Groupement automatique pour l'analyse du spectre singulier

Thomas MOREAU[1, 2], Laurent OUDRE[3, 2], Nicolas VAYATIS[1, 2]

[1]CMLA - ENS Cachan, CNRS - 61 Avenue du Président Wilson, 94230 Cachan

[2]COGNAC-G - Université Paris Descartes, CNRS - 45, rue des Saints-Pères, 75006 Paris

[3]L2TI - Université Paris 13 - 99 Avenue Jean Baptiste Clément, 93430 Villetaneuse

thomas.moreau@cmla.ens-cachan.fr, laurent.oudre@univ-paris13.fr, nicolas.vayatis@cmla.ens-cachan.fr

Résumé – Cet article introduit différentes stratégies de groupement automatique pour les composantes issues d'une analyse du spectre singulier (SSA). Cette étape, cruciale à la reconstruction de composantes haut niveau, permet de séparer les différentes dynamiques présentes dans un signal. De nouvelles stratégies sont comparées aux méthodes de la littérature dans un cadre d'évaluation unifié afin de mettre en lumière les spécificités de chacune.

Abstract – This paper introduce several automatic grouping strategies for Singular Spectrum Analysis (SSA) components. This step is useful to retrieve meaningful insight about the temporal dynamics of the series. A unifying framework is proposed to evaluate and compare the efficiency of different original methods compared to the existing one.

1 Introduction

L'analyse du spectre singulier (en anglais SSA : Singular Spectral Analysis) est une technique introduite par Vautard et al [6] permettant d'obtenir de manière simple une décomposition du signal en composantes de tendance, de saisonnalité (harmoniques) et de bruit. Elle est principalement utilisée en finance ou en météorologie où les séries étudiées sont souvent courtes et bruitées. De nombreuses publications ont montré la pertinence de la SSA dans ces contextes, où cette décomposition permet d'apporter de l'information sur les mécanismes produisant ces séries mais aussi de formuler des modèles de prédiction. Néanmoins, cette technique a été rarement utilisée sur d'autres signaux, principalement à cause de son manque d'automatisation. La pertinence des décompositions obtenues dépend en effet énormément du regroupement des composantes extraites, et résoudre cette tâche de façon manuelle peut s'avérer laborieux. Dans cet article, nous proposons un cadre permettant l'évaluation et la comparaison de différentes stratégies de regroupement automatique (existantes et originales).

Rappels sur la SSA. On considère un signal fini échantillonné $x = (x_1, \dots, x_T)$ et un entier $K \in \{1, \dots, T/2\}$. La matrice de K-trajectoire $X^{(K)} \in \mathbb{R}^{K \times L}$ de la série x est définie par :

$$X^{(K)} = \begin{bmatrix} x_1 & x_2 & \dots & x_L \\ x_2 & x_3 & \dots & x_{L+1} \\ \dots & & & \\ x_K & x_{K+1} & \dots & x_T \end{bmatrix} \tag{1}$$

avec $L = T - K + 1$. Cette matrice contient toutes les sous séries de longueur K de x. Une décomposition en valeur singulière (en anglais SVD pour Singular Value Decomposition) de la matrice de K-trajectoire $X^{(K)}$ de x permet d'obtenir une décomposition de la matrice de trajectoire sous la forme d'une somme de matrices de rang 1 :

$$X^{(K)} = \sum_{i=1}^{K} \lambda_i U_i V_i^T \tag{2}$$

où les U_i sont des motifs maximisant récursivement la covariance avec les sous séries de longueur K de x et les $\lambda_1 \geq \dots \geq \lambda_K$ sont les valeurs singulières associées. En faisant la moyenne le long des anti-diagonales, on peut transformer chaque matrice de trajectoire $\lambda_i U_i V_i^T$ en un signal temporel $x^{(i)}$, et une décomposition de x sous la forme $x = \sum_{i=1}^{K} x^{(i)}$ où $x^{(i)}$ est la composante associée au triplet propre (λ_i, U_i, V_i). L'utilisation de la SVD permet de mettre en évidence différents régimes de dépendance linéaire associés aux composantes de tendance, d'oscillation et de bruit de la série. Ceci la rend particulièrement adaptée pour les séries pseudo-périodiques [5].

