УДК 519.246.87

Н.В. Абалов, В.В. Губарев

АВТОМАТИЧЕСКАЯ ГРУППИРОВКА КОМПОНЕНТ РАЗЛОЖЕНИЯ ВРЕМЕННОГО РЯДА ПРИ СИНГУЛЯРНОМ СПЕКТРАЛЬНОМ АНАЛИЗЕ

Абалов Н.В., Губарев В.В. **Автоматическая группировка компонент разложения временного ряда при сингулярном спектральном анализе.**

Аннотация. Сингулярный спектральный анализ (ССА) является сравнительно новым методом анализа временных рядов. ССА представляет особый интерес в приложении к анализу нестационарных, коротких и зашумлённых рядов. Одной из слабых сторон метода является то, что простые гармонические колебания, как и более сложные компоненты, анализируемого временного ряда раскладываются на более чем одну компоненту, что приводит к необходимости группировки связанных компонент для дальнейшего анализа. Данная проблема частично рассматривается в работе Александрова и Голяндиной (2005), преимущественно в приложении к проблеме идентификации чистых гармонических колебаний.

В данной работе предлагается более гибкий и обобщённый алгоритм для автоматической группировки компонент (а также его модификация), позволяющий группировать не только компоненты, соответствующие гармоническим колебаниям, но и компоненты, соответствующие амплитудно-модулированным колебаниям, затухающим колебаниям и др. Алгоритм был апробирован на искусственных наборах данных, содержащих в себе следующие распространенные формы компонент: гармоническое, амплитудно-модулированное и экспоненциально-затухающее колебания, сумма двух кривых Гаусса, а также их различные аддитивные комбинации. Экспериментально получены оценки качества группировки и показано, что показатели качества группировки у предложенных алгоритмов в среднем лучше на 26%, чем показатели известного алгоритма.

Ключевые слова: сингулярный спектральный анализ; ССА; временные ряды; группировка; идентификация.

Abalov N.V., Gubarev V.V. **Automatic Grouping of Time Series Decomposition Components in Singular Spectrum Analysis.**

Abstract. Singular spectrum analysis (SSA) is a relatively new method of time series analysis. SSA is of particular interest in application to analysis of non-stationary, short and noise time series. One of the drawbacks of SSA is that both simple harmonic oscillations and complex components of analyzed time series are decomposed into more than one component, which leads to the necessity of grouping related components for further analysis. This problem was partially addressed by Alexandrov, Golyandina (2005), mainly in application to the problem of identification of harmonic oscillations. In this paper, we present a more agile and generalized algorithm for automated grouping of components, which allows grouping not only harmonic oscillations, but also components corresponding to amplitude-modulated oscillations, fading oscillations and other. The algorithm was tested on synthetic time series, composed of common components: harmonic, amplitude-modulated, and exponentially damped oscillations, sum of two Gaussians, and their linear combinations. Experimental results of quality of grouping were obtained, showing that the proposed algorithm gives on average 26% better grouping results than an existing algorithm.

Keywords: singular spectrum analysis, SSA, time series, grouping, identification.

1. Введение.

Нестационарное временное поведение характерно для различных естественных, социально-экономических и технических

процессов. При обработке на цифровых вычислительных машинах они представляются эмпирическими данными, которые можно описать нестационарными временными рядами (ВР). Относительно новым и набирающим распространение методом анализа таких ВР является метод сингулярного спектрального анализа (ССА). Его основой послужили методы главных компонент и теории динамических систем. Описание метода и ссылки на ключевые работы в этой области можно найти в [1–6].

Среди основных сильных сторон ССА можно отметить (см., например, [1, 3]) то, что он не предъявляет строгих требований к стационарности ряда, применим к зашумленным и коротким рядам, позволяет выделять как периодические, так и сложные нестационарные компоненты. Слабой стороной метода является то, что он не дает компактного аналитического модельного представления ряда и требует значительного объема ручной работы в диалоговом режиме.

Ранее (см., например, [7, 8]) нами был предложен метод автоматической идентификации нестационарных временны рядов на основе вариативного моделирования, объединяющий методы ССА и моделетеки. Одной из проблем, которая возникает при реализации этого метода, является автоматическая группировка компонент ССА разложения, относящихся к разным составляющим ряда. Она возникает из-за того, что метод ССА в общем случае раскладывает гармонические и более сложные, например затухающие, колебания более чем на одну компоненту. При этом на практике крайне желательно, чтобы компоненты разложения, соответствующие одной составляющей модельного представления ВР (такой как гармоническое и экспоненциально затухающее колебание, амплитудно-модулированное колебание или вейвлет и т.п.), были сгруппированы, позволяли компактно представить и воспроизвести связанную с ними составляющую исходного ряда, а также использовать их при дальнейшем анализе, исследовании ВР. Желательно чтобы такая группировка позволила значительно сократить время автоматического поиска и время подстройки моделей базовых компонент по типу того, как это делается в методе моделетеки, а также упростить результирующую модель.

