ADAPTING SINGULAR SPECTRUM ANALYSIS TO SPECTRAL ESTIMATION

Alan Zhao, asz2115

Columbia College, Columbia University

ABSTRACT

Singular spectrum analysis was introduced in the 1990s, and has proven to be a powerful technique for performing time series analysis. One of its applications is splitting up a time series into its components, which becomes useful when one component represents a trend or noise.

***Index Terms*—** Singular spectrum analysis

1. INTRODUCTION

The goal of this paper will be an exposition into modern advances of SSA. We will proceed as follows: §2 will discuss the basic SSA algorithm and §3 will discuss the state of the art. Then, §4 is from the project proposal, quickly making some remarks about sampling continuous signals and how SSA becomes useful in this scenario for separating noise. The theory of this separation is discussed in §5. In §6, we further this discussion by discussing optimal window length with respect to separating out noise. We then cover modifications of SSA that have come to light in recent years:

- §7: Oblique SSA and DerivSSA
- §8: SSA-AMUSE

For §7, we will refer to [1]. For §8, we will refer to [2].

2. SSA ALGORITHM

An exposition of the earliest version of SSA may be found in [3, Chapter 4]. The basic (univariate) algorithm currently consists of the following steps, given a time series of length N and window length L:

- Construct the *trajectory matrix* X, defined as the matrix whose $(m+1)$th row is $[x_{1+m} \dots x_{L+m}]$, where $0 \leq m \geq N - L + 1$. Set $K = N - L + 1$.
- Perform SVD on the trajectory matrix X to obtain U, Σ, and V.
- One can now write $X = \sum_{i=1}^{\text{rank}X} X_i$, where $X_i = \sqrt{\lambda_i} U_i V_i^t$, where λ_i is the ith singular value of X in decreasing order.
- Partition the set of indices $\{1 \dots \text{rank}X\}$ into r disjoint subsets $I_1, \dots, I_r$. Letting $X_I = \sum_{i \in I} X_i$, we have $X = \sum_{i=1}^{r} X_{I_i}$.
- Hankelize each X_{I_i}. Apply diagonal averaging to the resulting matrix X_{I_k}, which corresponds to a time series. Thus, we get the original time series as a sum of r time series.

Remark 2.0.1. Hankelization is a linear operator, and so the equality $X = \sum_{i=1}^{r} X_{I_i}$ still holds after hankelizing each term of the right-hand side.

Intuitively, SSA works because it allows one to decompose time series. We discussed previously that this decomposition can help separate noise from meaningful signal. But it can also be used to investigate structure. Take, for instance, the discussion of [4, §1.3.7], which separates a trend of U.S. combat deaths in the Indochina war into two components — one that describes the involvement of the U.S. in the war, and a modulating periodic component that corresponds to war intensity.

For the remainder of the paper, we will use the finest possible partition in the fourth step. The constants N, L, and K will be reserved as used in this context, and X will refer to a trajectory matrix unless specified otherwise. Let T be the time series that corresponds to X. ALl indices start at 0.

3. STATE OF THE ART

There have been a few papers in the recent literature developing the theory of SSA separability — for instance, the theory of [5], which is presented in §9, discusses the optimal window length for separating noise. Much of this theory centers around the production of orthogonality with respect to an inner product in Euclidean space.

4. SSA AND SAMPLING

The signal in question may be band-limited or otherwise. If it is band-limited, the Nyquist sampling theorem tells us how often we should sample the signal. The time series obtained in this manner uniquely represents the signal. Thus, the decomposition of this time series via the theory of §2 will allow one to reconstruct components of the signal. Of course, the

best case scenario is that the noise and (information-carrying) signal components of the signal can be separated.

However, in many cases, the signal we care about will not be band-limited. And for signals that are not band-limited, we no longer have the Nyquist sampling theorem. At whatever rate the signal is sampled, there will be aliasing. Fortunately, in many cases, the spectrum quickly tends to zero as frequency grows. The converse of this scenario comes in the presence of a Gaussian white noise channel. But even then, current theory in digital sampling tells us that the signal-to-noise ratio increases exponentially with respect to the number of bits used to quantize a signal.