Regroupement des composantes. Les oscillations harmoniques ou plus complexes, présentes dans la série initiale sont généralement décomposées sur plusieurs de ces composantes. Une étape d'identification et de regroupement est nécessaire pour produire une décomposition intéressante. Diverses approches ont été utilisées dans la littérature pour réaliser ce groupement, même si la plupart des travaux utilisant la SSA réalisent cette étape empiriquement [4, 5]. En particulier, certaines heuristiques basées sur la corrélation et le spectre de fréquence des composantes ont été développées pour aider au choix des groupes [5]. Ces indicateurs, d'abord utilisés pour le regroupement manuel, ont ouvert la voie au développement de méthodes de groupement automatisé [1, 3, 2]. Nous verrons dans la section 2 une description des principales méthodes existantes ainsi que de nouvelles stratégies, ainsi qu'un cadre unifié pour l'évaluation de ces méthodes dans la section 3. La section 4 présentera les résultats numériques de cette comparaison.

2 Méthodes

Dans cette section, nous allons présenter plusieurs méthodes de groupement automatique pour la SSA. Une description unifiée de celles-ci permet de les comparer dans un cadre simple, comme paramètres d'une stratégie globale.

2.1 Formulation générale

Les stratégies de regroupement peuvent être décrites en trois phases :

1. Sélectionner des composantes intéressantes de la SSA
2. Calculer une matrice d'adjacence entre ces composantes
3. Former les groupes I_j de composantes adjacentes

Les composantes finales sont obtenues en sommant les composantes regroupées $y_j = \sum_{i \in I_k} x^{(i)}$. La première étape est effectuée en supprimant les composantes de la SVD ayant une valeur singulière λ_i plus faible qu'un certain seuil (que l'on choisira ici adaptatif et de la forme $\tau_l.\lambda_2$ avec $\tau_l = 0.01$). On utilise la seconde valeur propre car la première, liée à la variance de la série, peut rejeter la plupart des composantes. La seconde étape utilise une mesure de similarité entre les composantes afin d'obtenir une matrice d'adjacence.Enfin, la troisième étape concerne la stratégie de formation des groupes à partir de la matrice d'adjacence. Cette stratégie permet de gérer la prévalence des composantes entre elles.

2.2 Mesures de similarité

2.2.1 Mesures basées sur la corrélation

Les composantes à regrouper sont celles qui sont produites par les mêmes phénomènes. Elles ne sont donc pas indépendantes. Il est donc intéressant d'observer la corrélation entre les composantes comme indicateur de similarité.

Corrélation (GG1). Deux composantes sont considérées comme adjacentes si la corrélation entre elles est plus haute qu'une valeur seuil ρ_c [1]. Pour éviter de grouper des composantes i, jd'importance trop éloignée, on impose de plus que le ratio de leur valeur singulière associée $\frac{\min(\lambda_i,\lambda_j)}{\max(\lambda_i,\lambda_j)}$ soit supérieur à un seuil $\rho_1 \in [0,1]$. Cela évite de grouper des phénomènes d'amplitude différentes ou du bruit. On définit donc la matrice d'adjacence $A = (a_{i,j})_{0 \leq i,j \leq K} \in \{0,1\}^{K \times K}$ par :

$$a_{i,j} = \begin{cases} 1 & \text{si } \dfrac{\min(\lambda_i,\lambda_j)}{\max(\lambda_i,\lambda_j)} \geq \rho_1 \text{ et } \mathrm{corr}(x^{(i)}, x^{(j)}) \geq \rho_c \\ 0 & \text{sinon} \end{cases} \tag{3}$$

W-Corrélation (GG3). La W-corrélation pour deux séries x, y de taille T et pour une longueur de fenêtre K est similaire à une corrélation classique mais utilise un produit scalaire pondéré $\langle x|y\rangle_w = \sum_{t=1}^{T} w_t x_t y_t$, $w_t = \min(t, T-t, K)$ qui permet de limiter les effets de bord. En théorie, deux séries sont séparables par la SSA si leur w-corrélation est nulle [5]. Il est donc possible d'utiliser la valeur de la w-corrélation comme indicateur de similarité entre deux composantes. On peut alors prendre une fonction de similarité identique à la précédente en remplaçant la corrélation par la w-corrélation.