Проблема группировки компонент возникает не только при совместном использовании ССА и моделетеки, но и при применении лишь ССА, так как исследователь должен вручную выбрать интересующие его компоненты и сгруппировать их. Зачастую при использовании ССА применяют укрупненную группировку, относя компоненты к одной из трех групп: тренд, колебания, шум. При таком крупном группировании, во-первых, понижается наглядность и информатив-

ность получаемых групп, во-вторых, сохраняется проблема компактности аналитического представления группы компонент, что ухудшает интерпретируемость и дальнейшее исследование результатов ССА.

Таким образом, возникает необходимость в алгоритме автоматической группировки связанных компонент разложения ССА. Цель работы – разработка алгоритма группирования компонент, позволяющего автоматизировать данный этап ССА и сократить объем вычислений в задачах вариативной идентификации [7]. В данной работе представлен разработанный алгоритм и проводится его сравнение с существующим алгоритмом.

2. Методы.

2.1. Сингулярный спектральный анализ. Для лучшего понимания предлагаемого метода, кратко рассмотрим на примере одномерного ряда Y_N основные этапы ССА, позволяющего разложить исходный временной ряд $Y_N = (y_0, y_1, \dots, y_{N-1})$ длины N, где $y_i = y(i\Delta t)$, $i = 1, \dots, N-1 = \overline{1, N-1}$, на набор аддитивных компонент. В ССА можно выделить два укрупненных этапа: разложение и восстановление.

*Разложение***.** Первым шагом этапа является преобразование (вложение) одномерного временного ряда Y_N в траекторную матрицу $\mathbf{X}$ размером $L \times K$, где L – длина скользящего вдоль ВР окна (гусеницы), $1 < L < N$, $K = N - L + 1$, $\mathbf{X} = [X_1, X_2, \dots, X_K]$, $X_h = (y_{h-1}, \dots, y_{h+L-2})^T$, $h = 1, \dots, K$. Следующий шаг – вычисление матрицы $\mathbf{X}\mathbf{X}^T$ и ее разложение по собственным векторам. Результатом шага является набор собственных троек $\left(\sqrt{\lambda_j}, U_j, V_j\right)$, $j = 1, \dots, d$, упорядоченных по убыванию ненулевых собственных чисел λ_j, где U_j – собственный вектор, а $V_j = \sqrt{\lambda_j}\mathbf{X}^T U_j$ – факторный вектор.

*Восстановление***.** На данном этапе производится группировка компонент в соответствии с интересами исследователя и восстановление сгруппированных компонент и всего ряда (численная замена исходного ряда новым, полученным суммированием значений отобранных и сгруппированных собственных (восстановленных) компонент). Для получения восстановленной компоненты RC_j, соответствующей одной j-ой тройке, необходимо вычислить диагональное усреднение (ганкелизацию) матрицы $\sqrt{\lambda_j}U_j{V_j}^T$. Рассматриваемые далее алгоритмы применяются после этапа разложения, когда получены все тройки $\left(\sqrt{\lambda_j}, U_j, V_j\right)$, а также, в отдельных случаях, вычислены соответствующие этим тройкам восстановленные компоненты RC_j.

2.2. Алгоритм группировки Ф. И. Александрова, Н. Э. Голяндиной. В работе [9] при решении задачи идентификации различных компонент (трендовой, колебательной или шумовой) косвенно,

как этап выполнения идентификации, решается задача группировки только пар компонент, соответствующих чистым гармоникам.

Пусть Π_Z – значение нормированной периодограммы заданного ряда Z, имеющего длину L. $\Pi_Z(k)$ – значение периодограммы, соответствующее частоте k/L, где $0 \le k \le L/2$ – индекс частоты. Суть предложенного Александровым и Голяндиной алгоритма можно кратко изложить в виде последовательности следующих действий:

1. Перебираем пары последовательных собственных троек $\left(\sqrt{\lambda_j}, U_j, V_j\right), \left(\sqrt{\lambda_{j+1}}, U_{j+1}, V_{j+1}\right), j = 1, \dots, d-1$.
2. Для каждой пары вычисляем показатель:

$$\rho_{j,j+1} = \frac{1}{2}\max_{0 \le k \le L/2}\left(\Pi_{U_j}(k) + \Pi_{U_{j+1}}(k)\right). \quad (1)$$

3. Если для текущего j значение $\rho_{j,j+1} \ge \rho_0$, где $\rho_0 \in [0,1]$ заданный пользователем предел, то тройки $\left(\sqrt{\lambda_j}, U_j, V_j\right)$ и $\left(\sqrt{\lambda_{j+1}}, U_{j+1}, V_{j+1}\right)$ считаются относящимися к одной компоненте (гармонике) и группируются, в противном случае – не относятся.
4. Переходим к следующему j, т.е. следующим нерассмотренным парам собственных троек, пока не переберем их все.

Согласно данному алгоритму, пара соседних $(j, j+1)$ собственных троек идентифицируется как относящаяся к одному гармоническому колебанию (пара группируется и восстанавливается как одна гармоника – составляющая ряда) при условии, что пики периодограмм Π_{U_j}, $\Pi_{U_{j+1}}$, вычисленных по собственным векторам U_j, U_{j+1} соответственно, приходятся на одинаковые частоты.