In sum, we see that the only limitation of the current theory regarding sampling of continuous signals is that arising from signals that are both not band-limited and whose spectrum does not decay sufficiently quickly. Given a sufficiently high number of quantization bits, we can be fairly confident that the received signal gives a good approximation of the original signal. However, one can only use so many quantization bits. In the presence of a huge amount of noise, sampling is insufficient for interpreting the original signal.

To sidestep this problem, we need to directly contend with the noise. By §2, we could use SSA to separate the noise component. We now present this theory.

5. SEPARABILITY OF UNIVARIATE SSA

5.1. Basics of the Theory

Separability means orthogonality, and refers to the ability to separate X_1 and X_2 in the equality $X = X_1 + X_2$ of trajectory matrices, where X_1 and X_2 may be written as the sum of some number of summands in the singular value decomposition. Then, the column spaces of X_1 and X_2 have readily known structure — they are the eigenvalues and eigenvectors of X. If these column spaces are orthogonal, then the eigenvectors of the trajectory matrix form an orthogonal basis of $\mathbb{R}^{\text{rank}X}$ that distinguishes windows of length L from one another via projection onto these basis elements.

It turns out separability comes in both a weak and strong flavor. The type of separability described above is known as *weak separability*. If the singular values corresponding to X_1 and X_2 are mutually exclusive, then we say that X_1 and X_2 are *strongly separable*. The point of the stronger flavor is to ensure that orthogonality does not depend on the choice of singular value decomposition of X

To make things rigorous, we declare these two terms in the following definitions, rephrased in terms of time series instead of their trajectory matrix.

Definition 5.1. *Fix a singular value decomposition of X. For the corresponding time series T and decomposition $T = T_1 + T_2$, we say that T_1 and T_2 are weakly L-separable if the corresponding decomposition $X = X_1 + X_2$ of trajectory matrices has the property that the column space of X_1 is orthogonal to the column space of X_2 in the standard inner product.*

Definition 5.2. *With notation as in the previous definition, T_1 and T_2 are strongly L-separable if X_1 and X_2 do not share any eigenvalues. Equivalently, there is no need to fix a singular value decomposition of X.*

Remark 5.2.1. Because the trajectory matrix is composed strictly of the time series entries $T = (t_i)$, weak L-separability is equivalent to vanishing relations of polynomials in the t_i. See [4, Proposition 6.1].

Because the trajectory matrix is composed strictly of the time series entries $T = (t_i)$, weak L-separability is equivalent to vanishing relations of polynomials in the t_i.

Orthogonality of two vectors in standard linear algebra can be detected by the cosine of the angle between the two vectors in the plane they span, with the cosine being zero if and only if the two vectors are orthogonal. To measure the failure of orthogonality, we can take the cosine between two windows. Let $T^{i,j} = (t_{i-1}, \ldots, t_{j-1})$. With notation as in Definition 5.1,

$$\rho^{i,j,M} := \frac{\langle T_1^{i,i+M-1}, T_2^{j,j+M-1} \rangle}{||T_1^{i,i+M-1}|| \cdot ||T_2^{j,j+M-1}||}.$$

For infinite time series, T_1 and T_2 are *weakly ε-separable* or *approximately separable* if

$$\rho^L := \max\left(\max|\rho^{i,j,L}|, \max|\rho^{i,j,K}\right) < \varepsilon$$

Let T^N be the first N terms of the series T.t Choosing window lengths $L(N)$, say that T_1 and T_2 are *asymptotically separable* if the $\rho_N := \rho^{L(N),K(N)} \to 0$ as $N \to \infty$. If this holds for any $L(N)$ such that $L(N) \to \infty$, we say that T_1 and T_2 are *regularly asymptotically separable*.

5.2. Separation of a Signal from Random Noise

Fix a probability space $(\Omega, P, \mathcal{F})$. We now proceed to introduce some theoretical results of SSA separability between two random infinite sequences F_1 and F_2. Say these two sequences are *stochastically separable* if approximate separability holds for these two sequences with probability 1.

