Une ou deux composantes : la réponse de l'analyse spectrale singulière

Jinane HARMOUCHE[1], Dominique FOURER[2], François AUGER[2], Pierre BORGNAT[1], Patrick FLANDRIN[1]

[1]Laboratoire de Physique, CNRS - École Normale Supérieure de Lyon

[2]LUNAM Université, Université de Nantes, IREENA, Saint-Nazaire

{jinane.harmouche,patrick.flandrin,pierre.borgnat}@ens-lyon.fr,
{dominique.fourer,francois.auger}@univ-nantes.fr

Résumé – L'analyse spectrale singulière (ou SSA, pour *Singular Spectrum Analysis*), est une technique de décomposition d'un signal en composantes périodiques, tendance polynomiale et bruit. Cet article présente une évaluation du pouvoir de séparation entre composantes stationnaires et non-stationnaires de cet algorithme. Il propose également une solution nouvelle à la sélection automatique des valeurs singulières qui permettent d'obtenir ces composantes.

Abstract – The singular spectrum analysis (SSA) expands a signal into periodic components, trend and noise. This paper first addresses the separability through the SSA of non-stationary components. A second objective is to propose a new solution to the automatic selection of the singular values that provide these components.

1 Introduction

L'analyse d'un signal à l'aide du spectre des valeurs singulières de sa "matrice de trajectoire" (ou SSA, pour *Singular Spectrum Analysis* [1, 2]) est une technique relativement récente [3] dont les caractéristiques sont les suivantes :

- elle est non paramétrique et purement dirigée par les données, car la description obtenue est entièrement déduite du signal analysé et ne s'appuie pas sur un modèle sous-jacent auquel le signal est supposé se conformer ;
- elle est globale, car elle s'appuie sur la décomposition en valeurs et vecteurs propres d'une "matrice de covariance" déduite de la totalité du signal analysé ;
- elle décompose le signal en une somme de composantes classées par niveau d'énergie décroissant ;
- elle n'a qu'un seul paramètre de réglage, la taille de la matrice de covariance (notée par la suite L), dont le choix est bien évidemment déterminant ;
- elle doit être associée à une procédure d'identification des composantes "significatives" qui forment l'espace signal, les autres composantes formant l'espace bruit.

De nombreuses publications (partiellement recensées dans [4]) ont démontré la possibilité d'utiliser avec succès cette technique dans des champs d'application très variés (météorologie, climatologie, océanographie, astronomie, économie, médecine ...) pour résoudre des problèmes de natures très diverses (analyse exploratoire, débruitage, prédiction, interpolation, estimation de paramètres, détection de défauts et ruptures ...).

Ce travail a été réalisé avec le soutien de l'ANR au titre du projet ASTRES, ANR-13-BS03-0002-01.

Dans le cadre de la théorie du signal, cette méthode a par contre fait l'objet d'assez peu de publications, bien qu'elle partage avec d'autres méthodes des buts communs. C'est en particulier le cas pour la séparation d'un signal en composantes non stationnaires, question qui a été récemment abordée par Décomposition Modale Empirique (ou EMD) [5] et par "synchrosqueezing" [6]. Un des objectifs de cet article est de contribuer à cette même question sous l'angle complémentaire de la SSA. Un deuxième objectif est de proposer une solution nouvelle à la sélection automatique de ces dernières.

Le principe de la SSA est rappelé en Section 2, ainsi que le lien avec d'autres méthodes d'analyse des signaux. La question de la séparabilité est abordée en Section 3, sur la base de modèles de signaux stationnaires ou non, en mode supervisé. Le passage à un mode non supervisé par une technique de classification ascendante hérarchique est discuté en Section 4.