Отметим, что в (1) вместо периодограмм Π_{U_j} и $\Pi_{U_{j+1}}$, вычисленных по собственным векторам, можно использовать значения периодограмм Π_{RC_j} и $\Pi_{RC_{j+1}}$, вычисленных по значениям восстановленных компонент RC_j и RC_{j+1} (при условии, что каждая восстановленная компонента RC_j соответствует одной j-ой собственной тройке).

Указанный алгоритм достаточно жесток, причем чем больше ρ_0, тем жестче условие. При больших ρ_0, это затрудняет применение алгоритма в условиях зашумленности ряда или наличия сложных нестационарных компонент. При низких же значениях ρ_0 возможны ложные выводы и некорректные результаты. Алгоритм, фактически, не предусматривает случаев группировки компонент, являющихся составными для амплитудно-модулированных и затухающих колебаний (поскольку авторами [9] изначально ставилась задача идентификации чистых гармонических колебаний). В дальнейшем будем обозначать этот алгоритм как HG (harmonic grouping).

Рассмотрим пример, как жесткость алгоритма может приводить к низкому качеству группировки. Положим, что $\rho_0 = 0{,}8$. Рассмотрим рисунок 1.

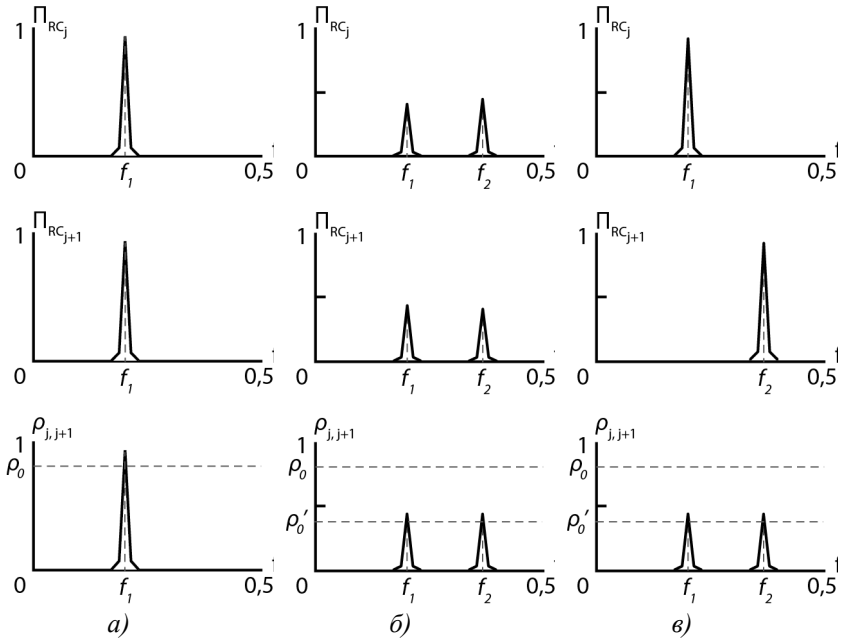

Рис. 1. Примеры группировки компонент

Случай *а)* является наиболее простым, именно для него предлагался алгоритм в [9]. Поскольку пики двух периодограмм приходятся на одну частоту, совпадающую с f_1, каждая нормированная периодограмма содержит только один пик (их значения, как следствие, близки к 1), то оценка $\rho_{j,j+1}$ имеет значение близкое к единице, превышающие ρ_0. В случае *б)*, поскольку каждая нормированная периодограмма содержит по два пика, то максимальное значение нормированной периодограммы не превышает 0,5. Тогда при вычислении показателя $\rho_{j,j+1}$ по (1) его значение также не может превышать 0,5, т.е. будет меньше, чем ρ_0. Как следствие данная пара, соответствующая одной компоненте, не будет автоматически сгруппирована. Чтобы автоматически группировались и такие компоненты, необходимо понизить ρ_0. Для группировки данной пары необходимо снизить значение ρ_0 и принять его равным, например, $\rho_0' = 0{,}4$. Тогда возникает ситуация *в)*, когда рассматривается пара, которая не должна быть сгруппирована, по-

скольку относится к двум разным чистым гармоникам с частотами f_1 и f_2 соотвественно. Поскольку каждая из $j, j+1$ периодограмм содержит только один пик, максимальное значение данных пиков будет равно 1 или близко к нему. Тогда $\rho_{j,j+1}$ в соответствии с (1) будет равно или незначительно меньше 0,5, т.е. больше, чем выбранное нами из-за случая *б)*, $\rho'_0 = 0{,}4$ и две разные гармоники будут ошибочно автоматически сгруппированы в одну компоненту.

2.3. Предлагаемый алгоритм. Ниже описывается более гибкий алгоритм, направленный на группировку собственных компонент, относящихся к таким элементарным составляющим исходного ВР как: гармонические, амплитудно-модулированные, экспоненциально затухающие колебания и т.п. В дальнейшем будем обозначать этот алгоритм как GG (generalized grouping).