Suppose F_1 is nonrandom, which we will thus now write as T_1. We want to consider the separability between $T_1 = (t_i)$ and $F_2 = (\tau_j)$. For any $\delta > 0$, introduce the random event

$$A(\delta) := \{\omega \in \Omega : \min_{0 \le j \le L-1} \frac{1}{K} \sum_{m=0}^{K-1} \tau_{m+j}^2 < \delta\}.$$

Let $A'(\delta)$ be the right-hand side with L and K reversed, and let

$$\tilde{f}_{i,k} := \frac{f_{i+k}}{\sqrt{\sum_{j=0}^{K-1} f_{i+k}^2}}$$

and

$$\kappa := \max_{i,j \le L-1} \left| K^{-1/2} \sum_{k=0}^{K-1} \tilde{f}_{i+k} \tau_{j+k} \right|,$$

with κ' being the right-hand side with L and K reversed.

We can now prove a sufficient condition for T_1 and F_2 to be stochastically separable. We first admit the following proposition without proof

Proposition 5.3. *If there exists $\delta > 0$ and window lengths $L = L(N)$ such that*

$$\max(P(A(\delta)), P(A'(\delta))) \to 0 \text{ and } \max(\kappa, \kappa') \to 0,$$

with probability 1 as $N \to \infty$, then T_1 and F_2 are stochastically separable.

Then, of particular interest to us will be the case where F_2 is Gaussian white noise.

Proposition 5.4. *If $L(N), K(N) \to \infty$ and $L(N)/K(N) \to a > 0$, then T_1 is stochastically separable from Gaussian white noise.*

Proof. Set $\delta = 1 - \Delta$, where $0 < \Delta < 1$. The analogue of the Chebyshev inequality for the fourth moment gives us that

$$P\left(\frac{1}{K} \sum_{m=0}^{K-1} \tau_{m+j}^2 < \delta \right) = O(K^{-2})$$

uniformly in j. Then, $P(A(\delta)) = O(L/K^2)$. Since $\tilde{f}_{i,0}^2 + \cdots + \tilde{f}_{i,K-1}^2 = 1$, we have that $C_{i,j,K} := \sum_{m=0}^{K-1} \tilde{f}_{i+k} \tau_{j+k} \in N(0,1)$. We then have that

$$P(|\kappa| > \varepsilon) \le 2L(1 - \Phi(K^{1/2})\varepsilon) \to 0$$

as $L, K \to \infty$ since Φ has exponential decay and $L/K \to a > 0$. □

6. WINDOW LENGTH

The goal of this section is to cover the results in [5], which suggests a window length of $\lfloor \frac{N+1}{2} \rfloor$ in some scenarios. We can compare this to past suggestions.

To start, [4] suggests that we should not choose any window length greater than $N/2$, and so the result of [5] is already a bit of a surprise (though not much, as it is approximately $N/2$). The reason for this is that the transpose of the trajectory matrix will be less than $N/2$ and singular spectrum analysis is not concerned with this symmetry: the transpose of a trajectory matrix is another trajectory matrix.

Another recommendation is explained in [3, §5.2], which is that the window size does not matter as long as it is small compared to N. It mentions that choosing $L = N/4$ is the most common practice when considering choosing L to be analogous to choosing the number of lags to use when constructing an autocorrelation function. To avoid getting into autocorrelation functions, it will suffice here to mention that they are essential in SSA separability analysis for stationary time series.

These two suggested values of $< N/2$ and $N/4$ seem to be largely out of heuristic. The results of [5] run counter to these heuristics, proving the need of the paper's goal, which is to establish theory behind the optimal window length. Specifically, optimization is done with respect to completely separating a noise component. To use the notation of the paper, consider a time series $Y = S + N$, where S is the desired signal and N can be random or deterministic (stationary, which was not covered in the previous section) noise.

Let us now return to [5]. The paper uses $\rho^{i,j,N}$ defined with respect to a weighted inner product $\langle \cdot, \cdot \rangle_w$ such that $\langle x, y \rangle = \sum w_i x_i y_i$, where $w_i = i + 1$ when $0 \le i < \min(L, K)$, $w_i = \min(L, K)$ when $\min(L, K) \le i < \max(L, K)$, and $w_i = N - i$ for $\max(L, K) \le i < N$. Denote this version of $\rho^{i,j,N}$ as $\rho^{(w)}$. The goal will be to minimize $\rho^{(w)}$.

