Scaling stratigraphic events using extreme occurrences

V. Pawlowsky-Glahn
Universitat de Girona, Departament d'Informàtica i Matemàtica Aplicada
Campus Montilivi – P1
E-17071 Girona, Spain
vera.pawlowsky@udg.es

J. J. Egozcue
Universitat Politécnica de Catalunya, Departament de Matemàtica Aplicada III
Jordi Girona 1–3
E-08008 Barcelona, Spain
juan.jose.egozcue@upc.es

1. Introduction

The objective of scaling techniques addressed in this paper is the estimation of optimal relative positions in time of biostratigraphic events. Available observations are either first or last occurrences of n taxa or microfossil species α_j, $j = 1, \ldots, n$, in several wells or outcrop sections. The number of common appearances of pairs of taxa, (α_j, α_k), in a well is important and not the number of sampling sites. This is so, because no direct measurement of position in time is possible. But quantitative information can be organized in terms of observed frequencies as follows: Let $m_{(jk)}$ be the number of times in which the j-th and the k-th fossil species appear in the same well, m_{jk} the number of times in which α_j precedes α_k, and *vice-versa* for m_{kj}. Simultaneous occurrences are allocated at random. Then, $m_{(jk)} = m_{jk} + m_{kj}$ and the total number of pairs examined is $\sum_{j,k} m_{(jk)} = m$. Furthermore, given a pair of taxa (α_j, α_k), let be p_{jk} the probability of the event $\{\alpha_j \text{ precedes } \alpha_k\}$. Then, $p_{kj} = 1 - p_{jk}$, and $\sum_{j,k} p_{(jk)} = n(n-1)/2$ is the total number of possible pairs.

Assume now the first or last occurrence in time of each taxon α_j is a continuous random variable, T_j, $j = 1, \ldots, n$, with sample space the real line and pdf $f_j(t - \tau_j)$, where τ_j is an unknown location parameter. The time scale is relative, and therefore the origin is fixed arbitrarily at τ_n. Other parameters are assumed to be given, as usually not enough data are available for estimation. Nevertheless, it would be easy to extend the approach presented herein to include unknown parameters of scale. Also, $f_j(t - \tau_j)$ is not assumed to be of the same type for each T_j.

Knowledge of the $f_j(t - \tau_j)$, $j = 1, \ldots, n$, allows to determine the pdf of $T_j - T_k$. Note that $\{T_j - T_k \leq 0\}$ is equivalent to $\{\alpha_j \text{ precedes } \alpha_k\}$. Then, considering $m_{jk}/m_{(jk)}$ to be an

initial estimate of $p_{jk} = \mathrm{P}[T_j - T_k \leq 0]$, it should be possible to determine, at least in certain cases, the parameters of the considered distribution in such a way that $p_{jk} \approx m_{jk}/m_{(jk)}$.

This fact led to the original approach of assuming the T_j's to follow each a normal distribution with identical variance σ^2 and unknown mean or *shift parameter* τ_j (Agterberg, 1990). Under this assumption, estimation of τ_j's is pretty straightforward, as $T_j - T_k \sim \mathcal{N}(\tau_j - \tau_k, 2\sigma^2)$ and $(T_j - T_\ell) - (T_k - T_\ell) \sim \mathcal{N}(\tau_j - \tau_k, 4\sigma^2)$. The solution is found using the quantiles corresponding to observed frequencies of precedence mentioned above. Critical points of this approach are the following: (a) $T_j - T_k$ and $(T_j - T_\ell) - (T_k - T_\ell)$ do not lead in general to identical estimates for $\tau_j - \tau_k$; (b) the assumption of normality with identical variance is not consistent with the fact that distributions of taxa are assumed to be skewed in general (Agterberg, 1990), an assumption supported by empirical evidence. Therefore, in a first step, given that available information consists of first or last occurrences of taxa, extreme value distributions were considered as an alternative to the normal (Gil-Bescós et al., 1998), keeping the algorithm identical. This approach was not completely satisfactory with respect to point (a) and led to the maximum likelihood (ML) approach presented here.