Алгоритм может быть записан в виде следующей последовательности действий:

1. В отличие от HG, для всех компонент рассчитываем матрицу близости собственных чисел $\Delta = [\delta_{ij}]$, $i, j = 1, \dots, d$, где

$$\delta_{ij} = \frac{\min\limits_{\lambda}(\lambda_i, \lambda_j)}{\max\limits_{\lambda}(\lambda_i, \lambda_j)}. \qquad (2)$$

2. Рассчитываем матрицу «смежности» компонент $\mathbf{G} = [g_{ij}]$, $i, j = 1, \dots, d$,

$$g_{ij} = \begin{cases} c_{ij}, & \delta_{ij} \geq \rho_1, \\ 0, & \text{иначе}, \end{cases}$$

где c_{ij} – бинарный показатель близости двух компонент (далее в работе будут рассмотрены два таких показателя).

Итогом выполнения операций 1–2 является матрица группировки $\mathbf{G}$ (аналогичная матрице смежности в теории графов), содержащая 1 в ячейках i, j, если пара компонент i, j принадлежат одной группе, иначе 0.

3. Объединяем в одну группу составляющей ряда те собственные компоненты, для которых $g_{ij} = 1$.

Первый показатель близости c'_{ij} основан на использовании степени коэффициента связи между значениями периодограмм восстановленных компонент (например, коэффициента корреляции Пирсона, Спирмена, конкоров Губарева, корреляционных отношений и т.п.).

Здесь $c'_{ij} = 1$, тогда и только тогда, когда $corr(RC_i, RC_j) \geq \rho_c$, $\rho_c \in [0,1]$.

Использование алгоритма GG с показателем в виде коэффициента корреляции Пирсона на рассматриваемых примерах, как будет показано ниже, в среднем дает лучшие результаты, чем алгоритм HG, особенно в случае наличия амплитудно-модулированных и других компонент, отличных от чистой гармоники. Чем выше значение ρ_c, тем жестче условие.

Второй показатель близости c''_{ij} основан на использовании «гибкого» сравнения множеств частот, соответствующих максимальным значениям периодограмм. Введем следующие обозначения:

$K_{RC_j} = \max_{o \le k \le L/2}\left(\Pi_{RC_j}(k)\right)$ – максимальное значение периодограммы восстановленной компоненты RC_j;

$$F_{RC_j} = \left\{k | \Pi_{RC_j}(k) \ge \rho_p K_{RC_j}\right\}, j = 1, \dots, d, \quad (3)$$

F_{RC_j} – упорядоченное по возрастанию значений множество из n индексов k частот, соответствующих первым n максимальным значениям периодограммы Π_{RC_j}, превышающим порог $\rho_p K_{RC_j}$, где ρ_p – задаваемая величина порога.

Величина порогового значения ρ_p ($\rho_p \in [0,1]$) позволяет регулировать исключение низких значений периодограмм. Например, в случае, когда периодограмма содержит лишь один «доминирующий» пик и остальные будут ниже порогового значения (для чистого гармонического колебания), независимо от n будет выбрана лишь одна частота. Значение $n = 2$ – является достаточным для выделения двух пиковых значений периодограммы, позволяющих идентифицировать амплитудно-модулированное колебание. Чем выше значение ρ_p, тем жестче условия на отбор «значимых» частот.

Здесь $c''_{ij} = 1$ тогда и только тогда, когда $\forall h = 1, \dots, m$ имеем $\frac{\left|F_{RC_i}(h) - F_{RC_j}(h)\right|}{L/2} \le \rho_2$, где $|\ |$ – модуль, $F_{RC_i}(h)$ – h-ое значение индекса частоты из множества F_{RC_i}, ρ_2 – порог предельного расхождения индексов частот, $\rho_2 \in [0,1]$, $m = \min\left(\#F_{RC_i}, \#F_{RC_j}\right)$, где # – обозначает мощность соответствующих множеств. Чем меньше значение ρ_2, тем жестче условие.

Использование данного показателя c''_{ij} должно позволить учесть размытие и слабое смещение («дрифт») пиковых значений периодограмм при повышении уровня соотношения сигнала к шуму.

Рассмотрим основные отличия предложенного алгоритма, являющегося расширением идей алгоритма Александрова и Голяндиной, от существующего алгоритма. Он является расширением алгоритма Александрова и Голяндиной. Они заключаются в следующем:

а) Вместо строго последовательного попарного рассмотрения еще не сгруппированных компонент предлагается рассматривать все комбинации пар компонент, имеющих близкие значения собственных числе в соответствии с оценкой (2).

б) Оценка близости частотного состава осуществляется не по совпадению значений частот, соответствующих только одному доминирующему пику в периодограммах собственных векторов, а по одному из двух предложенных критериев, первый из которых позволяет учесть весь частотный состав, второй – более одного пика в периодограмме и слабое смещение («дрифт») значений периодограммы из-за высокого уровня шума.

Первая особенность позволяет обобщить правило группирования и обеспечить большую гибкость алгоритма в условиях сильной зашумленности ряда или наличия сложных нестационарных компонент. Вторая особенность позволяет сместить фокус с группировки только чистых гармонических колебаний на группировку амплитудно-модулированных гармонических колебаний (периодограммы которых имеют пики на двух частотах), затухающих колебаний и других сложных по форме компонент.