2.2.2 Mesures basées sur le périodogramme

Le profil spectral des composantes obtenues peut aussi être utilisé comme un indicateur de similarité des séries [5]. En effet, les composantes contenant une même phase oscillante partagent des structures communes dans leur périodogramme qui peuvent permettre leur identification. On notera dans la suite $\Pi_x(k)$ le périodogramme d'une série x de longueur T.

Regroupement harmonique (HG). Une stratégie de regroupement efficace pour extraire les harmoniques exponentiellement modulées [2] définit une matrice d'adjacence qui regroupe deux composantes successives $(i, i+1)$ si

$$\frac{1}{2} \max_{0 \leq k \leq L/2} \left(\Pi_{U_i}(k) + \Pi_{U_{i+1}}(k)\right) \geq \rho_0 \tag{4}$$

pour $\rho_0 \in [0,1]$. Deux composantes $i, i+1$ représentant la même sinusoïde pure auront un même périodogramme ce qui donnera une valeur de cet indicateur proche de 1. Au contraire, si deux composantes successives ont des spectres disjoints, l'indicateur ne peut dépasser 0.5. Cette métrique rend compte du fait que les composantes successives ont des pics de densité spectrale aux mêmes fréquences.

Similarité des pics (GG2). Une autre métrique de similarité basée sur les périodogrammes est définie en prenant en compte la distance ℓ_∞ entre les fréquences non négligeables du périodogramme [1]. Pour une série x, on définit l'ensemble ordonné $F_x = \{k | \Pi_f(k) \geq \rho_p \|\Pi_f\|_\infty\}$ pour $\rho_p \in [0,1]$ et on notera $F_x(h)$ la h-ème valeur de cette ensemble. On définit alors la matrice d'adjacence par :

$$a_{i,j} = \begin{cases} 1 & \text{si } \dfrac{\min(\lambda_i,\lambda_j)}{\max(\lambda_i,\lambda_j)} \geq \rho_1 \\ & \text{et } \frac{|F_{x^{(i)}}(h) - F_{x^{(j)}}(h)|}{T/2} \leq \rho_2, \forall h \in \{1,\ldots,m\} \\ 0 & \text{sinon} \end{cases} \tag{5}$$

où $m = \min\left(|F_{x^{(i)}}|, |F_{x^{(j)}}|\right)$. Cette métrique rend plus robuste la détection de composantes ayant le même périodogramme en comparant non plus le pic maximal (HG) mais les supports spectraux des composantes.

Regroupement support harmonique (HGS). L'une des faiblesses de la similarité (HG) est qu'elle se concentre sur des composantes harmoniques. Ainsi, si les composantes extraites par la SSA ont une plage de fréquence plus large, cela peut nuire au groupement car la normalisation du spectre fera passer la métrique de similarité sous le seuil de regroupement.

Nous proposons pour pallier ce problème de considérer le support fréquentiel F_{U_i} des motifs associées au composantes en définissant leur fréquence fondamentale comme le centre de ce support. On sélectionne alors les composantes ayant un support de largeur inférieure à un seuil $\rho_s \in [0,1]$ et on considère que les composantes i, j sont adjacentes si l'écart entre leurs fréquences fondamentales est inférieur à $\rho_f \in [0,1]$. Ceci permet donc de grouper les composantes associées à des dictionnaires ayant des périodogrammes peu étalés et qui se chevauchent.

K-moyennes (KM). On peut aussi utiliser un algorithme de K-moyennes appliqué aux périodogrammes pour former le groupement [3]. Ceci est intéressant car on compare la distance euclidienne entre les périodogrammes et le choix des seuils est automatiquement fourni par les K-moyennes. Le choix du nombre de groupes C est un choix critique ici. Une estimation du nombre de groupe peut être faite en estimant le rang de la matrice $\Pi = \begin{bmatrix}\Pi_{U_1} & \dots & \Pi_{U_K}\end{bmatrix}^T$. On calcule les valeurs singulières $(\sigma_1 \geq \dots \geq \sigma_K)$ de Π et on considère le rang C comme le premier indice i tel que la valeur singulière σ_i soit inférieure à un seuil de la forme $\rho_r.\sigma_1$ avec $\rho_r \in [0,1]$. Ceci est intéressant car cela permet d'avoir une idée du nombre de composante indépendantes au sens du spectre de Fourier. Deux composantes sont adjacentes si elles appartiennent au même groupe formé par l'algorithme des K-moyennes.