The proof is by linear algebra. Let $\tilde{S}_L^r$ be the reconstructed series based on the first r singular values of the trajectory matrix. Then, one can show that $\tilde{S}_L^r = \tilde{S}_K^r$, and thus for our optimization problem, we only need to consider $L = 2, \ldots, \lfloor \frac{N+1}{2} \rfloor$.

The auxiliary lemma below records a type of linearity of the operator $T(A) = \text{trace}(AA^T)$.

Lemma 6.1. *Let A be an arbitrary $L \times K$ matrix and let B be its Hankelized form. Then, $T(A - B) = T(A) - T(B)$.*

Remark 6.1.1. This records a rare case where we observe a kind of weak linearity of an operator that involves matrix multiplication.

This lemma gives way to the proof for the following two theorems. Note that, from here on, a tilde on a matrix refers to the hankelized version of the matrix. Also, superscripts indicate the window length used.

Theorem 6.2. *$T(\tilde{S}^L)$ is an increasing function on $2 \le L \le \lfloor \frac{N+1}{2} \rfloor$, provided that there exists a Hankel matrix C such that*

$$T(S^L - C) \le T(S^L) - T(\tilde{S}^{L-m})$$

where $m \in \{1, \ldots, L-2\}$.

Theorem 6.3. *$T(S^L - \tilde{S}^L)$ is a decreasing function on $2 \le L \le \lfloor \frac{N+1}{2} \rfloor$, provided that there exists a Hankel matrix C such that*

$$T(S^L - C) \le T(S^{L-m}) - T(\tilde{S}^{L-m})$$

where $m \in \{1, \ldots, L-2\}$.

These two theorems can be combined with the auxiliary lemma to prove

Corollary 6.4. *The minimum value of w-correlation is attained at $L = \lfloor \frac{N+1}{2} \rfloor$ provided that there exists a Hankel matrix C satisfying the inequalities of the theorems.*

The existence of such a C does not come for free. A sufficient condition for existence is the following

Theorem 6.5. *Let $\sigma_\ell^2(S^L)$ be the ℓth secondary diagonal variance of the matrix S^L. If $\sigma_\ell(S^L) \leq \sigma_\ell^2(S^{L-m})$, then there are infinitely many such desired C.*

7. NON-ORTHOGONAL SSA

The definition of weak separability as orthogonality of column spaces in some Hilbert space is strict. The discussion of [1, §1] explains that, in practice, strong separability is needed. Indeed, weak separability is, in essence, "the existence of separability" with respect to a particular singular value decomposition. This would make life hard when conducting real-world analysis.

As strong separability, as perhaps indicated by its name, is hard to come across, [1] presents two modifications of SSA that weaken the separability conditions of univariate SSA. We now present the two modifications to SSA proposed in [1].

Remark 7.0.1. These two methods are meant to be used in a nested manner: first, use univariate SSA, and then use these methods on the resulting groups.

7.1. Oblique SSA (O-SSA)

O-SSA remedies the problem of the lack of weak separability. The main idea of O-SSA is to perform singular value decomposition in a non-orthogonal coordinate system, which we will explain. The derived technique of Iterative O-SSA (I-O-SSA) is similar to prewhitening in statistical preprocessing. The idea is that if we know the covariances between components, we can perform a linear transformation and obtain uncorrelated components. But note that the philosophy of SSA as a non-parametric time series analysis tool makes it particularly suited to the analysis of experimental time series. As such, the covariances are unknown. I-O-SSA helps remedy this problem.

We leave the discussion of the addition of the iterative feature to O-SSA and focus on O-SSA itself.

What Golyandina means by a non-orthogonal coordinate system is just a change of inner product using a positive-definite matrix A. That is, $\langle x, y\rangle_A = \langle Ax, y\rangle$, where the right-hand side is the standard Euclidean inner product between Ax and y.

When modifying the inner product, the corresponding change to be made on the SSA algorithm is replacing singular value decomposition with *restricted singular value decomposition*, which is made with respect to two matrices L and R (and hence are both positive-definite).