2. Maximum Likelihood Approach

The problem under study can be described as a multinomial experiment in the following terms: consider the random variables M_{jk}, $j, k = 1, \ldots, n$, $j \neq k$, which represent the number of times α_j precedes α_k whenever a pair of taxa is observed. This experiment is repeated m times, and there are $n^2 - n$ such variables. Then, in a single experiment, $m_{(jk)}/m$ is the probability of selecting the pair (α_j, α_k) and, thus,

$$\mathrm{P}[M_{jk} = 1] = \frac{m_{(jk)}}{m} p_{jk}; \quad \mathrm{P}[M_{kj} = 1] = \frac{m_{(jk)}}{m} p_{kj} = \frac{m_{(jk)}}{m}(1 - p_{jk}).$$

The sum of these probabilities over all the possible pairs of different subscripts is one and, therefore, we have a multinomial experiment with, for m pairs examined,

$$\mathrm{P}[M_{jk} = x_{jk}; j, k = 1, \ldots, n; j \neq k] = \frac{m!}{\prod_{j \neq k}(x_{jk}!)} \prod_{j \neq k} \left(\frac{m_{(jk)}}{m} p_{jk} \right)^{x_{jk}}. \tag{1}$$

As stated in the introduction, p_{jk} can be expressed in terms of the pdf of $T_j - T_k$ and, thus, in terms of the location parameters τ_j, τ_k. In fact, if we call $G_{jk}(t; \tau_j, \tau_k)$ the cdf of $T_j - T_k$, then

$$p_{jk} = \mathrm{P}[T_j - T_k \leq 0] = G_{jk}(0; \tau_j, \tau_k); \qquad p_{kj} = 1 - G_{jk}(0; \tau_j, \tau_k).$$

Substituting this expression in (**??**) we obtain the likelihood function of a sample as

$$L(\tau_1, \ldots, \tau_{n-1}) = \frac{m!}{\prod_{j \neq k}(m_{jk}!)} \prod_{j \neq k} \left(\frac{m_{(jk)}}{m} G_{jk}(0; \tau_j, \tau_k) \right)^{m_{jk}}. \tag{2}$$

To maximize this expression—or for Bayesian estimation of $\tau_1, \ldots, \tau_{n-1}$—constants can be suppressed, thus expressing that neither m nor $m_{(jk)}$ have a direct influence on the ML estimation. In fact, for Bayesian estimation they have an influence only as normalizing constants. After taking logarithms in (**??**), for $s = 2, \ldots, n-1$, we compute the gradient

$$\begin{aligned}\frac{\partial \ln L(\tau_1, \ldots, \tau_{n-1})}{\partial \tau_s} &= \sum_{k=1}^{s-1} \left[m_{sk} \frac{\frac{\partial}{\partial \tau_s}\left(G_{sk}(0; \tau_s, \tau_k)\right)}{G_{sk}(0; \tau_s, \tau_k)} - m_{ks} \frac{\frac{\partial}{\partial \tau_s}\left(G_{sk}(0; \tau_s, \tau_k)\right)}{1 - G_{sk}(0; \tau_s, \tau_k)} \right] \\ &\quad + \sum_{j=s+1}^{n} \left[m_{js} \frac{\frac{\partial}{\partial \tau_s}\left(G_{js}(0; \tau_j, \tau_s)\right)}{G_{js}(0; \tau_j, \tau_s)} - m_{sj} \frac{\frac{\partial}{\partial \tau_s}\left(G_{js}(0; \tau_j, \tau_s)\right)}{1 - G_{js}(0; \tau_j, \tau_s)} \right], \end{aligned} \tag{3}$$

whereas for $s = 1$ the first summatory does not appear. Thus, (**??**) and (**??**), enable us to determine τ_j's by using gradient methods of optimization. To apply those, computation of $G_{jk}(0; \tau_j, \tau_k)$ and partial derivatives appearing in (**??**) are required.