2 Quelques rappels sur l'analyse spectrale singulière

La SSA permet de décomposer un signal en la somme d'un nombre réduit de composantes grâce à la décomposition en valeurs singulières (SVD) d'une matrice spécifique construite à partir des données. L'algorithme comporte deux étapes :

- **décomposition** : cette première étape correspond à effectuer la SVD d'une matrice de Hankel appelée "matrice de trajectoire". Celle-ci est construite en associant à un signal $\mathcal{S}$ formé de N échantillons s_n $K = N - L + 1$ vecteurs-colonne de dimension L, le $k^{\text{ème}}$ vecteur étant donné par $\vec{s}_k = (s_k \ s_{k+1} \ ... \ s_{k+L-1})^T$. La matrice de

trajectoire (de dimension $L \times K$) s'en déduit alors par $X = [\overrightarrow{s_1}: \ ... \ :\overrightarrow{s_K}]$. Par SVD ([2], p. 219), cette matrice de rang $r \leq L$ se décompose en r matrices élémentaires X_i de rang 1. Soient σ_1, ..., σ_r les valeurs singulières non nulles de X classées dans l'ordre décroissant et $(U_1, V_1), ..., (U_r, V_r)$ les vecteurs singuliers gauche et droit associés. Les σ_i^2 sont aussi les valeurs propres de la matrice de covariance $C = X\,X^T$, associées aux vecteurs propres U_i, avec $V_i = X^T\,U_i/\sigma_i$. On obtient alors $X = \sum_{i=1}^{r} X_i$, avec $X_i = \sigma_i\,U_i\,V_i^T$.

– **reconstruction** : des signaux $\mathcal{X}_1$, ..., $\mathcal{X}_r$ de longueur N sont construits en moyennant les éléments des anti-diagonales des matrices X_1, ..., X_r. Un regroupement en composantes est ensuite effectué, en se référant généralement à l'analyse des valeurs singulières et/ou du comportement des vecteurs propres et/ou du taux de corrélation entre les $\mathcal{X}_i$. Il en résulte m composantes $\mathcal{Y}_j$, chacune étant associée à un groupe I_j :

$$\mathcal{S} = \sum_{j=1}^{m} \mathcal{Y}_j \quad \text{avec} \quad \mathcal{Y}_j = \sum_{i \in I_j} \mathcal{X}_i. \tag{1}$$

En tant que méthode basée sur la structure propre d'une certaine matrice de covariance construite sur les données, la SSA partage à la fois des principes et des propriétés avec d'autres méthodes d'estimation spectrale de type sous-espace (c'est en particulier le cas pour celle dite "de Cadzow" [7] dont on peut montrer que la SSA est essentiellement la première itération [8]). Cet algorithme est ainsi particulièment efficace pour identifier des signaux sinusoïdaux (éventuellement modulés en amplitude par une exponentielle) ou des fonctions polynômiales du temps, mais ces composantes doivent pour cela être clairement distinguées, ce qui conduit à la notion de séparabilité au sens de la SSA. Suivant [2], on conviendra que, dans le cas où $m = 2$, les composantes $\mathcal{Y}_1$ et $\mathcal{Y}_2$ de la décomposition (1) sont séparées si (*i*) elles sont orthogonales et (*ii*) les valeurs singulières des groupes I_1 et I_2 sont différentes.

3 Une étude spécifique de séparabilité

L'étude de la séparabilité peut être conduite de plusieurs manières, par exemple en discutant du rôle de L à modèle de signal donné. D'après [9], le meilleur choix de L lorsque N petit est $N/2$. Nous adopterons ici une perspective différente, en faisant varier les paramètres du modèle avec L fixé. Deux cas seront considérés, l'un stationnaire et l'autre non, en supposant dans un premier temps que la caractérisation est faite en mode supervisé.

3.1 Cas stationnaire

Le modèle stationnaire choisi est

$$s_n = \cos(2\pi\lambda_0 n) + H\,\cos(2\pi\lambda_1 n + \phi) + \sigma\epsilon_n, \tag{2}$$

avec ϵ_n un bruit blanc gaussien centré de variance unité, $N = 100$, $L = 50$, $\lambda_0 = 0.1$, $\lambda_1 \in]0, 0.1]$, $H \in [0.01, 100]$, $\phi = \frac{\pi}{2}$