Definition 7.1. *The restricted singular value decomposition given by the triple (X, L, R) is a singular value decomposition (U, S, V) such that, for $i = 1, \ldots, \mathrm{rank}X$ the U_i (resp. V_i form an orthonormal system with respect to L (resp. R).*

Before introducing the algorithm, make one more

Definition 7.2. *If the column space of L contains the column space of Y and the column space of R contains the row space of Y, then we will call (L, R) consistent with Y.*

We now have the following algorithm for Oblique SSA, where the input is some triple (X, L, R) such that (L, R) is consistent with X:

- Calculate O_L and O_R, where these two matrices have the property that $O_L O_L^T = L$ and $O_R O_R^T = R$. We can do this by the Cholesky decomposition, which only assumes that L and R are positive-definite.
- Compute $Y = O_L Y O_R^T$.
- Apply singular value decomposition to Y.
- Finally, $X = \sum_i \sigma_i u_i v_i^T$ such that σ_i is the ith singular value of X in decreasing order, $u_i = O_L^\dagger U_i$ and $v_i = O_R^\dagger V_i$, where $\dagger$ denotes pseudo-inverse.

We now state separability criteria. First make the following definitions.

Definition 7.3. *Two series T_1 and T_2 are called weakly (L, R)-separable if the column space of their trajectory matrices are L-orthogonal and the row spaces are R-orthogonal.*

Definition 7.4. *Two series T_1 and T_2 are called strongly (L, R)-separable if their trajectory matrices are weakly (L, R)-separable and share no singular values.*

We now have the following

Theorem 7.5. *Let $T = T_1 + T_2$ and let X have rank r and the X_i have rank r_i. If $r = r_1 + r_2$, then, there exists separating matrices L and R of rank r such that T_1 and T_2 are (L, R)-separable.*

Remark 7.5.1. This theorem provides a sufficient criterion for separability that involves just one linear relation ($r = r_1 + r_2$) versus that for weak separability, which entails the vanishing of a slew of degree 2 polynomials.

The proof of this theorem relies on the following

Proposition 7.6. *A set of linearly independent vectors $P_1, \ldots, P_r$ are A-orthonormal when $A = O_A^T O_A$, $O_A = P^\dagger$, and $P = [P_1 : \cdots : P_r]$.*

(of Theorem **??**. Let $\{P_{i,m}\}$ (resp. $\{Q_{j,m}\}$ be a basis for the column (resp. row) space of X_m. Then, set

$$P = [P_{1,1} \ldots P_{r_1,1} : P_{1,2} \ldots P_{r_2,2}]$$

and

$$Q = [Q_{1,1} \dots Q_{r_1,1} : Q_{1,2} \dots Q_{r_2,2}].$$

By the assumption, P and Q have rank r, and Proposition 7.6, $P^\dagger$ and $Q^\dagger$ orthonormalize the columns of P and Q, respectively. The choice $L = (P^\dagger)^T P^\dagger$ and $R = (Q^\dagger)^T Q^\dagger$ suffices. □

The effectiveness of O-SSA can be seen in the discussion in [1, §3.4.1], which illustrates how well the iterative version of O-SSA works in separating two sinusoids of very close frequency, and thus which are far from orthogonal. We also emphasize Golyandina's warning in [1, §3.2], which states that restricted singular value decomposition provides approximation in an inappropriate way, and so O-SSA cannot be used for the extraction of leading components, and thus for denoising.

7.2. DerivSSA

O-SSA covers the base where we do not have weak separability. In case we have weak, but not strong, separability, we can attempt to use DerivSSA.

DerivSSA proceeds via the following idea: a lack of strong separability is due to overlapping singular values of X_1 and X_2, and so maybe we can consider a time series $A_1T_1 + A_2T_2$ close to T, where we choose the coefficients A_1 and A_2 to make the singular values of A_1X_1 and A_2X_2 different.

A natural approach to doing this is to use the derivative of the time series, and thus the successive difference for discrete time. An example of this technique in the continuous case is the derivative of $\sin(2\pi\omega n + \phi)$ with respect to n, which is $2\pi\omega\cos(2\pi\omega n + \phi)$. Since a cosine wave is just a shifted sine wave, nothing should change in its statistical analysis. From this point of view, only the amplitude of the wave has changed.