As mentioned in the introduction, computing $G_{jk}(0; \tau_j, \tau_k)$ assuming normal distributions for T_j and T_k is straightforward, whereas assuming extreme value distributions this is only the case for iid Gumbel (type I) densities. Nevertheless, in all cases numerical methods have to be applied. We suggest to use Fourier transform (FT), in order to speed up the computing process. To do so, consider the pdf of $T_j - T_k$ to be

$$g_{jk}(z; \tau_j, \tau_k) = \int_{-\infty}^{+\infty} f_j(z + t - \tau_j) f_k(t - \tau_k) dt$$

where the convolution has been obtained assuming independence of the T_j's. Then, taking FT we obtain

$$g_{jk}^{\star}(\omega) = e^{-i\omega(\tau_j - \tau_k)} f_j^{\star}(\omega) \overline{f_k^{\star}(\omega)}, \tag{4}$$

where $(\ )^{\star}$ stands for FT. Back-FT of (**??**) allows a fast computation of

$$G_{jk}(0; \tau_j, \tau_k) = \int_{-\infty}^{0} g_{jk}(z) dz.$$

Gradient components can also be obtained following a similar procedure because

$$\frac{\partial}{\partial \tau_j} G_{jk}(0; \tau_j, \tau_k) = -\frac{\partial}{\partial \tau_k} G_{jk}(0; \tau_j, \tau_k) = \int_{-\infty}^{0} \frac{\partial}{\partial \tau_j} g_{jk}(z) dz$$

and

$$\left(\frac{\partial}{\partial \tau_j} g_{jk}(z) \right)^{\star} (\omega) = i\omega e^{-i\omega(\tau_j - \tau_k)} f_j^{\star}(\omega) \overline{f_k^{\star}(\omega)}. \tag{5}$$

Equations (**??**) and (**??**) show that it is enough to compute $f_j^{\star}(\omega)$, $j = 1, \ldots, n$, just once in the whole computation procedure. However, back-FT is carried out several times in each iteration of the gradient method of optimization.

Whenever enough samples are available, parameters of dispersion can be included in the model. For small sample sizes, one can compute the likelihood of the samples using the location parameters obtained, and repeat the process under different assumptions for the dispersion parameters or for different pdf's (*e.g.* different kind of extreme value distributions, normal, skew-normal). Repeating these steps would then lead to validate the more likely models.

3. Conclusions

Extreme occurrences of biostratigraphic events can be used to obtain an optimal relative time scale considering observations of precedence as realizations of a multinomial experiment. The multinomial model leads by definition to consistent estimates of the parameters of location τ_j. Thus, they are compatible in the relative time scale obtained. Probabilities of precedence of taxa are computed assuming different densities for the time of appearance of each taxon. Possible models for these densities are the normal, or any combination of extreme value distributions (Gumbel, Weibull, ...), which appear to be consistent with empirical evidence. Computation of maximum likelihood estimates of location parameters should be carried out numerically; gradient methods of optimization are appropriate. Also, Bayesian methods are allowed, once the likelihood expression (**??**) has been established. In these approach optimization techniques may be avoided but, as a counterpart, intensive Monte Carlo Simulation or integration techniques in a high dimensional space would be required.

REFERENCES

Agterberg, F. P. (1990). Automated Stratigraphic Correlation, Elsevier.

Gil-Bescós, E., Egozcue, J. J., Pawlowsky-Glahn, V. and Agterberg, F. P. (1998). An extreme value approach to scaling biostratigraphic events. In Proceedings of IAMG'98 — The fourth annual conference of the International Association for Mathematical Geology (eds A. Buccianti, G. Nardi, and R. Potenza) 767–772. De Frede Editore, Napoli (I).

RESUME

On veut estimer, sour une echèlle de temps relatif, la position de different microfossiles. On a des observations de precedence entre microfossiles. Nous utilisons un modèle multinomiel pour etablir la versemblance des données. Ce modèle admet des distributions temporèlles quelconques pour chaqu'n des microfossiles. De cette façon, on peut utiliser des distributions d'extremes comme alternative aux méthodes basés sur la distribution normale.