

Санкт-Петербургский государственный университет

Прикладная математика и информатика

Учебная практика 2 (проектно-технологическая) (семестр 3)

ПРИМЕНЕНИЕ МЕТОДА МОНТЕ КАРЛО SSA ДЛЯ АНАЛИЗА  
ВРЕМЕННЫХ РЯДОВ

Выполнил:

Потешкин Егор Павлович  
группа 24.М22-мм

Научный руководитель:

д. ф.-м. н., профессор  
Голяндина Нина Эдуардовна  
Кафедра Статистического Моделирования

Санкт-Петербург

2025

# Оглавление

<b>Введение . . . . .</b>	<b>4</b>
<b>Глава 1. Теория случайных процессов . . . . .</b>	<b>6</b>
1.1. Вспомогательные определения . . . . .	6
1.2. Процессы с длинной памятью . . . . .	7
1.3. Оценка параметров . . . . .	9
1.3.1. Maximum likelihood estimation (MLE) . . . . .	9
1.3.2. Whittle estimation . . . . .	10
1.3.3. Численное сравнение методов оценки параметров . . . . .	10
1.3.4. Сходимость оценок к истинным значениям . . . . .	13
<b>Глава 2. Метод Monte Carlo SSA . . . . .</b>	<b>17</b>
2.1. Проверка статистических гипотез . . . . .	17
2.1.1. Поправка неточных критериев . . . . .	17
2.1.2. Сравнение критериев . . . . .	18
2.2. Monte Carlo SSA . . . . .	18
2.2.1. Метод SSA . . . . .	18
2.2.2. Постановка задачи . . . . .	19
2.2.3. Множественный тест . . . . .	19
2.2.4. Ограничение на модель шума . . . . .	20
2.2.5. Используемый вариант MC-SSA . . . . .	21
2.2.6. Сравнение MC-SSA по мощности при разных моделях шума . . . . .	22
2.3. Применение MC-SSA на реальных временных рядах с длинной памятью	24
2.3.1. Nile Minima . . . . .	24
2.3.2. Ireland Wind . . . . .	25
<b>Глава 3. Метод autoMCSSA . . . . .</b>	<b>29</b>
3.1. Сравнение способов задания проекционных векторов . . . . .	29
3.2. SSA с проекцией . . . . .	32
3.3. Автоматическая группировка в SSA . . . . .	32
3.4. Метод autoMCSSA . . . . .	33

3.4.1.	Пример работы алгоритма . . . . .	34
3.4.2.	Подходы к выделению сигнала . . . . .	35
3.5.	Выбор оптимальных параметров autoMCSSA . . . . .	37
3.5.1.	Численное сравнение . . . . .	39
3.5.2.	Случай неизвестной амплитудной модуляции . . . . .	43
<b>Заключение</b>	. . . . .	45
<b>Список литературы</b>	. . . . .	46
<b>Приложение А. Графики</b>	. . . . .	48
A.1.	Сравнение arfima_mle и arfima . . . . .	48

## Введение

Метод Singular Spectrum Analysis (SSA) [1, 2] является мощным инструментом для анализа временных рядов. Он позволяет разложить ряд на интерпретируемые компоненты, такие как тренд, периодические колебания и шум, что значительно упрощает процесс анализа. Метод Monte Carlo SSA [3], в свою очередь, решает задачу обнаружения сигнала в шуме, проверяя соответствующую гипотезу.

Для наиболее распространенного варианта метода Monte Carlo SSA необходимо, чтобы спектральная плотность шума была строго монотонной. Такое ограничение связано с методом SSA и понятием строгой разделимости компонент, без которой оценить доминирующую частоту значимой компоненты не представляется возможным.

