J. exp. mar. Biol. Ecol., 1976, Vol. 25, pp. 67–75; © North-Holland Publishing Company

ÉTUDE DE LA DÉCROISSANCE DES VALEURS PROPRES DANS UNE ANALYSE EN COMPOSANTES PRINCIPALES: COMPARAISON AVEC LE MODÈLE DU BÂTON BRISÉ

SERGE FRONTIER

O.R.S.T.O.M., 24 rue Bayard, Paris, et Station Zoologique, Villefranche sur Mer, France

Résumé: La décroissance des valeurs propres successives dans un certain nombre d'analyses en composantes principales portant sur 20 variables et 44 individus, est comparée à la décroissance des segments dans le modèle du 'bâton brisé', où une quantité fixe (la longueur du bâton) est divisée simultanément en S par S–1 points au hasard.

On constate que les deux premières valeurs propres sont situées au dessus du modèle, et indiquent sans doute l'existence de facteurs agissant sur les variables. La variance résiduelle semble se partager entre les axes restant suivant un modèle proche du bâton brisé; cependant l'ajustement n'est pas parfait, car la variance non due aux facteurs se répartit sur l'ensemble des axes y compris les deux premiers, et il est impossible de discerner dans ces derniers variance factorielle et variance aléatoire. Le modèle semble cependant suffisant pour permettre de déterminer empiriquement le nombre d'axes significatifs.

Abstract: The decrease of successive eigenvalues in some principal component analysis of 20 variates and 44 individuals, is compared with that of the 'broken stick' model, where a fixed length is broken at random into a number of segments.

In the real analysis, the two first eigenvalues stay distinctly above the model values, indicating the occurrence of factors acting on the variables. Residual variance seems partitioned between other axes, according to a model which approaches the broken stick. The adjustment is not perfect, since an unknown part of the random variance also contributes to the first eigenvalues. The model, however, allows an empirical determination of the number of significant vectors.

Dans un travail précédent (Frontier, 1974a, b) nous avions proposé de juger de la significativité des premiers vecteurs propres dans une analyse en composantes principales, en comparant la décroissance des valeurs propres successives avec celle donnée par un partitionnement 'au hasard' de la variance totale en un nombre de parties égal au nombre des variables de départ.

Le problème du partage au hasard 'moyen' d'une quantité fixe en S parties est connu sous le nom de problème du bâton brisé. Le modèle constitué par l'espérance mathématique d'un tel partage a été introduit par Barton & David (1956), puis repris par MacArthur (1957), Barton & David (1959), Pielou (1969) pour comparer la répartition d'un lot d'individus en S espèces zoologiques distinctes, à ce que donnerait la répartition au hasard du meme lot en S catégories exclusives.

Le raisonnement utilisé est le suivant. Nous partons d'une quantité fixe, que nous pouvons assimiler à un 'bâton' de longueur unité; nous la partageons en S segments par S–1 points au hasard. Si les S segments pouvaient être distingués *a priori* (par exemple, numérotation de gauche à droite le long du bâton), tous auraient

même espérance mathématique – à supposer que la fonction de répartition des S–1 points de partage soit uniforme le long du bâton. Par contre, si on réarrange les segments après un partage en les ordonnant par longueurs décroissantes (ce qui revient à les distinguer *a posteriori*), l'ordre sera conservé par la moyenne, et l'on obtiendra une décroissance théorique moyenne, qu'il est possible de calculer.

Soient en effet $l_1, l_2, \ldots, l_s$ les longueurs des segments rangées en ordre décroissant. Effectuons un changement de variable, en considérant les différences entre longueurs des segments successifs: ces nouvelles grandeurs ne sont pas classées.

