Data handling of complex GC-MS chromatograms: characterization of n-alkane distribution as chemical marker in organic input source identification

Maria Chiara Pietrogrande, Mattia Mercuriali, Luisa Pasti, Francesco Dondi

Department of Chemistry, University of Ferrara, Via L. Borsari, 46, 44100 Ferrara, Italy;

APPENDIX

Combination of $EACVF_o(\Delta t)$ and $EACVF_e(\Delta t)$ to obtain $EACVF_{tot}(\Delta t)$

In the following, mathematical equations are derived to relate the Experimental Autocovariance Function $EACVF_{tot}(\Delta t)$ computed on all the terms (tot=odd+even) of a homologous series to the separated contributions of $EACVF_o(\Delta t)$ and $EACVF_e(\Delta t)$ computed on the odd and even terms of the series, respectively. In particular, the computation of $EACVF_{tot}(kb)$ at $\Delta t = kb$ for even and odd k values is investigated in order to extract information on the distribution of the odd/even terms of the series. In order to simplify the preliminary treatment, at first the terms of the series are considered as made by pure spikes and therefore as chromatographic peaks of a given shape (e.g. gaussian)

Peaks represented by spikes.

For the sake of simplicity, we consider two distinct signals representing the odd and even terms of a homologous series: each of them is formed by N_p data points sampled at frequency equal to $1/\tau$, with values $a_{o,h,j}$ and $a_{e,h,j}$ ($j = 1$ to N_p). Each series contains n spikes, having height values significantly different from 0 and located as an ordered sequence at fixed and constant inter-distance $2b$, all the other points being set equal to zero (hypothesis #1) . Consequently, the spikes are located at position $2kb$ for each series. Moreover, it is assumed that $N_p >> n$ (hypothsis #2). The two series (odd and even) are shifted from each other, by a lag time b. For each series, the $EACVF$ computed at a given time lag ($\Delta t = s\tau$) according to eq. 1 is given by:

$$EACVF_o(s) = \sum_{j=1}^{N_p - s} (a_{o,h,j} - \overline{a}_o) \cdot (a_{o,h,j+s} - \overline{a}_o) \qquad s = 0,1,2...M-1 \tag{a1}$$

$$EACVF_e(s) = \sum_{j=1}^{N_p - s} (a_{e,h,j} - \overline{a}_e) \cdot (a_{e,h,j+s} - \overline{a}_e) \qquad s = 0,1,2...M-1 \tag{a2}$$

where $\overline{a}_o$ and $\overline{a}_e$ are the average signal values of the two series of the n spikes, N_p being the number of points in the signal, and M the truncation point in the $EACVF$ computation.

Under the above mentioned hypotheses #1 and #2, we can assume that the average signal value is equal to zero:

$$\overline{a}_o \approx \overline{a}_e \approx 0 \tag{a3}$$

Note that the above described model of the spike series is a good approximation for an equivalent experimental chromatogram displaying a number of peaks significantly lower than the number of baseline points and by assuming that the baseline signal has been properly subtracted to achieve a mean value close to zero. In general, the latter condition can be easily achieved for any signal by applying a mean centering pre-treatment.

Under the above mentioned approximation (see eq. a3), eqs. a1 and a2 become:

$$EACVF_o(s) = \sum_{i=1}^{N_p - s} (a_{o,h,j} \cdot a_{o,h,j+s}) \qquad s = 0,1,2...M-1 \tag{a4}$$

$$EACVF_e(s) = \sum_{i=1}^{N_p - s} (a_{e,h,j} \cdot a_{e,h,j+s}) \qquad s = 0,1,2...M-1 \tag{a5}$$

An additional hypothesis is now introduced: spike heights and positions are mutually independent (hypothesis #3). Eqs a4 and a5 can be rewritten only considering the sum of the $a_{h,i}$ terms located at position $2kb$ of the n spikes, since the $a_{h,j}$ values at the N_p points other than the spike locations are zero, according to hypotesis #2. Moreover, according to hypothesis #3, the quantities $a_{h,j}$ and $a_{h,j+s}$ are statistically equal and thus we obtain [1A]:

$$EACVF_o(2kb) = \sum_{i=1}^{n-k} a_{o,h,i}^2 \qquad k = 0,1,2...M-1 \tag{a6}$$

$$EACVF_e(2kb) = \sum_{i=1}^{n-k} a_{e,h,i}^2 \qquad k = 0,1,2...M-1 \tag{a7}$$

In the general case of n spikes, where the peak height a_i of each i^{th} peak is randomly distributed around the mean value $\overline{a}$ displaying a standard deviation σ_a, the following equation can be written to estimate σ_a^2:

$$\sigma_a^2 = \frac{\sum_i^n (a_i - \overline{a})^2}{n} \tag{a8}$$

where $\overline{a}$ is the mean value:

$$\overline{a} = \sum_i^n a_i \ / n \tag{a9}$$

Equation a8 can be rewritten as:

$$\sigma_a^2 = \frac{\sum_i^n a_i^{\,2}}{n} - \overline{a}^2 \tag{a10}$$

which can be rearranged to give:

$$\sum_i^n a_i^2 = (\sigma_a^2 + \overline{a}^2)n \tag{a11}$$

This equation, a11, is strictly true for $n \to \infty$: in the case of the studied chromatographic signal, it corresponds to a "sufficiently large" number n of spikes representing the chromatographic peaks, i.e., $n \geq 30$. The equations a11 can be applied to the case of the two ordered sequences formed by $a_{o,h,i}$ and $a_{e,h,i}$ signals, and introduced in eqs. a6 and a7 to give as final result:

$$EACVF_o(2kb) = \sum_i^{n-k} a_{o,h,i}^2 = (\sigma_{a_{o,h,i}}^2 + \overline{a}_{o,h,i}^2)(n-k) = (n-k)\overline{a}_{o,h,i}^2 \left(\frac{\sigma_{a_{o,h,i}}^2}{\overline{a}_{o,h,i}^2} + 1 \right) \tag{a12}$$

$$EACVF_e(2kb) = \sum_i^{n-k} a_{e,h,i}^2 = (\sigma_{a_{e,h,i}}^2 + \overline{a}_{e,h,i}^2)(n-k) = (n-k)\overline{a}_{e,h,i}^2 \left(\frac{\sigma_{a_{e,h,i}}^2}{\overline{a}_{e,h,i}^2} + 1 \right) \tag{a13}$$

The $n-k$ values in eqs a12 and a13 correspond to the actual values of the n term of eq. a11 (see the upper limit of the sums in equations in a6 and a7).

Let us now consider the superimposition of the odd and the even series, i. e. the total series. The pertinent expression for the total series, analogous to eq a1 or a2, is:

$$EACVF_{tot}(s) = \sum_{j=1}^{N_p} (a_{o,h,j} + a_{e,h,j} - (\overline{a}_o + \overline{a}_e)) \cdot (a_{o,h,j+s} + a_{e,h,j+s} - (\overline{a}_o + \overline{a}_e)) \qquad \text{(a14)}$$

In fact, by comparing eqs. a14 and a1 one can see that, at each i position ($j = 1..N_p - s$), the signal is the sum of the even and odd values, whereas the average values of the total series are the sum of the average values for the two series (cfr. eq. a14 to eq 1).

The sum in eq 14 is now developed. The terms odd×odd and even×even are collected and thus they become equal to eq a4 and a5, respectively. Moreover hypothesis #2 ($\overline{a}_o \approx \overline{a}_e \approx 0$) is assumed and eqs. a14, a4 and a5 are combined to give:

$$EACVF_{tot}(s) = EACVF_o(s) + EACVF_e(s) + \sum_{j=1}^{N_p - s} (a_{o,h,j} \cdot a_{e,h,j+s}) + \sum_{j=1}^{N_p - s} (a_{e,h,j} \cdot a_{o,h,j+s}) \qquad \text{(a15)}$$

In eq a15, only the s positions located at $\Delta t = kb$ are interesting, the other s positions being equal to zero (see at eqs a6, a7 derivation), eq. a15 becomes:

$$EACVF_{tot}(kb) = EACVF_o(kb) + EACVF_e(kb) + \sum_{i=1}^{n-k} (a_{o,h,i} \cdot a_{e,h,i+s}) + \sum_{i=1}^{n-k} (a_{e,h,i} \cdot a_{o,h,i+s}) \qquad \text{(a16)}$$

The general equation a16 assumes simplified forms when computed at $\Delta t = kb$ values.

