR Results

2.1 FFT process

The sample *CTL-ctl_19-02-16* was process using the *Chenomx* suite version 7.7, Chenomx Inc. (2013). Automatic phase correction followed by fine manual correction was applied to the FID. Also was employed automatic cubic spline baseline correction with 20 points. The region belonging to the interval from 4.63702 ppm to 5.11702 ppm was removed because in this region the remaining water spectrum is present. The spectrum is shown in figure (1). Also in this figure is exhibited some metabolites that were identified, using the Chenomx Inc. (2013) suite and the TOCSY experiment what is part of the study. the number meaning can be found in table (1).

2.2 FDM process

In figure (2) is shown the full NMR spectrum of the former sample. In this case, the FDM was used to process using the parameters of table (2).

Number	Name	Concentration
1	2'-Deoxyadenosine	53.1
2	5-Hydroxylysine	24.2
3	O-Phosphoserine	15.0
4	Saccharopine	14.0
5	n-Methylhistidine	10.8
6	4-Hydroxybutyrate	3.1
7	ADP	6.6
8	Mannose	13.3

Table 1: Some metabolites identified for the sample *CTL-ctl_19-02-16*.

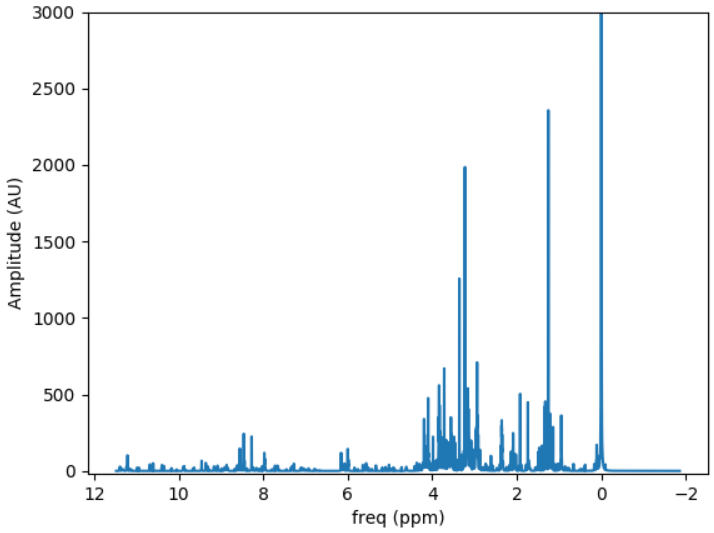

Figure 2: Full NMR spectrum for the sample *CTL-ctl_19-02-16*. The spectrum was build using the FDM.

Name	Value	Unit
N	32 000	
L	100	
ρ	1.196 875 00	
Φ_a	−4000	Hz
Φ_b	4000	Hz
$\Delta\Phi$	20.915 032 679 738 562	Hz
W	765	Hz
LimInf	1.0×10^{-9}	

Table 2: Parameters used to build the spectrum shown in figure (2). N is the number of points in the FID, and LimInf is the eigenvalues cutoff value.

References

E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen. *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, third edition, 1999. ISBN 0-89871-447-8 (paperback). URL `https://epubs.siam.org/doi/pdf/10.1137/1.9780898719604`. 1.3

J. H. Chen, V. A. Mandelshtam, and A. J. Shaka. Regularization of the two-dimensional filter diagonalization method: Fdm2k. *Journal of Magnetic Resonance*, 146(2):363–368, Oct. 2000. doi: 10.1006/jmre.2000.2155. 1.1, 1.3, 1.3

Chenomx Inc. Chenomx nmr suite 7.7, 2013. URL `https://www.chenomx.com/`. 2.1

A. Geroldinger and F. Halter-Koch. Complex analysis. In *Non-Unique Factorizations Algebraic, Combinatorial and Analytic Theory*, Pure and Applied Mathematics, pages 659–669. Chapman and Hall/CRC, jan 2006. ISBN 978-1-4200-0320-8. doi: 10.1201/9781420003208.axb. 1.3

M. Goldman. *Quantum Description of High-Resolution NMR in Liquids (International Series of Monographs on Chemistry)*. Oxford University Press, 1991. ISBN 019855639X. 1.1

C. Magon. *A inversão harmônica do espectro de ressonância magnética: uma solução para o problema dos autocampos.* Livre docencia, Instituto de Física da USP de São Carlos, 2007. 1.3, 1.3

V. A. Mandelshtam and H. S. Taylor. Harmonic inversion of time signals and its applications. *Journal of Chemical Physics*, 107(17):6756–6769, Nov. 1997. doi: 10.1063/1.475324. 1.3

D. Neuhauser. Bound-state eigenfunctions from wave-packets - time-energy resolution. *Journal of Chemical Physics*, 93(4):2611–2616, Aug. 1990. doi: 10.1063/1.458900. 1.1

A. Neumaier. Solving ill-conditioned and singular linear systems: A tutorial on regularization. *SIAM Review*, 40(3):636–666, 1998. ISSN 00361445. URL `http://www.jstor.org/stable/2653234`. 1.3