Posons,

$$\left|\begin{array}{ll} d_0 & = l_S \\ d_1 & = l_{S-1} - l_S \\ d_2 & = l_{S-2} - l_{S-1} \\ & \vdots \\ d_{S-1} & = l_2 - l_1 \end{array}\right. \quad \text{d'ou} \quad \left|\begin{array}{ll} l_S & = d_0 \\ l_{S-1} & = d_0 + d_1 \\ l_{S-2} & = d_0 + d_1 + d_2 \\ & \vdots \\ l_1 & = d_0 + d_1 + d_2 + \ldots + d_{S-1} \end{array}\right.$$

En aditionnant membre à membre les égalités de droite on obtient:

$$1 = Sd_0 + (S-1)d_1 + (S-2)d_2 + \ldots + d_{S-1}$$

Le segment de longueur 1 est donc maintenant divisé en S parties de longueur $(S-i)d_i$, non classées et avec comme seule contrainte $\sum_i (S-i)d_i = 1$. L'espérance mathématique de chacune d'elles est $1/S$. On en déduit:

$$\left|\begin{array}{ll} \mathrm{E}(d_0) = \dfrac{1}{S^2} & \\ \mathrm{E}(d_1) = \dfrac{1}{S(S-1)} & \\ \mathrm{E}(d_2) = \dfrac{1}{S(S-2)} & \\ \vdots & \end{array}\right. \quad \text{d'où} \quad \left|\begin{array}{ll} \mathrm{E}(l_S) & = \dfrac{1}{S^2} \\ \mathrm{E}(l_{S-1}) & = \dfrac{1}{S^2} + \dfrac{1}{S(S-1)} \\ \mathrm{E}(l_{S-2}) & = \dfrac{1}{S^2} + \dfrac{1}{S(S-1)} + \dfrac{1}{(S-2)} \\ & \vdots \end{array}\right.$$

et généralement:

$$\mathrm{E}(l_j) = \frac{1}{S} \sum_{i=0}^{S-j} \frac{1}{j+i}$$

Le tableau I donne la valeur de ces termes, exprimés en pourcentages, pour S variant de 2 à 20[1]. Lors d'une analyse en composantes principales portant sur 20 variables, nous comparerons la décroissance des valeurs propres de la matrice des corrélations à celle réalisée dans la colonne de droite du tableau. Les premiers vecteurs seront considérés comme significatifs s'ils extraient nettement plus de variance que ne le prévoit le modèle aléatoire ainsi construit; au delà d'un certain rang on admettra

[1] Les auteurs cités plus haut ordonnent les espèces par abondances croissantes, et donnent:

$$\mathrm{E}(l_j) = \frac{1}{S} \sum_{i=1}^{j} \frac{1}{S+1-i}$$

TABLEAU I

Longueurs moyennes des segments, rangés par ordre décroissant, issus d'un partage en S d'une quantité égale à 100 conformément au modèle du 'bâton brisé'.

$S=$	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	75,00	61,11	52,08	45,67	40,83	37,04	33,97	31,43	29,29	27,45	25,86	24,46	23,23	22,12	21,13	20,23	19,42	18,67	17,99
	25,00	27,78	27,08	25,67	24,17	22,76	21,47	20,32	19,29	18,36	17,53	16,77	16,08	15,45	14,88	14,35	13,86	13,41	12,99
		11,11	14,58	15,67	15,83	15,61	15,22	14,77	14,29	13,82	13,36	12,92	12,51	12,12	11,75	11,41	11,08	10,78	10,49
			6,25	9,00	10,68	10,85	11,06	11,06	10,96	10,79	10,58	10,36	10,13	9,90	9,67	9,45	9,23	9,02	8,82
				4,00	6,11	7,28	7,93	8,28	8,46	8,51	8,50	8,44	8,34	8,23	8,11	7,98	7,84	7,71	7,57
					2,78	4,42	5,43	6,06	6,46	6,70	6,83	6,90	6,92	6,90	6,86	6,80	6,73	6,65	6,57
						2,04	3,35	4,21	4,79	5,18	5,44	5,62	5,73	5,79	5,82	5,82	5,81	5,78	5,74
							1,56	2,62	3,36	3,88	4,25	4,52	4,71	4,84	4,92	4,98	5,01	5,03	5,02
								1,23	2,11	2,75	3,21	3,56	3,81	4,00	4,14	4,25	4,32	4,37	4,40
									1,00	1,74	2,29	2,70	3,02	3,26	3,45	3,59	3,70	3,78	3,84
										0,83	1,45	1,93	2,30	2,60	2,82	3,00	3,15	3,26	3,34
											0,69	1,23	1,65	1,99	2,26	2,47	2,64	2,78	2,89
												0,59	1,06	1,43	1,73	1,98	2,18	2,34	2,47
													0,51	0,92	1,25	1,53	1,75	1,93	2,09
														0,44	0,81	1,11	1,35	1,56	1,73
															0,39	0,71	0,98	1,21	1,40
																0,35	0,64	0,88	1,09
																	0,31	0,57	0,79
																		0,28	0,51
																			0,25