3. Методология тестирования и сравнения алгоритмов. Для тестирования алгоритмов используется метод идеального сигнала. Многократно повторялись опыты по группировке собственных компонент для различных значений соотношения уровня сигнала к шуму $s \in S$ и различных наборов составляющих исходного ряда C^h, $h \in H$. Для оценки качества группировки используется следующий подход.

Для каждой пары искусственного набора компонент и заданного соотношения уровня сигнала к шуму $\langle h, s\rangle \in H \times S$, $H = \{0,1, \dots ,5\}$, $S = \{0{,}05; 0{,}1; 0{,}25; 0{,}5\}$:

1. Повторить (200 раз) следующие шаги:

1.1. Генерируется искусственный временной ряд $x(t) = \sum_{i=1}^{N_h} c_i^h(t) + \varepsilon(t)$, где $N_h = \#C^h$ – количество составляющих (искусст-

венно заданных компонент) в h-ом наборе компонент, $c_i^h(t)$ – случайная реализация i-ой компоненты h-ого набора компонент C^h, $\varepsilon(t)$ – случайный шум, имеющий нормальное распределение с нулевым средним и стандартным отклонением $\sigma_\varepsilon = s \cdot \sigma_\Sigma$, где σ_Σ – стандартное отклонение $\sum_{i=1}^{N_h} c_i^h(t)$.

1.2. Производится автоматическая группировка собственных компонент разложения $x(t)$ рассматриваемыми алгоритмами группировки.

1.3. Заложенные в искусственный имитируемый ряд составляющие сопоставляются с компонентами, полученными в результате группировки. Для этого для каждой составляющей $c_i^h(t)$, заложенной в ряд, находится такая собственная компонента из результата группировки $g_j^h(t)$, которая обеспечивает наибольше значение коэффициента детерминации $R^2{}_i = R^2\left(c_i^h(t), g_j^h(t)\right)$.

1.4. Вычисляется среднее по всем N_h компонентам значение $\overline{R^2}$ полученных для каждого алгоритма коэффициентов детерминации $R^2{}_i$, $i = 1, \dots, N_h$.

2. Для каждого опыта и алгоритма вычисляются следующие оценки показателя качества группировки: $\mu_{\overline{R^2}}$ – среднее значение $\overline{R^2}$ и $\sigma_{\overline{R^2}}$ – СКО $\overline{R^2}$, полученные по 200 повторениям опыта.

Коэффициент детерминации R^2 для каждой составляющей (искусственной компоненты) отражает долю её дисперсии, объясняемой рассматриваемой моделью (выделенной группой собственных компонент).

Пусть X~$U(a; b)$ обозначает, что X – случайная переменная, непрерывно равномерно распределенная на открытом интервале $(a; b)$. Для всех компонент $x = 1, \dots, len$. Если не оговорено иное, то φ_i~$U(-\frac{\pi}{2}; \frac{\pi}{2})$, $len = \lfloor L \rfloor$, где L~$U(182; 730)$. Случайные переменные остаются постоянными для одного опыта (повторения). Новое значение переменной выбирается случайно в соответствии с ограничениями при каждом новом опыте. Описание случайных компонент $c_i^h(t)$, соответствующих тестовому набору C^h, приведено в таблице 1.

Напомним обозначения: HG – алгоритм Александрова и Голяндиной; GG1 – предложенный в работе алгоритм, использующий первый показатель; GG2 – предложенный в работе алгоритм, использующий второй показатель.