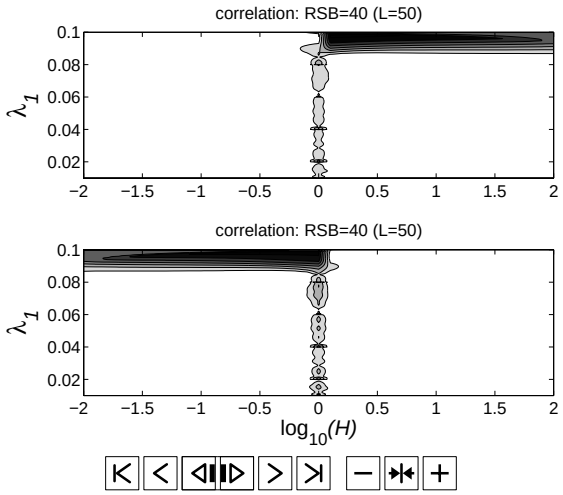

FIGURE 1 – Coefficient de corrélation entre les composantes d'origine x_1 (en haut) et x_2 (en bas) et celles reconstruites $\mathcal{Y}_1$ et $\mathcal{Y}_2$. Le blanc indique une corrélation totale et le noir une décorrélation totale.

et σ tel que le rapport signal-sur-bruit (RSB) varie de -14 à $+40$ dB. Lorsque l'algorithme SSA les distingue, chaque composante sinusoïdale conduit à deux valeurs singulières identiques, les autres étant liées au bruit. Deux composantes $\mathcal{Y}_1$ et $\mathcal{Y}_2$ associées aux deux premières paires de valeurs singulières sont alors reconstruites, et la qualité de la reconstruction est mesurée en calculant la corrélation entre $\mathcal{Y}_j$ et la sinusoïde qui lui correspond. Le résultat est montré sur la figure 1 pour $L = N/2$ et $L = N/4$. dans le cas supervisé. À fort RSB (40 dB), la SSA permet de séparer les deux sinusoïdes sauf dans deux situations : (*i*) quand $\lambda_0 \approx \lambda_1$ (fréquences très proches), on n'obtient qu'une seule paire significative de valeurs singulières et (*ii*) quand $H = 1$ (amplitudes égales), la condition de valeurs propres distinctes n'est pas respectée. La présence de bruit rajoute aux valeurs singulières significatives un fond continu dont le niveau augmente lorsque le RSB décroît, limitant de fait l'identification correcte aux situations où le RSB est positif.

3.2 Cas non-stationnaire

Le modèle non stationnaire choisi est le suivant :

$$s_n = \cos(2\pi\lambda_0 n) + H\cos(\varphi(n)),\ n = 1, ..., N \tag{3}$$

$$\text{avec } \varphi(n) = 2\pi\left(\lambda_1\, n + (\delta\lambda/2N)n^2\right). \tag{4}$$

Si une fréquence pure est caractérisée par une *paire* de valeurs singulières, un "chirp" linéaire tel que celui associé à la phase quadratique (4) est caractérisé par un *plateau* de valeurs singulières, ce qui peut se justifier par leur interprétation spectrale. En effet, une étude théorique montre que, sous la condition $K \gg L$, chaque valeur propre de la matrice de covariance normalisée ($C/K = XX^T/K$) peut être approchée par la valeur moyenne d'une portion de la puissance spectrale de la série,

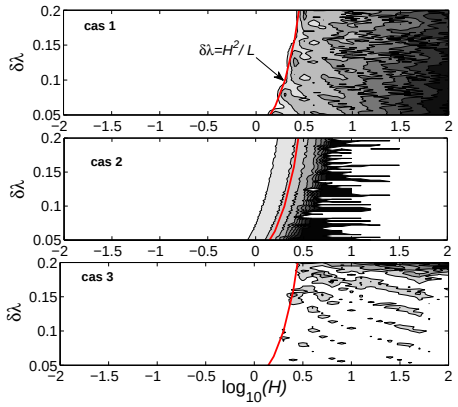

FIGURE 2 – Coefficient de corrélation entre la composante sinusoïdale pure et sa reconstruction obtenue en mode supervisé.