В большинстве работ, посвященных методу Monte Carlo SSA, в качестве шума используется модель красного шума — процесса авторегрессии первого порядка с положительным коэффициентом. Такая модель шума обладает строго монотонной спектральной плотностью, однако она плохо описывает временные ряды, обладающие длинной памятью, то есть ряды, автоковариационная функция которых убывает медленней, чем экспоненциальное затухание. Процессы с длинной памятью довольно распространены в реальном мире, например, в работе [4] обнаружена длинная память в таких среднегодовых гидрологических временных рядах, как количество осадков, температура и данные о речном стоке. В работе [5] на наличие длинной памяти исследовалась скорость ветра в Ирландии, в работе [6] исследовался эффект длинной памяти у сейсмических данных. Помимо геофизики, длинная память встречается также в финансах [7, 8].

Помимо проблемы выбора модели шума, существует проблема оценки параметров рассматриваемой модели. В реальных задачах редко встречается ситуация, когда параметры известны, поэтому параметры необходимо наилучшим образом оценить на основе исходного временного ряда. Неправильно оцененные параметры модели могут значительно повлиять на результат Monte Carlo SSA.

После проверки гипотезы о наличии сигнала в шуме возникает задача выделения сигнала, если он обнаружился. Выделяться сигнал можно с помощью SSA, однако один из шагов метода подразумевает визуальный анализ для определения компонент сигнала. В связи с этим возникает потребность в автоматизации этого шага. Этой проблеме посвящены, например, работы [9, 10, 11, 12].

В данной работе была расширена применимость метода Monte Carlo SSA на временные ряды с длинной памятью. Также, поскольку истинные параметры модели для конкретного временного ряда неизвестны почти всегда, было произведено численное сравнение различных методов оценки параметров. Кроме того, был предложен метод автоматической идентификации значимых компонент, использующий метод Monte Carlo SSA для определения значимых частот.

Для реализации рассматриваемых методов использовался язык программирования R [13] и пакет Rssa [14].

# Глава 1

## Теория случайных процессов

### 1.1. Вспомогательные определения

Для начала введем некоторые обозначения, которые будем использовать в дальнейшем.

**Определение 1.1.** Случайный процесс  $\{Y_t : t \in \mathbb{Z}\}$  называют стационарным (в широком смысле), если

1.  $EY_t \equiv \text{const}$  (среднее постоянно по времени);
2.  $\text{cov}(Y_t, Y_{t+h}) = \gamma(h)$  (ковариация зависит только от лага  $h$ ).

**Замечание 1.1.** Поскольку  $\gamma(0) = \text{cov}(Y_t, Y_t) = D Y_t$ , то дисперсия также не меняется со временем.

**Замечание 1.2.** Далее под стационарностью будет подразумеваться именно стационарность в широком смысле.

**Определение 1.2.** Случайный процесс  $\{\varepsilon_t\}$  называют белым шумом  $WN(0, \sigma^2)$ , если он стационарный,  $E\varepsilon_t = 0$ ,  $\gamma(h) = 0 \forall h \neq 0$  и  $D\varepsilon_t = \sigma^2$ .

**Определение 1.3.** Моделью ARMA( $p, q$ ), где  $p, q \in \mathbb{N} \cup \{0\}$  называют случайный процесс  $\{X_t\}$ , удовлетворяющий соотношению

$$X_t = \varepsilon_t + \sum_{i=1}^p \phi_i X_{t-i} + \sum_{i=1}^q \theta_i \varepsilon_{t-i},$$

где  $\{\varepsilon_t\} \sim WN(0, \sigma^2)$ .

**Замечание 1.3.** Модель ARMA( $p, q$ ) является стационарным и обратимым процессом, если корни характеристических полиномов

$$\Phi(L) = 1 - \sum_{i=1}^p \phi_i L^i, \quad \Theta(L) = 1 + \sum_{i=1}^q \theta_i L^i$$

лежат вне единичной окружности  $\{z : |z| = 1\}$  [15, Section 3.4.1].