Таблица 1. Описание тестовых наборов

k	Аналитическая запись компоненты	Примечания
0	$c_1(t) = \sin(2\pi t/T_1 + \varphi_1)$, $T_1 \sim U(5; 5 + \frac{1}{2} len)$;	Простое гармоническое колебание;
1	$c_1(t) = \sin(2\pi t/T_1 + \varphi_1)\sin(2\pi t/T_2 + \varphi_2)$, $T_1 \sim U(5; 5 + \frac{1}{3} len)$, $T_2 \sim U(\frac{1}{3} len; \frac{2}{3} len)$;	Амплитудно-модулированное колебание;
2	$c_1(t) = \sin(2\pi t/T_1 + \varphi_1)\, e^{-\gamma t/len}$, $T_1 \sim U(5; 5 + \frac{1}{2} len)$; $\gamma \sim U(\frac{1}{2}; 5)$;	Затухающее колебание;
3	$c_1(t) = a_1 e^{-\left(\frac{t-b_1}{c_1}\right)^2} + a_2 e^{-\left(\frac{t-b_2}{c_2}\right)^2}$, $a_1, a_2 \sim \pm U(\frac{1}{2}; 1)$, $b_1, b_2 \sim U(0; \frac{1}{2} len)$, $c_1, c_2 \sim U(0; \frac{1}{3} len)$;	Сумма двух гауссовых кривых;
4	$c_1(t) = a_1 \sin(2\pi t/T_1 + \varphi_1)$, $a_1 \sim U(7; 9)$, $T_1 \sim U(5; 10)$;	Сумма нескольких гармонических колебаний;
	$c_2(t) = a_2 \sin(2\pi t/T_2 + \varphi_2)$, $a_2 \sim U(1; 4)$, $T_2 \sim U(28; 32)$;	
	$c_3(t) = a_3 \sin(2\pi t/T_3 + \varphi_3)$, $a_3 \sim U(10; 18)$, $T_2 \sim U(13; 15)$;	
	$c_4(t) = a_4 \sin(2\pi t/T_4 + \varphi_4)$, $a_4 \sim U(15; 21)$, $T_2 \sim U(58; 62)$;	
5	$c_1(t) = a_1 \sin\left(2\pi t/T_1 + \frac{1}{2}\right)$, $a_1 \sim U(\frac{9}{2}; \frac{13}{2})$, $T_1 \sim U(2 \cdot 365; 3 \cdot 365)$;	Ряд, содержащий все рассмотренные выше компоненты; $len = 2 \cdot 365$.
	$c_2(t) = a_2 \sin(2\pi t/T_2 + 1)$, $a_2 \sim U(\frac{3}{2}; \frac{5}{2})$, $T_2 \sim U(5; 7)$;	
	$c_3(t) = -a_3 \sin(2\pi t/T_{31} + 1)\sin(2\pi t/T_{32})$, $a_3 \sim U(\frac{7}{2}; \frac{9}{2})$, $T_{31} \sim U(10; 12)$, $T_{32} \sim U(150; 170)$;	
	$c_4(t) = -a_4 \sin(2\pi t/T_4 + 1)$, $a_4 \sim U(\frac{1}{2}; \frac{3}{2})$, $T_4 \sim U(29; 31)$;	
	$c_5(t) = a_{51}\exp\left(-\left(\frac{t-b_{51}}{c_{51}}\right)^2\right) + a_{52}\exp\left(-\left(\frac{t-b_{52}}{c_{52}}\right)^2\right)$, $a_{51}, a_{52} \sim U(\frac{1}{2}; \frac{3}{2})$, $b_{51}, b_{52} \sim U(0; \frac{1}{2} len)$, $c_{51}, c_{52} \sim U(0; \frac{1}{3} len)$;	

Выбор параметров осуществлялся на основе предварительных экспериментов и рекомендаций из [9] для алгоритма HG. Опыты проводились при следующих значениях параметров. Для алгоритма HG: $\rho_0 = 0{,}8$. Для алгоритма GG1 и GG2 $\rho_1 = 0{,}8$, $\rho_c = 0{,}8$, $\rho_p = 0{,}8$, $\rho_2 = 0{,}05$.

Таблица 2. Результаты экспериментов

h	s	HG $\mu_{\overline{R^2}}$	HG $\sigma_{\overline{R^2}}$	GG1 $\mu_{\overline{R^2}}$	GG1 $\sigma_{\overline{R^2}}$	GG2 $\mu_{\overline{R^2}}$	GG2 $\sigma_{\overline{R^2}}$
0	0,05	0,84	0,11	0,98	0,06	1,00	0,00
	0,1	0,82	0,10	0,98	0,06	1,00	0,00
	0,25	0,83	0,12	0,99	0,06	1,00	0,00
	0,5	0,83	0,11	0,98	0,06	1,00	0,00
1	0,05	0,50	0,07	0,76	0,16	1,00	0,01
	0,1	0,49	0,10	0,71	0,20	1,00	0,03
	0,25	0,50	0,08	0,73	0,16	0,99	0,04
	0,5	0,50	0,07	0,67	0,14	0,99	0,02
2	0,05	0,81	0,09	0,93	0,12	1,00	0,00
	0,1	0,80	0,09	0,93	0,12	1,00	0,00
	0,25	0,80	0,09	0,94	0,12	1,00	0,01
	0,5	0,81	0,08	0,96	0,10	1,00	0,01
3	0,05	0,70	0,20	0,73	0,19	0,81	0,15
	0,1	0,71	0,21	0,75	0,21	0,83	0,15
	0,25	0,69	0,21	0,73	0,20	0,81	0,17
	0,5	0,70	0,21	0,73	0,21	0,82	0,15
4	0,05	0,86	0,07	0,99	0,02	0,95	0,14
	0,1	0,87	0,06	0,99	0,02	0,96	0,14
	0,25	0,84	0,07	0,95	0,08	0,86	0,18
	0,5	0,75	0,12	0,81	0,18	0,70	0,21
5	0,05	0,64	0,04	0,77	0,04	0,79	0,03
	0,1	0,66	0,04	0,77	0,04	0,79	0,02
	0,25	0,64	0,04	0,74	0,03	0,78	0,02
	0,5	0,62	0,05	0,69	0,05	0,73	0,05
Среднее:		0,72	0,10	0,84	0,11	0,91	0,06

$\mu_{\overline{R^2}}$ – усредненное по 200 повторениям опыта значение оценок качества группировки для заданного набора данных и уровня шума.

$\sigma_{\overline{R^2}}$ – СКО оценок качества группировки для заданного набора данных и уровня шума.