dont la largeur est d'environ K/L [10] :

$$\sigma_i^2/K \approx \frac{1}{l}\sum_{j=il-l}^{il-1} |\hat{s}_j|^2, \qquad l=\frac{K}{L}, \quad i=1,...,L, \qquad (5)$$

en notant $\hat{s}_j$ la transformée de Fourier de s_n. Dans le cas d'un chirp linéaire de largeur de bande (normalisée) $\delta\lambda$, il suit de (5) que le spectre de puissance $|\hat{s}_j|^2$ est plat et répartit uniformément sur $\lfloor L\delta\lambda \rfloor$ valeurs propres égales sur l'intervalle des fréquences positives. elon qu'il y ait ou non un recouvrement des fréquences des composantes, trois cas sont possibles :

1. $\lambda_0 < \lambda_1$
2. $\lambda_1 \leq \lambda_0 \leq \lambda_1+\delta\lambda$
3. $\lambda_0 > \lambda_1+\delta\lambda$

Afin de caractériser la séparabilité, il est possible de fixer λ_0 et λ_1 dans chacun de ces 3 cas et d'analyser le résultat en fonction de H et $\delta\lambda$. C'est ce qui est représenté en figure 2, où est tracée la corrélation entre la sinusoïde de référence, supposée connue, et la composante reconstruite sur la paire de valeurs singulières assurant une erreur minimale. Les paramètres de l'étude sont $L = 40$, $N = 1000$, $\lambda_1 = 0.11$ et $0.05 \leq \delta\lambda \leq 0.2$, la fréquence λ_0 prenant respectivement les valeurs 0.1, $\lambda_1 + \delta\lambda/2$ et 0.32 dans les 3 cas cités précédemment. Une transition est observée dans les trois cas et correspond au passage d'une zone où la séparation des composantes réussit partout à une zone d'ambiguité. La courbe frontière entre les deux zones correspond à la situation où le niveau de la paire des valeurs singulières associées à la sinusoïde pure atteint le plateau des valeurs singulières associées au chirp. L'interprétation spectrale énoncée ci-dessus confirme et quantifie cette intuition. En effet, à un facteur $1/l$ près, les valeurs propres associées à la sinusoïde sont égales à $1/2$, alors que le niveau du plateau des valeurs propres associées au chirp est égal à $(H^2/2)/(L\delta\lambda)$. En égalant ces deux quantités, il vient immédiatement que : $\delta\lambda = H^2/L$. C'est cette courbe parabolique qui est matérialisée (en rouge) sur la figure 2. La transition observée dans le premier et le troisième cas, où il n'y a pas intersection entre les spectres des composantes, est nette, alors qu'un certain élargissement de la frontière existe dans le cas du recouvrement spectral.

4 Sélection automatique des composantes

À la suite de l'analyse spectrale singulière, une phase de sélection et de regroupement automatique des composantes succède immanquablement. Les méthodes existantes utilisent le plus souvent des hypothèses sur la nature des composantes recherchées. Ainsi des solutions ont été proposées séparément, notamment pour des composantes périodiques ou des tendances (sinusoïdes de basse fréquence) [11] et pour des composantes chaotiques bruitées [12].

4.1 Approche non supervisée par classification ascendante hiérarchique

Nous proposons d'utiliser un algorithme de classification automatique non supervisée pour regrouper les signaux $\mathcal{X}_i$ produits par l'analyse spectrale singulière en composantes significatives. Pour cela, une mesure de dissimilarité entre séries chronologiques, à valeurs dans $[0;1]$, va être déduite de leur intercorrélation de Pearson

$$d(\mathcal{X}_i,\mathcal{X}_j)=1-\frac{|<\mathcal{X}_i,\mathcal{X}_j>|}{||\mathcal{X}_i||\cdot||\mathcal{X}_j||}, \text{ avec } <\mathcal{X}_i,\mathcal{X}_j>=\sum_{n=1}^{N} x_{i,n}x_{j,n}$$

et $||\mathcal{X}_i|| = \sqrt{<\mathcal{X}_i,\mathcal{X}_i>}$. La méthode de classification automatique que nous avons choisie est la classification ascendante hiérarchique [13, 14]. Cet algorithme est initialisé en affectant chaque série chronologique à une classe différente. Ensuite, l'algorithme recherche les deux classes distinctes les plus proches et les fusionne pour donner une nouvelle classe. Cette opération est répétée jusqu'à atteindre le nombre de classes souhaité ou lorsque la dissimilarité maximale entre deux éléments de la même classe est atteinte. Pour comparer deux classes c_1 et c_2 comportant plusieurs séries chronologiques, la mesure utilisée est la dissimilarité minimale entre deux éléments de chacune des classes :

$$d(c_1,c_2) = \min_{\mathcal{X}_i\in c_1, \mathcal{X}_j\in c_2} d(\mathcal{X}_i,\mathcal{X}_j). \qquad (6)$$