Thus, if we had a linear combination

$$a_1 \sin(2\pi\omega_1 n) + a_2 \sin(2\pi\omega_2 n),$$

where $\omega_1 \neq \omega_2$ and the a_i are chosen to make the singular values of each wave overlap, taking the derivative would separate these singular values. This is the example given in [1, §4.3].

The DerivSSA algorithm operates on groupings coming from univariate SSA (it is nested by default) takes two inputs: X and a "weight of derivative" $\gamma > 0$. Then, it does the following:

- Let $\Phi(X) = [X_2 - X_1 : \dots X_K - X_{K-1}]$, where X_i is the ith column of X. Construct the matrix $Y = [X : \gamma\Phi(X)]$.
- Perform the singular value decomposition of Y. Obtain $Y = \sum \sigma_i u_i v_i^T$.
- Construct the following decomposition of X into the sum of elementary matrices: $X = \sum u_i u_i^T X$.
- Partition the indices to obtain $X = X_1 + \dots + X_k$.
- Obtain a refined decomposition $T = T_1 + \dots T_k$ such that T_i corresponds to the hankelization of $X - i$.

This algorithm directly coincides with the O-SSA algorithm given in §7.1, but just with a specific choice of pair (L, R). We state this in the following

Proposition 7.7. *The left singular vectors of the singular value decomposition of Z coincide with those of the restricted singular value decomposition corresponding to (L, R) when L is the identity matrix and $R = I + \gamma^2 F^T F$, where*

$$F = \begin{pmatrix} -1 & 1 & 0 & 0 & \dots & 0 \\ 0 & -1 & 1 & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & -1 & 1 & 0 \\ 0 & \dots & 0 & 0 & -1 & 1 \end{pmatrix}$$

8. A BRIEF POINTER TO OTHER MODIFICATIONS: SSA-AMUSE

We finish off the paper by briefly looking at [2], which is an example of research that finds improvement to SSA that arises when combining SSA with other types of analysis (in this case, principal component analysis). This area of study if relatively new compared to the papers looked at in the main sections. We will summarize the comparison of SSA-AMUSE and DerivSSA given in [2, §4], as both are used to weaken the conditions of strong separability.

Golyandina explains that one advantage of SSA-AMUSE over DerivSSA is that it has the ability to exactly separate components, and thus need not strictly be used in a nested fashion. In the case of approximate or asymptotic separation of harmonics, DerivSSA requires some conditions on the amplitudes to be satisfied. SSA-AMUSE requires none. One disadvantage of SSA-AMUSE is that its resulting orthogonal expansion does not use Frobenius matrices.

9. CONCLUSION

In this paper, we have developed the the theory of SSA separability and modifications to this theory that lend to greater practical applications for SSA. Finally, we briefly touched upon other ways to modify SSA and compared one of them to a more pure modification of SSA. Indeed, the numerous advantages of SSA-AMUSE of DerivSSA indicate that the thread of research discussed in §9 ought to be further pursued.

10. REFERENCES

[1] N. Golyandina and A. Shlemov, "Variations of singular spectrum analysis for separability improvement: non-orthogonal decompositions of time series," *arXiv preprint arXiv:1308.4022*, 2013.

[2] N. E. Golyandina and M. A. Lomtev, "Improvement of separability of time series in singular spectrum analysis using the method of independent component analysis," *Vestnik St. Petersburg University: Mathematics*, vol. 49, no. 1, pp. 9–17, 2016.

[3] J. B. Elsner and A. A. Tsonis, *Singular spectrum analysis: a new tool in time series analysis*. Springer Science & Business Media, 1996.

[4] N. Golyandina, V. Nekrutkin, and A. A. Zhigljavsky, *Analysis of time series structure: SSA and related techniques*. CRC press, 2001.

[5] H. Hassani, R. Mahmoudvand, M. Zokaei, and M. Ghodsi, "On the separability between signal and noise in singular spectrum analysis," *Fluctuation and noise letters*, vol. 11, no. 02, p. 1250014, 2012.