que les vecteurs se partagent au hasard la variance résiduelle, ne décrivant qu'un bruit. Nous donnons ailleurs *loc. cit.* le résultat de ces comparaisons pour 20 analyses réalisées sur des ensembles de 18 à 20 variables mesurées en 44 points. Nous concluons alors que deux, parfois trois axes, au plus, doivent être considérés comme significatifs.

La comparaison entre le résultat d'une analyse réelle et le modèle se fait ici globalement: on compare deux profils. Blanc & Laurec (1976) contestent cette approche, et estiment que la comparaison entre une valeur propre et la taille relative d'un segment du bâton brisé devrait se faire de proche en proche: le pourcentage de variance résiduelle extrait par le $n^{\text{ième}}$ vecteur étant à comparer avec le premier segment d'un bâton brisé d'ordre $S-n+1$. Les résultats attendus sont alors différents.

En effet dans cette hypothèse, la première valeur propre étant extraite, la variance résiduelle se trouve partagée en $S-1$ segments (2e colonne de droite dans le tableau I); on en déduit la valeur attendue de λ_2. Puis, les deux premières valeurs propres réelles étant retranchées, ce qui reste de la variance est partagé en $S-2$ (comparaison avec la troisième colonne) et ainsi de suite. Les proportions successives déterminées par le modèle sont situées le long de la première ligne du tableau I, et sont différentes des progressions décrites par les colonnes (à titre d'exemple, les deux derniers segments du bâton d'ordre 3 sont entre eux dans un rapport de 27,78 à 11,11, soit 71,43 et 28,57 %, différents des rapports trouvés dans un bâton d'ordre 2 : 75 et 25 %).

Cette distinction peut sembler paradoxale au premier abord: pourquoi après extraction du segment le plus grand d'un bâton brisé d'ordre S, ce qui reste n'est pas divisé comme un bâton brisé d'ordre $S-1$? C'est que le partage d'une quantité fixe en S parties simultanément, ou bien progressivement (avec la contrainte que les segments extraits successivement soient dans un ordre décroissant) sont deux problèmes mathématiques différents. On retrouve ici la nécessité d'une précaution fondamentale dans les problèmes de détermination d'une distribution: définir parfaitement les conditions concrètes de détermination de la variable aléatoire et de la probabilité élémentaire. On sait que la méconnaissance de cette condition peut aboutir, dans le domaine des probabilités géométriques, à certains paradoxes[1].

Dans le cas du partitionnement d'une variance totale en un nombre de valeurs propres égal à celui des variables de départ, ce partage est déterminé d'emblée et simultanément, comme implication de la matrice des corrélations. L'algorithme qui extrait les valeurs propres procède successivement, mais ces dernières sont déterminées avant tout calcul. Le modèle 'progressif', donné par la première ligne du tableau I, semble illogique en la circonstance. Nous étaierons cette opinion par l'examen des résultats obtenus dans nos 20 analyses portant sur le zooplancton de la baie d'Ambaro (Frontier, 1974b).

[1] Par exemple, l'espérance mathématique de la longueur d'une corde d'un cercle de rayon 1 varie, selon qu'on détermine la corde en fixant une de ses extrémités et choisissant l'autre au hasard sur la circonférence; ou bien que l'on choisit au hasard entre 0 et 1 la distance de la corde au centre du cercle; ou encore que l'on choisit au hasard le milieu de la corde sur la surface intérieure au cercle (problème de Bertrand, in Borel *et al.*, 1960).