2.3 Stratégies de formation des groupes

Méthode uniforme (MU). La stratégie de formation des groupes à partir de la matrice d'adjacence la plus simple est de considérer que toutes les composantes ont le même intérêt. On peut alors considérer que deux composantes sont dans le même groupe dès lors qu'elles sont adjacentes. Cette méthode est utilisée dans la plupart des stratégies qui ont été proposés jusque là [1, 3].

Méthode hiérarchique (MH). Nous proposons une autre méthode de formation des groupe donnant une importance relative au différentes composantes. L'importance de chaque composante peut être mesurée par la part de variance expliquée par cette composante. On peut donc considérer qu'on ajoute une composante dans un groupe si elle est adjacente à la composante dominante du groupe (au sens de la variance). Cela permet de donner un poids plus important lors du groupement aux composantes les plus intéressantes.

3 Méthode d'évaluation

3.1 Génération de signaux

Les stratégies de regroupement ont été évaluées sur des signaux artificiels générés aléatoirement. L'objectif du regroupement est de séparer la tendance des différentes composantes périodiques et du bruit. Les signaux tests sont échantillonnés à 100Hz selon le modèle :

$$f(t) = \underbrace{b_0 t^p}_{c_0} + \sum_{i=1}^{K} \underbrace{b_i e^{-\alpha_i t}\sin\left(2\pi f_i t + \phi_i\right)}_{c_i} + \epsilon_t \qquad (6)$$

où ϵ_t est une composante de bruit blanc gaussien d'écart type $\sigma = s.\sigma_f$. Les différents paramètres sont choisis de manière aléatoire avec $b_i \in [0,1]$, $p \in [0,5]$, $\phi_i \in \left[\frac{-\pi}{2}, \frac{\pi}{2}\right]$, $\alpha_i \in \left[0, \frac{1}{2}\right]$, $f_i \in [0,50]$Hz et $s \in [0,40]$dB.

Trois classes de signaux sont définies. La première classe fixe $b_0 = \alpha_i = 0$ (harmoniques + bruit) , la seconde $\alpha_i = 0$ (tendance + harmoniques + bruit) et la troisième correspond au modèle général (6).

3.2 Métriques d'évaluation

L'une des métriques utilisée pour évaluer la qualité du groupement est le coefficient de détermination r^2 [1]. Ce coefficient rend compte de l'erreur commise par rapport à la variance de la composante. Pour une série x de moyenne $\overline{x}$ et pour un estimateur $\widehat{x}$, le coefficient de détermination est calculé suivant $r^2(x,\widehat{x}) = 1 - \frac{\|x-\widehat{x}\|^2}{\|x-\overline{x}\|^2}$. Cette quantité correspond au rapport entre l'erreur d'estimation et la variance d'un signal.

On calcule alors le rappel et la précision du regroupement pour r^2

$$R = \frac{1}{N}\sum_{i=1}^{N} \min_{1\leq j\leq M} r^2(c_i, y_j) \quad P = \frac{1}{M}\sum_{j=1}^{M} \min_{1\leq i\leq N} r^2(c_i, y_j) \qquad (7)$$

où les $(y_j)_{1\leq j\leq M}$ sont les différentes composantes obtenues lors du groupement et les $(c_i)_{1\leq i\leq N}$ sont les composantes initiales formant le signal artificiel.

Nous proposons de fusionner ces deux concepts est de considérer le score obtenu pour une allocation optimale entre les composantes recherchées et celles formées par le regroupement :

$$S = \frac{1}{N}\min_{\sigma\in\mathfrak{S}(M)} \sum_{i=1}^{N} r^2(c_i, y_{\sigma(i)}) \qquad (8)$$

où $\mathfrak{S}(M)$ est le groupe des permutations de $\{1,\dots,m\}$. La qualité du regroupement obtenu dépend fortement de la qualité des composantes de départ. Il est donc important de considérer non pas les valeurs brutes de ces scores mais l'amélioration apportée par le groupement. Pour cela, on calcule les métriques R, P et S pour le groupement évalué et pour les composantes avant groupement R_0, P_0, S_0. Le taux d'augmentation entre ces deux scores $\frac{R-R_0}{1-R_0}$ permet de mesurer l'apport d'un groupement par rapport aux composantes de départ. Ces métriques sont moins sensibles aux variations produites par une mauvaise décomposition de départ. On les notera respectivement R_r, P_r et S_r.