**Определение 1.4.** Процесс  $\{X_t\}$  называют красным шумом с параметрами  $\phi$  и  $\sigma^2$ , если  $\{X_t\}$  — стационарная модель ARMA( $p, q$ ) с  $p = 1$ ,  $q = 0$  и  $\phi = \phi_1 \in (0, 1)$ .

**Определение 1.5.** Спектральной плотностью стационарного процесса называется такая функция  $f(\omega)$ , что

$$\gamma(h) = 2 \int_0^{1/2} e^{2\pi h \omega i} f(\omega) d\omega.$$

**Определение 1.6.** Пусть  $\{Y_t\}$  — стационарный процесс. Функцию

$$I(\omega) = \frac{1}{n} \left| \sum_{j=1}^n Y_j e^{-2\pi\omega(j-1)i} \right|^2 \quad (1.1)$$

называют периодограммой выборки размера  $n$  процесса  $\{Y_t\}$ .

**Замечание 1.4.** Для любой фиксированной частоты  $\omega_0$

$$\begin{aligned} E(I(\omega_0)) &\rightarrow f(\omega_0), \quad n \rightarrow \infty; \\ D(I(\omega_0)) &\rightarrow f^2(\omega_0) \neq 0, \quad n \rightarrow \infty. \end{aligned}$$

Таким образом периодограмма является асимптотически несмещенной, но несостоительной, оценкой спектральной плотности [16, Раздел 4.5].

## 1.2. Процессы с длинной памятью

**Определение 1.7.** Говорят, что стационарный процесс  $\{Y_t\}$  обладает длинной памятью, если

$$\sum_{h=0}^H |\gamma(h)| \rightarrow \infty,$$

при  $H \rightarrow \infty$ . Иначе говорят, что  $\{Y_t\}$  обладает короткой памятью:

$$\sum_{h=0}^{\infty} |\gamma(h)| < \infty.$$

Существуют и альтернативные определения процессов с длинной памятью, которые можно найти в [17, Раздел 3.1]. Там же показано, что они согласованы с определением 1.7.

**Пример 1.1.** Процессом с короткой памятью является, например, стационарная модель ARMA( $p, q$ ), поскольку  $|\gamma(h)| \leq CR^h$ , где  $C > 0$  и  $0 < R < 1$  [15, Section 10.4].

Введем понятие дробного интегрирования  $(1 - L)^d$ , где  $L$  — оператор сдвига. Например, для  $d = 1$  имеем  $(1 - L)Y_t = Y_t - Y_{t-1}$ , для  $d = 2$  —  $(1 - L)^2 Y_t = Y_t - 2Y_{t-1} + Y_{t-2}$ , и так далее. Обобщим этот оператор для нецелых  $d$  с помощью разложения в ряд Тейлора функции  $(1 - x)^d$  в нуле:

$$\begin{aligned}(1 - x)^d &= 1 - dx - \frac{d(1-d)}{2}x^2 - \frac{d(1-d)(2-d)}{3!}x^3 - \dots \\ &= \sum_{j=0}^{\infty} \pi_j(d)x^j = \sum_{j=0}^{\infty} \binom{d}{j}(-1)^j x^j,\end{aligned}$$

где  $\binom{d}{j}$  — обобщенный биномиальный коэффициент. Коэффициенты  $\pi_j(d)$  удовлетворяют соотношению

$$\pi_j(d) = (-1)^j \binom{d}{j} = \frac{j-1-d}{j} \pi_{j-1}(d) = \frac{\Gamma(j-d)}{\Gamma(j+1)\Gamma(-d)}, \quad (1.2)$$

где  $\Gamma(x)$  — гамма функция. Заметим, что второе равенство в формуле (1.2) верно для любых  $d$ , третье же верно только для  $d \notin \mathbb{N} \cup \{0\}$ , поскольку гамма функция не определена для неположительных целых чисел.