Les décroissances des valeurs propres dans les 10 analyses fondées sur 20 variables de départ seront comparées à plusieurs modèles: 1) bâton brisé 'simultané' appliqué à la variance totale divisée en 20 segments ou 2) à la variance résiduelle après soustraction de la première ou des deux premières valeurs propres; 3) 'bâton brisé progressif' c'est-à-dire pourcentages théoriques donnés par les segments no. 1 des bâtons d'ordre 20, 19, 18 . . .; 4) valeur attendue pour chaque segment, en fonction de la variance résiduelle réelle à chaque étape, et de la proportion théorique du segment no. 1 à l'étape correspondante (produit de la variance résiduelle après extraction des k premières valeurs propres, par la proportion donnée par le premier segment du bâton d'ordre $S-k$).

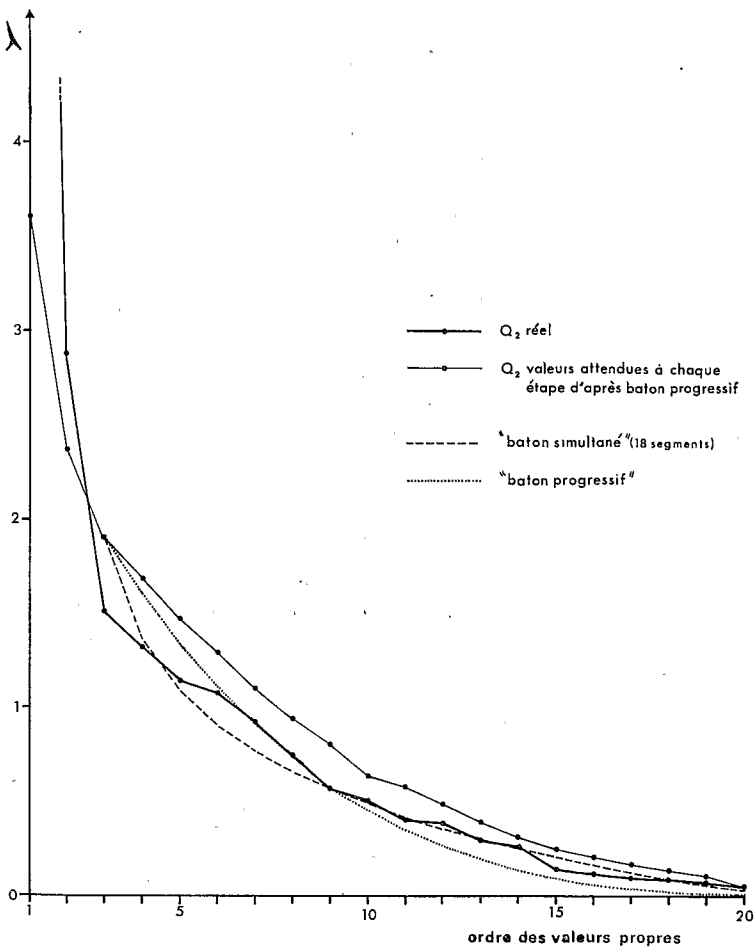

Fig. 1. Décroissance des valeurs propres dans une analyse en composantes principales (zooplancton de baie d'Ambaro: quadrillage Q_2), trait épais: calcul des valeurs attendues, à chaque étape, d'après la variance résiduelle réelle et la proportion donnée par le modèle 'bâton brisé progressif', trait fin: tireté, modèle 'bâton brisé simultané' appliqué à la variance subsistant après soustraction des deux premières valeurs propres: pointillé, modèle 'bâton brisé progressif'.

Nous donnons dans un travail antérieur la comparaison des données réelles avec le modèle (1) pour les 20 analyses (Frontier, 1974b). La figure 1 ci-après compare l'analyse du quadrillage Q_2 (trait épais) aux modèles (2) et (3) après élimination des deux premières valeurs propres (trait interrompu et trait pointillé respectivement), et au modèle (4) (trait fin). Les deux premières valeurs propres réelles se situent au dessus des modèles aléatoires: on en déduit que les axes 1 et 2 sont significatifs, extrayant une variance en partie due à des facteurs communs à un certain nombre de variables de départ. Les vecteurs 3 à 20 semblent se partager la variance résiduelle selon un modèle proche du bâton brisé 'simultané' (en fait, légèrement différent: noter en particulier la valeur basse de λ_3). Le modèle (4) donne toujours une valeur supérieure à la valeur trouvée. Nous n'avons représenté que pour mémoire le modèle (3) correspondant à la première ligne du tableau I, appliquée à la variance $\lambda_3+\ldots+\lambda_{20}$.