At $\Delta t = kb$ for even k values the cross correlation terms are equal to zero since the two series are shifted by the same quantity $\Delta t = b$, i.e.:

$$\text{If} \quad a_{o,h,i} = 0 \quad a_{e,h,i+kb} \neq 0 \quad \forall i \quad k = 0,2,4,...$$

$$\text{If} \quad a_{e,h,i} \neq 0 \quad a_{o,h,i+kb} = 0 \quad \forall i \quad k = 0,2,4,...$$

and eq. a16 assumes the simplified form:

$$EACVF_{tot}(kb) = EACVF_o(kb) + EACVF_e(kb) \qquad k = 0,2,4,... \qquad \text{(a17)}$$

At $\Delta t = kb$ for odd k values, one has:

$$\text{If} \quad a_{o,h,i} = 0 \quad a_{o,h,i+kb} \neq 0 \quad \forall i \quad k = 1,3,5,...$$

$$\text{If} \quad a_{e,h,i} \neq 0 \quad a_{e,h,i+kb} = 0 \quad \forall i \quad k = 1,3,5,...$$

and eq. a16 contains only the cross correlation terms which, in this case are not equal to zero:

$$EACVF_{tot}(kb) = \sum_{i=1}^{n-k} (a_{o,h,i} \cdot a_{e,h,i+kb}) + \sum_{i=1}^{n-k} (a_{e,h,i} \cdot a_{o,h,i+kb}) \qquad k = 1,3,5,... \qquad \text{(a18)}$$

Eq. a18 can be further simplified by considering the symmetry property of the product of two random viariables (i.e. $a_{o,h,i}$ and $a_{e,h,i}$) under the above mentioned hypothesis #3.

$$EACVF_{tot}(kb) = 2\sum_{i=1}^{n-k}(a_{o,h,i} \cdot a_{e,h,j}) \tag{a19}$$

Let us introduce the following hypothesis: the height distributions of the two spikes are linearly related, which can be expressed by:

$$a_{e,h,i+s} = \frac{a_{o,h,i}}{R} \tag{a20}$$

where R is a constant.

By combining eq a19 and a20, one has:

$$\sum_{i=1}^{n-k}\left(a_{o,h,i}a_{e,h,i+s}\right) = \sum_{i=1}^{n-k}\left(a_{o,h,i}\frac{a_{o,h,i}}{R}\right) = \frac{EACVF_o}{R} \tag{a21}$$

For even k values, eq a19 becomes:

$$EACVF(k) = 2\frac{EACVF_o(k)}{R} \qquad k \text{ even} \tag{a22}$$

and for odd k values, eq a19 becomes:

$$EACVF(k) = EACVF_o(k) + \frac{EACVF_o(k)}{R^2} \qquad k \text{ odd} \tag{a23}$$

Gaussian shape peaks.

In the most general case representing experimental chromatograms, the chromatographic signals of a homologous series can be regarded as a series of pulses whose parameters are random variables. The pulses are the single-component peaks, whose shape is supposed to be constant along the chromatogram. The position of the peak is deterministic (equal to 2*kb*, see hypothesis #2) whereas the height *h* of the pulses is a random variable. By assuming the shape of the pulses to be a Gaussian function of constant standard deviation (σ) and random abundance ($a_{h,i}$), distribution is given by:

$$Y_j = \sum_{i=1}^{m} a_{h,i} \cdot e^{\left[-\frac{1}{2}\left(\frac{j-2ib}{\sigma}\right)^2\right]} \tag{a24}$$

It has been demonstrated in Ref. 23 that in this case the $EACVF_{tot}(kb)$ at the repeated interdistances $\Delta t = kb$ is given by eq 3 in the main text:

$$EACVF_{tot}(bk) = \frac{\sqrt{\pi}\sigma a_h^2 (n_{max} - k)}{X}\left[\frac{\sigma_h^2}{a_h^2} + 1\right] \quad k = 0,1,2,...n_{max} - 1 \tag{a25}$$

where n_{max} is the number of SCs present in the mixture. Eq a25 differs from the analogous eq. a12 (or a13) only in the multiplicative term $\sqrt{\pi}\sigma/X$, that takes into account the peak shape (σ): the part of the $EACVF_{tot}$ accounting for the peak height distribution is independent of peak shape. Consequently, it is possible to splitt and separately calculate the effects of the peak shape and the peak height distribution (hypothesis #3).