4. Результаты и их обсуждение. Полученные результаты экспериментов представлены в 2. Как видно из таблицы 2, предложенный алгоритм в большинстве случаев показал лучшие результаты: усредненная по многократным повторениям оценка качества группировки $\mu_{\overline{R^2}}$ (средняя оценка $\overline{R^2}$) для предлагаемого алгоритма в большинстве опытов превышает аналогичную для существующего алгоритма. Особо это заметно в случае опытов с амплитудно-модулированной компонентой. Поскольку в данном случае значение $\rho_{j,j+1}$ в (1) мало, т.к. значения периодограммы почти равномерно распределяется между двумя пиками и получаемое значение зачастую значительно меньше значения порога ρ_0,

значительное снижение порогового значения может привести к ошибкам группировки. Кроме того среднее значение СКО оценки качества группировки для алгоритма GG2 ниже такового у алгоритма HG в среднем на 32%, что позволяет предположить, что предлагаемый алгоритм более устойчив к статистическим погрешностям.

5. Заключение. В работе был предложен алгоритм автоматической группировки компонент разложения для метода CCA и его две модификации GG1 и GG2. Алгоритм был апробирован на искусственных данных. Предложенный алгоритм был сравнен по качеству группировки с существующим алгоритмом. Алгоритмы GG1 и GG2 показали лучшее качество группировки (R^2 для GG1 в среднем на 16,67% выше, а для GG2 – на 26,39% выше, чем для существующего алгоритма). Это особо заметно в случаях, когда временной ряд содержит амплитудно-модулированные колебания.

Одним из направлений дальнейшей работы по исследованию алгоритма является выработка рекомендаций по выбору значений пороговых параметров в зависимости от характера составляющих ряда и уровня шума, экспериментальная проверка алгоритмов на примере других составляющих ряда.

Литература

1. *Данилов Д.Л., Жиглявский А.А.* Главные компоненты временных рядов: метод Гусеница // СПб: Издательство Санкт-Петербургского университета. 1997. 307 с.
2. Time series analysis and forecasting, Caterpillar SSA method. URL: http://www.gistatgroup.com/ (дата обращения: 10.04.2014).
3. *Vautard R., Yiou P., Ghil M.* Singular-spectrum analysis: A toolkit for short, noisy chaotic signals // Phys. D Nonlinear Phenom. Elsevier. 1992. vol. 58. no. 1-4. pp. 95–126.
4. *Hassani H.* Singular Spectrum Analysis: Methodology and Comparison // J. Data Sci. University Library of Munich. Germany. 2007. vol. 5. no. 4991. pp. 239–257.
5. *Hassani H., Thomakos D.* A review on singular spectrum analysis for economic and financial time series // Stat. Interface. 2010. vol. 3. pp. 377–397.
6. *Ghil M., Taricco C.* Advanced spectral analysis methods // In Past and Present Variability of the Solar-Terrestrial System: Measurement, Data Analysis and Theoretical Models. 1997. pp. 137–159.
7. *Абалов Н.В., Губарев В.В., Альсова О.К.*. Использование методов сингулярного спектрального анализа и моделетеки при идентификации временных рядов // Труды СПИИРАН. 2014. Вып. 35. С. 49–63.
8. *Gubarev V.V., Alsova O.K., Abalov N.V., Melnikov G.A.* Use of variative modeling for the identification of random signals // Proc. of 7th International Forum on Strategic Technology (IFOST). Tomsk. 2012. vol. 1. pp. 739–742.
9. *Alexandrov Th., Golyandina N.* Automatic extraction and forecast of time series cyclic components within the framework of SSA // Proc. 5th St. Petersburg. Work. Simulation. 2005. pp. 45–50.

References

1. Danilov D., Zhigljavsky A.A. *Glavnye komponenty vremennyh rjadov: metod Gusenica* [Principal Components of Time Series: the Caterpillar Method]. SPB: St. Petersburg University. 1997. 307 p. (In Russ.).

2. Time series analysis and forecasting. Caterpillar SSA method. Available at: http://www.gistatgroup.com/ (accessed: 10.04.2014).
3. Vautard R., Yiou P., Ghil M. Singular-spectrum analysis: A toolkit for short, noisy chaotic signals. *Phys. D Nonlinear Phenom.* Elsevier. 1992. vol. 58. no. 1-4. pp. 95–126.
4. Hassani H. Singular Spectrum Analysis: Methodology and Comparison. *J. Data Sci.* University Library of Munich, Germany. 2007. vol. 5. no. 4991. pp. 239–257.
5. Hassani H., Thomakos D. A review on singular spectrum analysis for economic and financial time series. *Stat. Interface*. 2010. vol. 3. pp. 377–397.
6. Ghil M., Taricco C. Advanced spectral analysis methods. *In Past and Present Variability of the Solar-Terrestrial System: Measurement, Data Analysis and Theoretical Models.* 1997. pp. 137–159.
7. Abalov N.V., Gubarev V.V., Alsova O.C. [Use of Methods of Singular Spectral Analysis and Modeleteka for the Identification of Time Series]. *Trudy SPIIRAN – SPIIRAS Proceedings.* 2014. vol. 35. pp. 49–63. (In Russ.).
8. Gubarev V.V., Alsova O.K., Abalov N.V., Melnikov G.A. Use of variative modeling for the identification of random signals. Proc. of 7th International Forum on Strategic Technology (IFOST). Tomsk. 2012. vol. 1. pp. 739–742.
9. Alexandrov Th., Golyandina N. Automatic extraction and forecast of time series cyclic components within the framework of SSA. Proc. 5th St. Petersburg. Work. Simulation. 2005. pp. 45–50.