L'arbre binaire obtenu par les agglomératitions successives constitue le dendrogramme de la classification (voir figure 3). Cette représentation permet de visualiser simultanément l'ordre de fusion des séries chronologiques désignées par leur indice (en abscisse) et leur dissimilarité relative (en ordonnée).

4.2 Retour sur la séparabilité en mode non supervisé

L'application de l'algorithme de classification ascendante hiérarchique appliqué dans le cas du modèle non-stationnaire et avec les mêmes paramètres que précédemment permet d'obtenir la figure 4. Ce résultat montre la corrélation obtenue dans

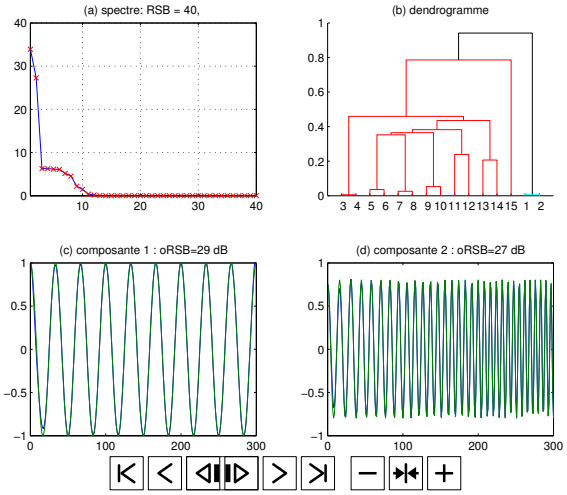

FIGURE 3 – Séparation d'un signal de $N = 300$ échantillons composé d'une sinusoïde ($\lambda_0 = 0.03$) et d'une sinusoïde modulée linéairement en fréquence ($H = 0.8$, $\lambda_1 = 0.06$, $\delta_\lambda = 0.09$). (a) présente le spectre de la matrice de trajectoire. (b) est le dendrogramme obtenu avec l'algorithme proposé pour un nombre maximal de classes fixé à 2 (une couleur distincte par classe). Le signal reconstruit $\hat{s}_i$ associé à chaque classe est représenté sur les figures (c) et (d). L'animation permet de voir les résultats obtenus lorsque le m'elange est bruité par un bruit blanc gaussien avec un RSB allant de 40 à -20 dB. La qualité de reconstruction de chaque composante s_i est donnée par $\text{oRSB} = 10\log_{10}(||s_i||^2/||s_i - \hat{s}_i||^2)$.

les trois cas ainsi que la frontière théorique. Ainsi on constate que la sélection automatique proposée permet d'atteindre les performances optimales que l'on obtient naturellement en mode supervisé en respectant la frontière théorique. Une différence majeure existe cependant avec le mode supervisé dans le cas du recouvrement spectral (cas 2) : la zone d'ambiguité est transformée en une zone où la séparation échoue partout.

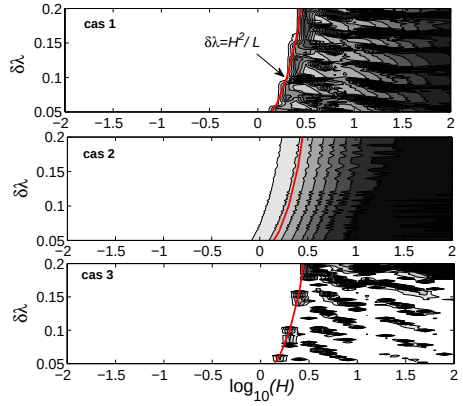

FIGURE 4 – Corrélation entre la composante sinusoïdale pure et sa reconstruction obtenue en mode non supervisé.