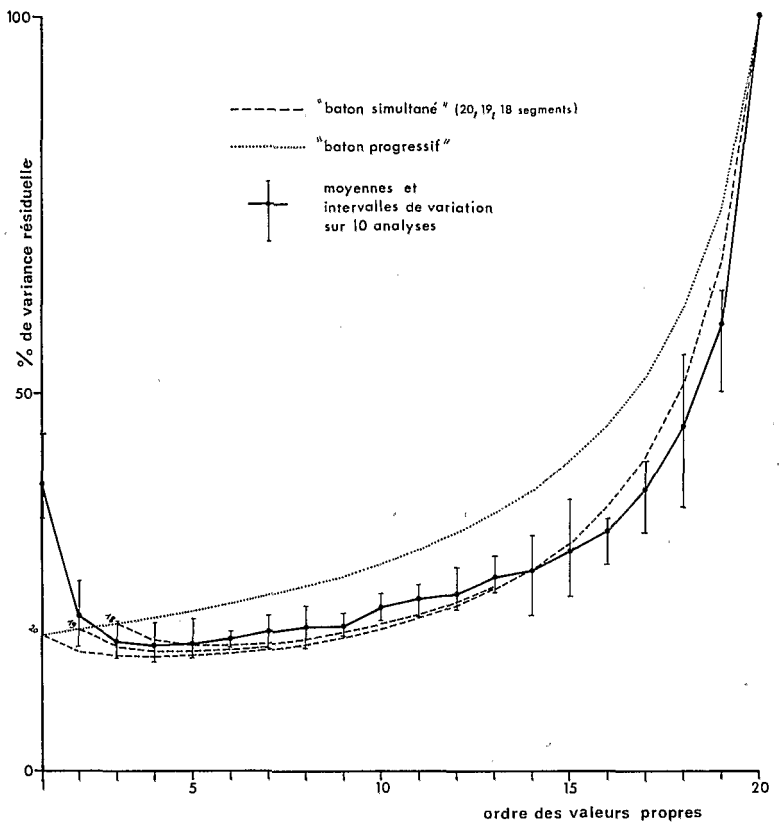

Fig. 2. Proportion de variance résiduelle représentée, après soustraction des n premières valeurs propres, par la valeur propre $(n+1)$: trait continu, dans 10 analyses portant sur 20 variables (moyennes, intervalles de variation); trait interrompu, dans le modèle du bâton brisé pour 20, 19, 18 segments; trait pointillé, dans le modèle 'progressif' (premiers segments des bâtons brisés d'ordre 20 à 2).

La comparaison peut également être faite sur la base des proportions de variance résiduelle à chaque étape, dans les différents cas (Fig. 2). Les résultats concernant les analyses réelles (résultats trouvés et modèle (4) peuvent donc être comparés d'une analyse à l'autre, et nous en indiquons la moyenne et l'intervalle de variation. Les résultats réels sont toujours situés nettement au dessous que ceux prévus par le modèle (4), qui ne convient visiblement pas. La succession des valeurs réelles ne s'éloigne pas considérablement de celle d'un bâton brisé 'simultané' d'ordre 19 ou 18: une ou deux valeurs propres seulement sont donc à attribuer à autre chose qu'à un partage au hasard de la variance totale. Néanmoins l'ajustement au modèle n'est pas parfait.