Абалов Николай Владимирович — аспирант кафедры вычислительной техники, ФГБОУ ВПО Новосибирский государственный технический университет. Область научных интересов: интеллектуальный анализ данных, вариативное моделирование. Число научных публикаций — 5. nickabalov@yahoo.com; пр. К. Маркса, 20, Новосибирск, 630073; р.т.: +7-913-714-97-03.

Abalov Nikolay Vladimirovich — Ph.D student of computer sciences department, Novosibirsk State Technical University (NSTU). Research interests: intellectual data analysis, variative modeling. The number of publications — 5. nickabalov@yahoo.com; 20, Prospekt K. Marksa, Novosibirsk, 630073; office phone: +7-913-714-97-03.

Губарев Василий Васильевич — д-р техн. наук, профессор, заслуженный деятель науки Российской Федерации, заслуженный работник высшей школы Российской Федерации, профессор кафедры вычислительной техники, ФГБОУ ВПО Новосибирский государственный технический университет (НГТУ). Область научных интересов: идентификация, измерение характеристик, имитация и прогнозирование случайных сигналов; вероятностное моделирование реальных объектов; статистические прикладные информационные системы; системный анализ в экспериментальных исследованиях; интеллектуальный анализ данных и вариативное моделирование; концептуальные основы информатики. Число научных публикаций — 500. gubarev@vt.cs.nstu.ru; пр. К. Маркса, 20, Новосибирск, 630073; р.т.: +7(383)346-11-33.

Gubarev Vasiliy Vasilyevich — Ph.D., Dr. Sci., professor, honored scientist of Russian Federation, honored worker of higher school of Russian Federation, professor of computer sciences department, Novosibirsk State Technical University (NSTU). Research interests: identification, measurement of characteristics, simulation and prediction of random signals; probabilistic modeling of real objects; applied statistical information systems; system analysis in experimental research; intellectual data analysis and variative modeling; conceptual foundations of informatics. The number of publications — 500. gubarev@vt.cs.nstu.ru; 20, Prospekt K. Marksa, Novosibirsk, 630073, Russia; office phone: +7(383)346-11-33.

РЕФЕРАТ

Абалов Н.В., Губарев В.В. **Автоматическая группировка компонент разложения временного ряда при сингулярном спектральном анализе.**

Статья посвящена рассмотрению проблемы автоматической группировки компонент разложения при сингулярном спектральном анализе (ССА). В работе предложен алгоритм для автоматической группировки компонент при ССА. Приведены результаты его апробации на искусственных данных и сравнения с существующим алгоритмом.

ССА является сравнительно новым методом анализа временных рядов. ССА представляет особый интерес в приложении к анализу нестационарных, коротких и зашумлённых рядов. Одной из слабых сторон метода является то, что простые гармонические колебания, как и более сложные компоненты, анализируемого временного ряда раскладываются на более чем одну компоненту, что приводит к необходимости группировки связанных компонент для дальнейшего анализа.

В работе рассматривается существующий алгоритм группировки гармонических компонент, предложенный Ф. И. Александровым, Н. Э. Голяндиной в приложении к задаче идентификации тренда и чистых гармонических колебаний во временных рядах. На ряде примеров показано, что существующий алгоритм жесток и малопригоден для решения задачи группировки сложных компонент нестационарных временных рядов.

Предложен алгоритм, направленный на группировку собственных компонент, относящихся к таким составляющим исходного временного ряда как: гармонические, амплитудно-модулированные, экспоненциально затухающие колебания и т.п.

Приведены результаты апробирования предложенного алгоритма на искусственных наборах данных. Экспериментально получены оценки качества группировки и показано, что показатели качества группировки у предложенных алгоритмов в среднем лучше на 26%, чем показатели известного алгоритма.

SUMMARY

Abalov N.V., Gubarev V.V. **Automatic Grouping of Time Series Decomposition Components in Singular Spectrum Analysis.**

The paper discusses the problem of automated grouping of decomposition components in singular spectrum analysis (SSA). In the paper, a new algorithm for automated grouping of decomposition comments in SSA is presented. The results of its approbation on synthetic time series and comparison to existing algorithm are presented.

SSA is a relatively new method of time series analysis. SSA is of great interest in application to analysis of non-stationary, short and noisy time series. One of the drawback of SSA is the fact that simple harmonic components and complex components of analyzed time series are decomposed into more than one component, which leads to a need for a grouping of such related comments for further analysis.

In the paper, an existing algorithm of grouping, proposed by Alexandrov Th., Golyandina N. in application to identification of trend and pure harmonic components, is considered. Several examples are provided to show that this algorithm is strict and might be unsuitable for solving the problem of grouping of complex components in non-stationary time series.

An algorithm is proposed for automated grouping of such components as harmonic, amplitude-modulated, and exponentially damped oscillations, etc.

Results of approbation of the algorithm on synthetic data are provided. Experimental results of quality of grouping were obtained, showing that the proposed algorithm gives on average 26% better grouping results than an existing algorithm.