5 Conclusion

Classiquement, l'analyse spectrale singulière est vue comme une technique de décomposition d'un signal en oscillations périodiques, tendance et bruit. Cet article s'intéresse au cas où les composantes peuvent être non stationnaires, se proposant d'étudier et de caractériser leur séparabilité sous l'angle de l'interprétation spectrale du spectre singulier. Ainsi, une courbe théorique dépendant du paramètre de réglage de l'algorithme de la SSA, à savoir la dimension L, et des paramètres du signal choisi a permis de délimiter la zone de bonne séparation d'un chirp et d'une sinusoïde. L'extraction de ces composantes est réalisée dans un premier temps par leur connaissance préalable, puis dans un second temps par un algorithme de classification automatique non supervisée proposé à cet effet. Les résultats obtenus ont montré que l'on obtient en mode non supervisé les performances souhaitées. La méthodologie présentée ici permet d'évaluer la capacité du SSA à analyser des signaux non-stationnaires. Elle pourrait être appliquée à d'autres types de non-stationnarités que les chirps étudiés ici. Des éléments permettant de reproduire et d'approfondir les résultats présentés ici sont disponibles sur le site [15].

Références

[1] J. Elsner and A. Tsonis, *Singular Spectrum Analysis, A New Tool in Time Series Analysis*. Plenum Press, 1996.

[2] N. Golyandina, V. Nekrutkin, and A. Zhigljavsky, *Analysis of Time Series Structure : SSA and Related Techniques*. Chapman & Hall/CRC, 2001.

[3] R. Vautard and M. Ghil, "Singular Spectrum Analysis in nonlinear dynamics, with applications to paleoclimatic time series," *Physica D*, vol. 35, pp. 395–424, 1989.

[4] The SSA-MTM group. (2012, May) Publications relating to singular spectrum analysis. [Online]. Available : http://web.atmos.ucla.edu/tcd//ssa/ssa-reference.html

[5] G. Rilling and P. Flandrin, "One or two frequencies ? The Empirical Mode Decomposition answers," *IEEE Trans. Signal Process*, vol. 56, pp. 85–95, 2008.

[6] H.-T. Wu, P. Flandrin, and I. Daubechies, "One or two frequencies ? The synchrosqueezing answers," *Advances in Adaptive Data Analysis*, vol. 3, no. 1 & 2, pp. 29–39, 2011.

[7] J. Cadzow, "Signal enhancement : A composite property mapping algorithm," *IEEE Trans. on ASSP*, vol. 36, no. 2, 1988.

[8] J. Gillard, "Cadzow's basic algorithm, alternating projections and SSA," *Statistics and its Interface*, vol. 3, pp. 335–343, 2010.

[9] N. Golyandina, "On the choice of parameters in singular spectrum analysis and related subspace-based methods," *Statistics and Its Interface*, vol. 3, pp. 259–279, 2010.

[10] E. Bozzo, R. Carniel, and D. Fasino, "Relationship between Singular Spectrum Analysis and Fourier analysis : Theory and application to the monitoring of volcanic activity," *Computers and Mathematics with Applications*, vol. 60, pp. 812–820, 2010.

[11] T. Alexandrov and N. Golyandina, "Automatic extraction and forecast of time series cyclic components within the framework of SSA," *Proceedings of the 5th St.Petersburg Workshop on Simulation*, pp. 45–50, 2005.

[12] R. Vautard, P. Yiou, and M. Ghil, "Singular Spectrum Analysis : A toolkit for short, noisy chaotic signals," *Physica D*, vol. 58, pp. 95–126, 1992.

[13] J. H. Ward, "Hierarchical grouping to optimize an objective function," *JASA*, vol. 58, pp. 236–244, 1963.

[14] L. Lebart, M. Piron, and A. Morineau, *Statistique exploratoire multidimensionnelle*. Dunod, 2006.

[15] [Online]. Available : http://www.ens-lyon.fr/PHYSIQUE/Equipe3/ANR_ASTRES/resources.html