**Определение 1.8.** Пусть процесс  $\{Y_t\}$  определен соотношением

$$Y_t = (1 - L)^{-d} X_t = \sum_{k=0}^{\infty} \pi_k(-d) X_{t-k}, \quad d < 1/2,$$

где  $\pi_k(-d)$  из формулы (1.2),  $\{X_t\}$  — стационарная и обратимая модель ARMA( $p, d$ ). Процесс  $\{Y_t\}$  называют дробно интегрированной моделью ARMA или ARFIMA( $p, d, q$ ).

**Предложение 1.1.** Процесс  $\{Y_t\}$  из определения 1.8 является стационарным процессом с нулевым средним. Его спектральная плотность определяется выражением

$$\begin{aligned}f_Y(\omega) &= 4^{-d} \sin^{-2d}(\pi\omega) f_X(\omega) \\ &= 4^{-d} \sin^{-2d}(\pi\omega) \sigma^2 \frac{|\Theta(e^{-2\pi\omega i})|^2}{|\Phi(e^{-2\pi\omega i})|^2}, \quad \omega > 0 \\ &\sim \omega^{-2d} \sigma^2 \frac{|\Theta(1)|^2}{|\Phi(1)|^2}, \quad \omega \rightarrow 0,\end{aligned} \quad (1.3)$$

где  $\Phi(L)$ ,  $\Theta(L)$  — характеристические полиномы процесса  $\{X_t\}$ .

*Доказательство.* См. [16, Proposition 6.1]. □

**Замечание 1.5.** Из формулы (1.3) видно, что монотонность спектральной плотности процесса  $\{Y_t\}$  зависит от поведения спектральной плотности процесса  $\{X_t\}$ .

**Следствие 1.1.** В условиях предложения 1.1 при  $0 < d < 1/2$

$$\gamma(h) \sim C_{\gamma,d} h^{2d-1}, \quad h \rightarrow \infty,$$

где

$$C_{\gamma,d} = \sigma^2 \frac{|\Theta(1)|^2}{|\Phi(1)|^2} \frac{\Gamma(1-2d)}{\Gamma(d)\Gamma(1-d)}.$$

*Доказательство.* См. [16, Corollary 6.1].  $\square$

**Замечание 1.6.** Из следствия 1.1 сразу следует, что ARFIMA( $p, d, q$ ) с  $d \in (0, 1/2)$  обладает длинной памятью.

### 1.3. Оценка параметров

Пусть  $Y_t = (1 - L)^{-d} X_t$ ,  $d < 1/2$ . Будем считать, что  $\{X_t\}$  представляет собой модель ARMA( $p, q$ ) с нормально распределенным белым шумом  $\{\varepsilon_t\}$ . Тогда представим его спектральную плотность в параметрическом виде:  $f_X(\omega) = f_X(\omega; \psi, \sigma)$ , где

$$\psi = (\phi_1, \dots, \phi_p, \theta_1, \dots, \theta_q)^T.$$

Поставим задачу оценить параметры  $\varphi^T = (d, \psi^T)$  и  $\sigma^2$ .

#### 1.3.1. Maximum likelihood estimation (MLE)

Поскольку  $\{\varepsilon_t\}$  — гауссовский белый шум, вектор

$$\mathbf{Y} = (Y_1, \dots, Y_n)^T \sim \mathcal{N}_n(\mathbf{0}, \Sigma_n),$$

где  $\Sigma_n = (\gamma(|i - j|))_{i,j=1}^n$  — ковариационная матрица  $\mathbf{Y}$ . Совместная плотность распределения  $\mathbf{Y}$  равна

$$(2\pi)^{-n/2} |\Sigma_n|^{-1/2} \exp \left\{ -\frac{1}{2} \mathbf{Y}^T \Sigma_n^{-1} \mathbf{Y} \right\}.$$

Рассмотрим логарифм функции правдоподобия. Отбрасывая аддитивные константы, получаем

$$\ell(\varphi, \sigma^2) = -\frac{1}{2} \ln |\Sigma_n| - \frac{1}{2} \mathbf{Y}^T \Sigma_n^{-1} \mathbf{Y}.$$