4 Résultats

Pour toutes les expériences, les composantes de la SSA sont calculées avec une valeur de fenêtre $K = T/2$. Les regroupements sont ensuite réalisés avec les paramètres $\rho_0 = 0.8$, $\rho_1 = 0.8$, $\rho_2 = 0.05$, $\rho_c = 0.8$, $\rho_p = 0.8$ et $\rho_r = 0.4$ fixés comme ceux utilisés dans les publications originales [1, 3]. Pour nos similarités (HGS, GG3), les paramètres $\rho_f = 0.001$ et $\rho_s = 0.6$ ont été fixés empiriquement sur quelques exemples.

La Figure 1 présente un résultat de décomposition par la SSA d'un signal de la classe 3 composé d'une tendance, de quatre harmoniques exponentiellement modulées et d'un bruit blanc gaussien (SNR de 9.7dB). On observe que la décomposition avec un groupement (HG)-(MH) permet de mettre en évidence 4 des 5 composantes, qui sont retrouvées avec une faible déformation. Le coefficient de détermination r^2 moyen est de 0.96 pour ces 4 composantes alors que sans le groupement, le coefficient moyen est de 0.21. La cinquième composante harmonique

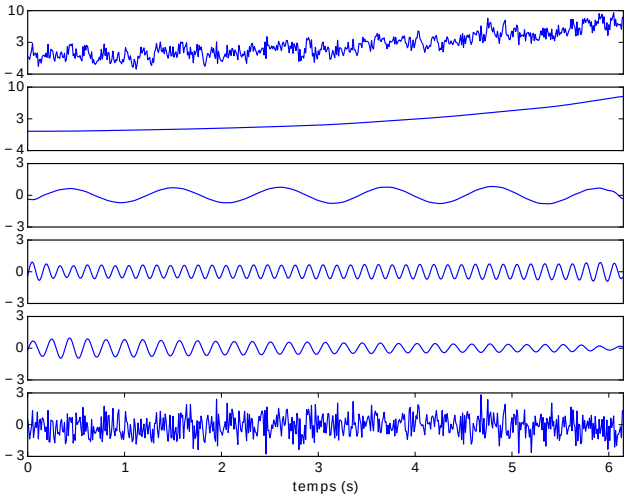

FIGURE 1 – Décomposition d'un signal artificiel (*haut*). Les quatre composantes du milieu correspondent aux groupes formés après la SSA par (HG)-(MH). Le signal du bas correspond au résiduel de la décomposition. On notera que les échelles des quatre signaux du bas ont été adaptées pour ne pas écraser les signaux.

Méthodes	GG1	GG2	GG3	HG	KM	HGS
R_r - MU	0.426	0.33	0.436	0.402	0.395	0.475
R_r - MH	0.427	0.367	0.437	**0.484**	0.396	0.459
P_r - MU	0.338	0.329	0.351	0.693	0.804	0.409
P_r - MH	0.339	0.347	0.351	0.705	**0.804**	0.439
S_r - MU	0.361	0.373	0.373	0.567	0.639	0.422
S_r - MH	0.361	0.389	0.373	0.592	**0.639**	0.437

TABLE 1 – Évaluation des différentes méthodes de groupement moyennée sur l'ensemble des signaux de la base de test. Les 3 premières lignes présentent la stratégie de formation (MU) et les 3 suivantes la stratégie (MH).

est rejetée comme du bruit par le filtrage des composantes initiales car son amplitude est faible par rapport aux autres composantes. Il est en effet difficile de retrouver les composantes avec une amplitude faible devant le bruit.