The $EACVF_{tot}(kb)$ computed on the total signal can be expressed as a function of $EACVF_o(kb)$ and $EACVF_e(kb)$ separately computed on n odd and even terms of the series. In this case, based on the addictivity of $EACVF_e(kb)$ and $EACVF_{tot}(kb)$ for even k values, the eq. a17 assumes the following expression:

$$EACVF_{tot}(bk) = \frac{\sqrt{\pi}\sigma a_{o,h}^2 (n-k)}{X}\left[\frac{\sigma_{o,h}^2}{a_{o,h}^2}\right] + \frac{\sqrt{\pi}\sigma a_{e,h}^2 (n-k)}{X}\left[\frac{\sigma_{e,h}^2}{a_{e,h}^2}\right] = \tag{a26}$$

$$\frac{\sqrt{\pi}\sigma (a_{o,h}^2 + a_{e,h}^2)(n-k)}{X}\left[\frac{\sigma_h^2}{a_h^2} + 1\right] \qquad k = 0,2,4,...2n-2$$

Such a relationship is also based on the assumption that both the odd and even terms display the same peak abundance distribution described by peak height dispersion ratio, i.e., (eq. 6 in the main text)

$$\frac{\sigma_{o,h}^2}{a_{o,h}^2} \approx \frac{\sigma_{e,h}^2}{a_{e,h}^2} \approx \frac{\sigma_h^2}{a_h^2} \tag{a27}$$

The condition is usually met in real samples since the compound abundances generally follow the same distribution for even and odd terms, yielding a constant relative standard deviation for the height distribution (i.e. σ^2/a^2).

Moreover, eq. a18 — which shows that $EACVF_{tot}(kb)$ for odd k values contains only the cross correlation terms between the two sequences — assumes the following expression in the case of Gaussian peak shape and linearly dependent series:

$$EACVF_{tot}(bk) = \frac{\sqrt{\pi}\sigma 2(a_{o,h} \cdot a_{e,h})(n-k)}{X}\left[\frac{\sigma_h^2}{a_h^2}+1\right] \qquad k = 1,3,5,...2n-1 \qquad \text{(a28)}$$

Determination of n_{max}

A procedure has been developed for the computation of the terms of the homologous series, $n_{max} = 2n$, in order to make it more robust in the presence of odd/even prevalence of the terms of the series. It is based on the $EACVF_{tot}$ values at $\Delta t = bk$ for even k values, since they are related to the quantity $(a_{o,h}^2 + a_{e,h}^2)$ (eq. 7 and a26) and therefore are independent of peak abundance distribution of odd and even terms.

For two subsequent $EACVF_{tot}$ deterministic peaks at $\Delta t = kb$ for even k terms, the general equation of $EACVF_{tot}$ at $\Delta t = kb$ and $\Delta t = b(k+2)$ can be computed according to the following equations:

$$EACVF_{tot}(kb) = \frac{\sqrt{\pi}\sigma}{X}(a_{o,h}^2 + a_{e,h}^2)\left[\frac{\sigma_h^2}{a_h^2}+1\right](n-k) \qquad \text{(a29)}$$

$$EACVF_{tot}((k+2)b) = \frac{\sqrt{\pi}\sigma}{X}(a_{o,h}^2 + a_{e,h}^2)\left[\frac{\sigma_h^2}{a_h^2}+1\right](n-(k+2)) \qquad \text{(a30)}$$

These equations at $\Delta t = kb$ for even k contain the number n of odd or even terms of the series, that is $n = \frac{n_{max}}{2}$.

By dividing the $EACVF_{tot}$ values of the two subsequent peaks for even k the following equation is obtained:

$$\frac{EACVF_{tot}(kb)}{EACVF_{tot}((k+1)b)} = \frac{n-k}{n-(k+2)} \qquad \text{(a31)}$$

This is the basis for computing the n value:

$$n = \frac{EACVF_{tot}(kb)}{EACVF_{tot}((k+2)b)} + k \qquad \text{(a32)}$$

From it we obtain n_{max}

$$n_{max} = 2 \cdot n \qquad \text{(a33)}$$

It must be underlined that a correct estimation of n_{max} makes it possible to remove the approximation introduced in estimating R by using the approximate eq. 13 instead of the rigorous eq. 12: therefore by computing n_{max} by eq. a33 and introducing it in eq. 12 it is also possible to achieve a correct estimation of R from the $EACVF_{tot}$ values.

Under hypothesis #3 for both even and odd series, equation a32 is true $\forall k$. A more complex data treatment is required for different peak abundance distributions in each series and in view of the fact that peak height and peak position are dependent variables. In such a case, the $EACVF_{tot}$ should be modified to take into account the statistical dependence between peak height and peak position. As a first approximation, the n values were estimated at different k values and the mean value of the obtained results was considered. It is an approximate way of extending the applicability of eq. a32 to the cases where a correlation exists between peak height and position. With such a procedure the method robustness has been enhanced to achieve an accurate estimation of n_{max} .

References

[1A]W. Feller, in *An introduction to probability theory and Its Applications*, ed. J. Wiley & Sons, 1968, vol. I, p. 236.