Положим  $\Gamma_n = \Sigma_n / \sigma^2$  и, максимизируя  $\ell$  по  $\sigma^2$ , получаем

$$\ell_c(\varphi) = -\frac{n}{2} \ln (S(\varphi)/n) - \frac{1}{2} \ln g_n(\varphi), \tag{1.4}$$

где  $S(\varphi) = \mathbf{Y}^T \Gamma_n \mathbf{Y}$ ,  $g_n(\varphi) = |\Gamma_n|$ . Тогда

$$\hat{\varphi}_{\text{ML}} = \underset{\varphi}{\operatorname{argmax}} \ell_c(\varphi), \quad \hat{\sigma}_{\text{ML}}^2 = S(\hat{\varphi}_{\text{ML}}).$$

**Замечание 1.7.** В случае ненулевого матожидания  $EY_t = \mu$ , для получения  $\hat{\varphi}_{\text{ML}}$  и  $\hat{\sigma}_{\text{ML}}^2$  вместо  $\mathbf{Y}$  рассматривается  $\mathbf{Z} = \mathbf{Y} - \mu$ .

**Замечание 1.8.** Для вычисления  $\ell_c$  можно использовать алгоритм Левинсона-Дурбина, имеющий временную трудоемкость  $O(n^2)$  [18].

### 1.3.2. Whittle estimation

Метод максимального правдоподобия применим, когда известно матожидание  $\mu$ . При неизвестном  $\mu$  обычно используют его оценку  $\bar{\mathbf{Y}}$ , однако, помимо этого, существует проблема вычислительной сложности метода при больших  $n$ .

Обе эти проблемы можно решить, используя оценку Уиттла (Whittle): вместо логарифма функции правдоподобия рассматривается ее оценка (с точностью до константы) [19]. Пусть  $f(\omega; \varphi, \sigma^2)$  — спектральная плотность  $\{Y_t\}$ ,  $I(\omega)$  — периодограмма  $\mathbf{Y}$ , тогда

$$\ell_W(\varphi, \sigma^2) = -\frac{1}{m} \sum_{j=1}^m \left( \ln f(\omega_j; \varphi, \sigma^2) + \frac{I(\omega_j)}{f(\omega_j; \varphi, \sigma^2)} \right),$$

где  $m = \lfloor (n-1)/2 \rfloor$ ,  $\omega_j = j/n$ ,  $j = 1, 2, \dots, m$ . Заметим, что  $f(\omega; \varphi, \sigma^2) = \sigma^2 g(\omega; \varphi)$ .

Тогда, максимизируя  $\ell_W$  по  $\sigma^2$ , получаем

$$\hat{\varphi}_W = \underset{\varphi}{\operatorname{argmax}} Q(\varphi), \quad \hat{\sigma}_W^2 = \frac{1}{m} \sum_{j=1}^m \frac{I(\omega_j)}{g(\omega_j; \hat{\varphi}_W)},$$

где

$$Q(\varphi) = -\ln \frac{1}{m} \sum_{j=1}^m \frac{I(\omega_j)}{g(\omega_j; \varphi)} - \frac{1}{m} \sum_{j=1}^m \ln g(\omega_j; \varphi).$$

**Замечание 1.9.** Такой метод оценки параметров можно использовать при неизвестном среднем, поскольку при ее вычислении не используется значение периодограммы в нуле.

**Замечание 1.10.** Периодограмму временного ряда можно вычислить за  $O(n \log n)$  с помощью быстрого преобразования Фурье, что делает этот метод значительно быстрее MLE для больших  $n$ .

### 1.3.3. Численное сравнение методов оценки параметров

Сравним качество оценок параметров следующих моделей:

1.  $d = q = 0, p = 1$  — модель AR(1);

2.  $p = q = 0$  — модель ARFIMA(0, d, 0).