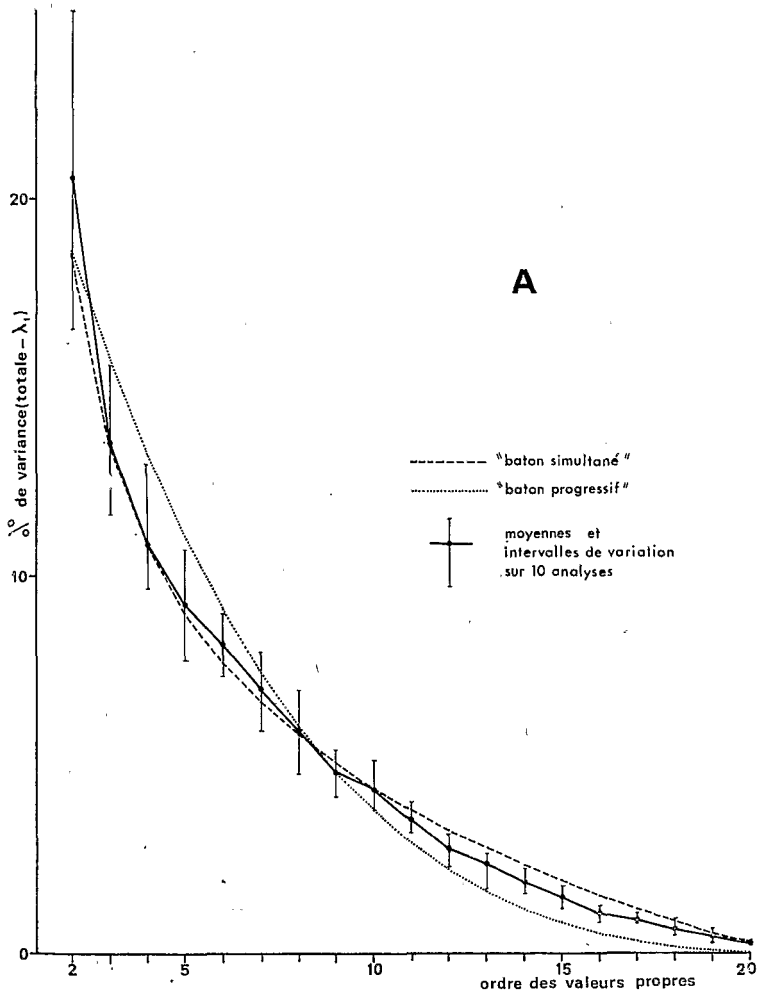

Fig. 3A.

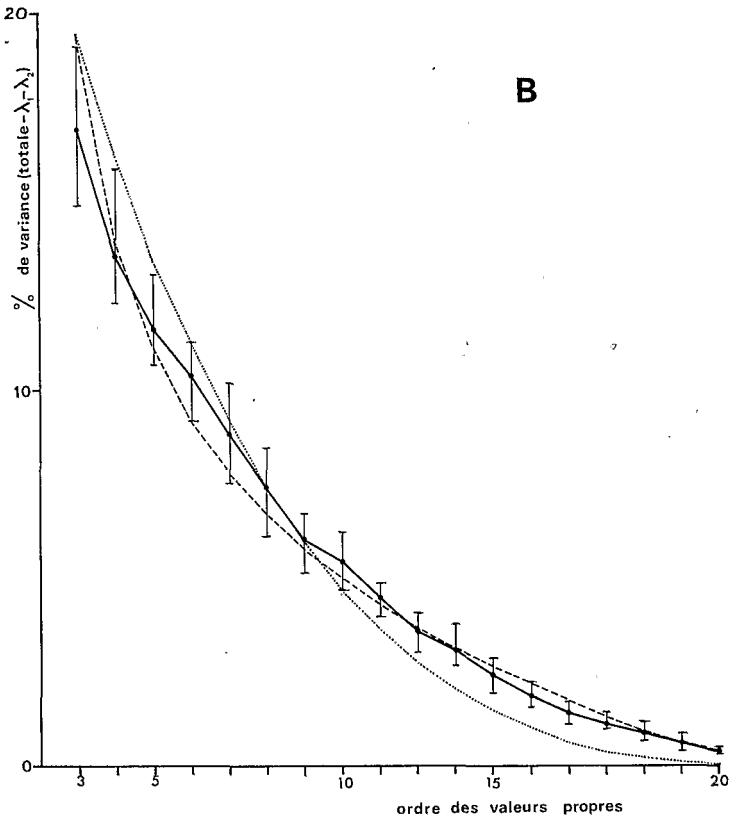

Fig. 3. Décroissance des valeurs propres; pourcentages de la variance subsistant après soustraction de la première valeur propre (A) ou des deux premières (B); trait continu, dans 10 analyses portant sur 20 variables (moyennes, intervalles de variation); trait interrompu, dans le modèle du bâton brisé d'ordre 19 (A) ou 18 (B); trait pointillé, dans le modèle 'progressif'.