Les méthodes ont été testées sur 1000 signaux pour chacune des 3 classes, et les résultats moyennés sont reportés dans le Tableau 1. Pour la métrique R_r calculée sur l'ensemble de la base de signaux tests (3 classes), l'utilisation de la méthode (HG)-(MH) donne un résultat statistiquement meilleur (au sens du test de Friedman) que les autres méthodes avec une amélioration moyenne de 48.4%. Pour P_r et S_r, la méthode (KM) est statistiquement meilleure que les autres. Elle permet d'obtenir une amélioration moyenne de P_r de 80.4% et de 63.9% pour S_r. L'estimation précise du nombre de composantes dans (KM) explique la précision largement meilleure que les autres, ce qui se traduit aussi dans S_r. Pour toutes les similarités, l'utilisation de la stratégie (MH) introduite dans cet article permet une amélioration des résultats. Cet effet est assez faible pour les performances des méthodes basées sur la corrélation et pour (KM). Pour les méthodes basées sur l'étude du périodogramme autre que (KM), l'utilisation de (MH) améliore les résultats sauf pour le rappel de (HGS). Cependant, dans ce cas, les résultats des deux méthodes sont statistiquement équivalents. Il est donc intéressant d'utiliser cette stratégie pour le groupement automatique.

On observe pour la classe 1 que (HG)-(MH) est statistiquement équivalente en rappel à (HGS)-(MU) et (HGS)-(MH). Pour les deux autres métriques, (HG)-(MH) et (HGS)-(MH) sont équivalentes et donnent un meilleur score que (KM). Ces deux similarités sont particulièrement adaptées à la reconstruction d'harmoniques pures ce qui explique leurs performances sur la classe 1. Pour des composantes au spectres très concentrés, elles donnent des résultats similaires ce qui explique leur proximité. Pour les classes 2 et 3, (KM) est équivalente à (HGS)-(MH) pour R_r et statistiquement meilleure que les autres méthodes pour les deux autres métriques.

L'analyse des résultats sur chaque type de composante explique en parti ces observations. Le coefficient de détermination r^2 moyen obtenu pour la composante c_0 de tendance dans les classes 2 et 3 est en probabilité plus grand avec la méthode (KM) que pour les autres méthodes. Cette méthode observe l'intégralité du contenu fréquentiel pour effectuer le groupement ce qui est un avantage pour retrouver la composante de tendance qui peut avoir un périodogramme étalé. Pour les composantes harmoniques $(c_i)_{i>0}$ dans les classes 1 et 2, l'alliance de nos deux stratégies (HGS)-(MH) donne statistiquement un meilleur score r^2 que les autres méthodes exceptée (HG)-(MH) qui lui est équivalente. Ces composantes sont particulièrement adaptées à une similarité basée sur leurs périodogrammes car ils ne comporte qu'un pic. Pour les composantes exponentiellement modulées, toutes les méthodes donnent des résultats statistiquement équivalents et en dessous des ceux produit pour les composantes harmoniques. Les différentes méthodes ont donc du mal à distinguer ce genre de composantes.

Le cadre général introduit pour la comparaison des stratégies de regroupement permet de voir que la stratégie de regroupement à choisir est dépendante de l'importance donnée à l'extraction de chaque type de composantes. La méthode (KM) apparaît comme la plus adaptative et permet une estimation précise du nombre de composantes. L'utilisation de la stratégie (MH) proposée pour la formation des groupes permet une légère amélioration des performances. Ces résultats prometteurs sont en cours d'adaptation pour l'extraction de phases oscillantes dans des signaux physiologiques d'oculométrie.

Références

[1] N.V. Abalov and V.V. Gubarev. Automated grouping of decomposition components of time series for singular spectrum analysis. In *Proceedings of the 9th International Forum on Strategic Technology (IFOST), Chittagong, India*, 2014.

[2] Th. Alexandrov and N. Golyandina. Automatic extraction and forecast of time series cyclic components within the framework of ssa. In *Proceedings of the Workshop on Simulation, St. Petersburg , Russia*, pages 45–50, 2005.

[3] A.M. Alvarez-Mesa, C.D. Acosta-Medina, and G. Castellanos-Dominguez. Automatic singular spectrum analysis for time-series decomposition. In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium*, pages 131–136, 2013.

[4] M. Ghil, M.R. Allen, M.D. Dettinger, K. Ide, D. Kondrashov, M.E. Mann, A.W. Robertson, A. Saunders, Y. Tian, and F. Varadi. Advanced spectral methods for climatic time series. *Reviews of geophysics*, 40(1) :3–1, 2002.

[5] N. Golyandina, V. Nekrutkin, and A.A. Zhigljavsky. *Analysis of time series structure : SSA and related techniques*. CRC Press, 2010.

[6] R. Vautard and M. Ghil. Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series. *Physica D : Nonlinear Phenomena*, 35(3) :395–424, 1989.