3.  $p = 1, q = 0$  — модель ARFIMA(1, d, 0).

Для оценки параметров этих моделей на языке R были реализованы функции `arfima_mle` и `arfima_whittle`, которые соответствуют методам MLE и Whittle соответственно. Для реализации MLE были использованы функции `tacvfARFIMA` из пакета `arfima` [20] и `DLLLoglikelihood` из пакета `ltsa` [18], вычисляющие автоковариационную функцию модели ARFIMA и функцию  $\ell_c$  (1.4) соответственно.

Помимо функций `arfima_mle` и `arfima_whittle`, будем использовать:

1. Функцию `arima` из пакета `stats`, соответствующую MLE модели ARMA;
2. Функцию `fracdiff` из пакета `fracdiff` [21], соответствующую аппроксимации MLE модели ARFIMA, описанной в работе [5] (обозначим ее за H&R).

**Замечание 1.11.** Помимо вышеперечисленных функций, в пакете `arfima` есть функция `arfima`, которая вычисляет MLE модели ARFIMA. Но данная реализация MLE в некоторых случаях дает оценки хуже, чем `arfima_mle`. Сравнение оценок обоих реализаций можно найти в разделе A.1.

Поскольку для реальных временных рядов матожидание  $\mu$  неизвестно, будем рассматривать MLE с известным средним и с его оценкой — выборочным средним (будем обозначать их  $\text{MLE}(\mu)$  и  $\text{MLE}(\bar{x})$  соответственно). Не умаляя общности, пусть  $\mu = 0$ .

На рис. 1.1 и 1.2 изображены среднеквадратичное отклонение, смещение и дисперсия оценок параметров  $\phi$  и  $d$  моделей AR(1) и ARFIMA(0, d, 0). Отметим, что все оценки имеют в большинстве своем отрицательное смещение и отличаются между собой в основном степенью смещения. Как и ожидалось, оценка параметров методом максимального правдоподобия с известным средним дает оценку с наименьшим MSE. С другой стороны, если использовать вместо известного среднего выборочное среднее, оценки становятся сильно смещенными. Whittle же, в свою очередь, дает менее смещенную оценку, чем  $\text{MLE}(\bar{x})$ , а в случае оценки  $d$  имеет смещение даже меньше, чем у  $\text{MLE}(\mu)$ . Однако оценки Whittle обладают наибольшей дисперсией среди всех рассмотренных методов, но разница не такая значительная, как в случае смещения.

В таблицах 1.1 и 1.2 представлены значения среднеквадратичной ошибки и смещения оценок параметров  $d$  и  $\phi$  модели ARFIMA(1, d). Заметим, что в оценках присутствует смещение, для  $\phi = 0.1$  и  $\phi = 0.5$  оценки  $d$  имеют отрицательное смещение, а