La figure 3 représente la décroissance moyenne, pour les mêmes analyses, des valeurs propres, comparées aux modèles 2 (tiretés) et 3 (pointillés) appliqués à 19 variables (Fig. 3A) ou à 18 variables (Fig. 3B). Il semble que le meilleur ajustement au bâton brisé soit réalisé après extraction des deux premières valeurs propres plutôt que de la première seule. Néanmoins la première valeur propre n'est en général pas située très au dessus du modèle aléatoire, et montre de plus une forte variabilité: le deuxième axe semble ne pas être toujours significatif. Par ailleurs on constate qu'après soustraction des deux premières valeurs propres, la troisième est constamment au dessous du modèle, et se conformerait plutôt à un deuxième segment de bâton d'ordre 19 qu'à un premier segment de bâton d'ordre 18. Cela tendrait à indiquer que la variance qui se trouve partagée suivant le modèle du bâton brisé n'est ni la variance totale, ni celle-ci diminuée des deux premières valeurs propres, mais d'une quantité intermédiaire provenant de ce que, dans la variance représentée

par ces deux premières valeurs propres, une partie seulement est déterminée par les 'facteurs' supposés, l'autre étant aléatoire. Il est impossible d'estimer les proportions dans lesquelles s'effectue cette partition, et donc de proposer un modèle rigoureux.

Le modèle (4) n'a pas été porté sur la figure, car la valeur attendue pour une valeur propre dépend des étapes antérieures, ce qui donne au profil une variabilité excessive.

Il semble, en conclusion, que le modèle du bâton brisé dans sa conception initiale (partage simultané) convienne approximativement pour rendre compte du partage de la variance subsistant après soustraction des valeurs propres les plus élevées. Il convient mieux, en tous cas, qu'un modèle 'progressif'.

Intuitivement, on distingue autant de tendances mutuellement indépendantes dans le corps de données, qu'il y a de valeurs propres visiblement éloignées du modèle aléatoire. Un test serait bienvenu à ce niveau du raisonnement. Mais ce test ne peut être progressif et appliqué aux valeurs propres les unes après les autres: la partie de la variance non déterminée par des 'facteurs' est en effet partagée d'emblée et au hasard sur la totalité des axes. On ne comparera donc que des profils. Un test χ^2 serait applicable si l'on calculait la variance des segments dans le modèle – ce qui n'a pas encore été fait. Quoiqu'il en soit l'incertitude subsistera au moins pour la première valeur propre non trop éloignée du modèle.

Références

Barton, D. E. & F. N. David, 1956. Some notes on ordered random intervals. *J. roy. stat. Soc., ser. B*, Vol. 18, pp. 79–94.

Barton, D. E. & F. N. David, 1959. The dispersion of a number of species. *J. roy. stat. Soc., ser. B*, Vol. 21, pp. 190–194.

Blanc, F. & A. Laurec, 1976. De l'heuristique au thaumaturgique dans l'analyse des données en écologie marine. *Cah. O.R.S.T.O.M., sér. Océanogr.*, T. 14, sous presse.

Borel, E., R. Deltheil & R. Huron, 1960. *Probabilités, erreurs.* Collection Armand Colin, Paris, 220 pp.

Frontier, S., 1974a. L'analyse factorielle est-elle heuristique en écologie du plancton? *Cah. O.R.S.T.O.M., sér. Océanogr.*, T. 12, pp. 77–81.

Frontier, S., 1974b. Contribution à la connaissance d'un écosystème néritique tropical: étude descriptive et statistique du peuplement zooplanctonique de la région de Nosy Be (Madagascar). Thèse Univ. Marseille, multigr. 268 pp.

MacArthur, B. H., 1957. On the relative abundance of bird species. *Proc. natn. Acad. Sci. U.S.A.*, Vol. 43, pp. 293–295.

Pielou, E. C., 1969. *An introduction to mathematical ecology.* Wiley Interscience, New York, 285 pp.