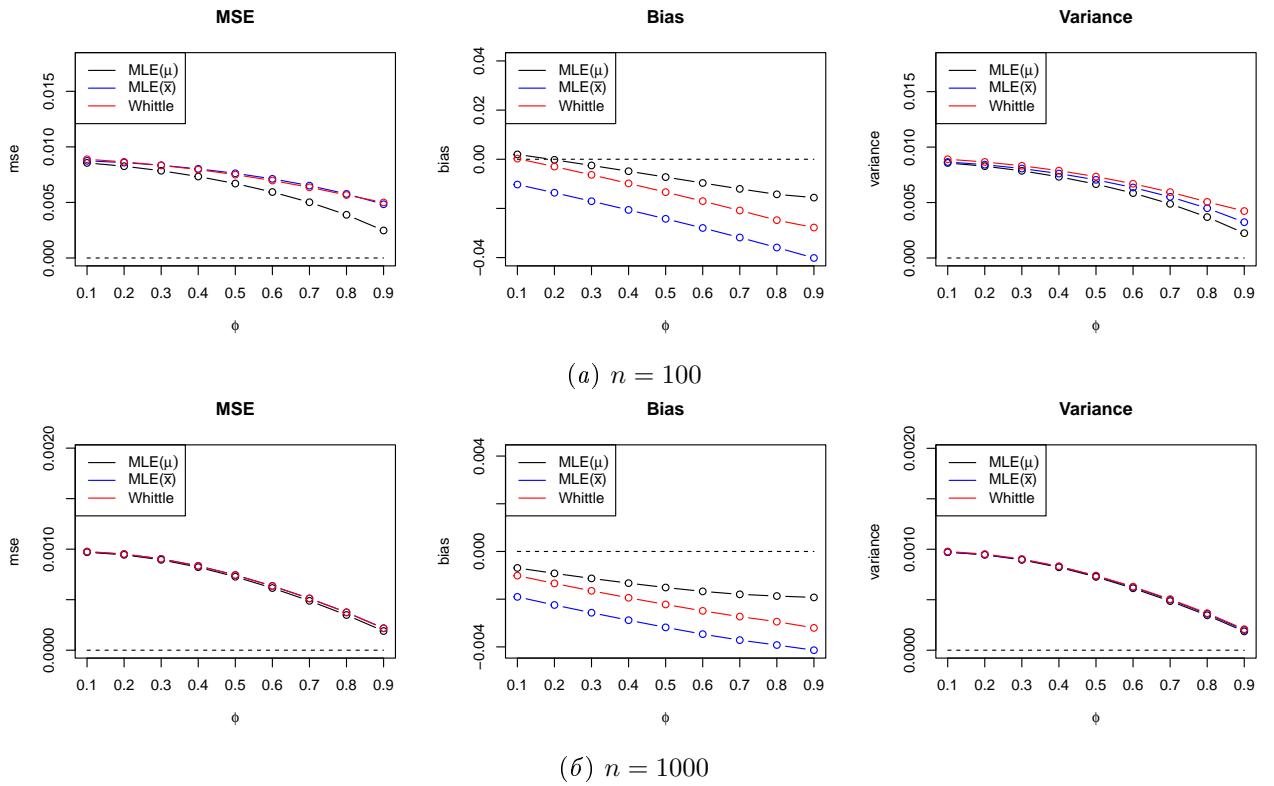


Рис. 1.1. Среднеквадратичное отклонение, смещение и дисперсия оценок параметра  $\phi$  модели AR(1) (500 повторений)

оценки  $\phi$ , наоборот, положительное. Также в таблицах синим цветом выделена лучшая по строке оценка  $d$ , а красным — лучшая оценка  $\phi$ . Видно, что в случае коротких рядов ( $n = 100$ ) метод H&R в большинстве случаев дает оценки с наименьшим MSE, однако наименьшее смещение имеют оценки MLE( $\mu$ ). В случае же длинных рядов ( $n = 1000$ ) наименьшую среднеквадратичную ошибку и смещение дает MLE( $\mu$ ). Отметим, что даже для длинных рядов оценки MLE( $\bar{x}$ ) имеют, в большинстве случаев, наибольшее смещение и MSE. Оценки H&R, хотя и дают наименьшее после MLE( $\mu$ ) MSE, также сильно смещены. Оценки методом Whittle выглядят наиболее привлекательными, поскольку имеют наименьшее после MLE( $\mu$ ) смещение и имеют MSE меньше, чем MLE( $\bar{x}$ ).

Подведем итоги численного сравнения. Если для рассматриваемого ряда известно его матожидание  $\mu$  (что, конечно, редкость на практике), наилучшим методом является MLE( $\mu$ ). Если же среднее неизвестно, для коротких рядов оценивать параметры моделей AR(1) и ARFIMA( $0, d, 0$ ) следует методом Whittle, а параметры модели ARFIMA( $1, d, 0$ ) — методом H&R. В случае длинных рядов параметры рассматриваемых моделей следует оценивать методом Whittle.

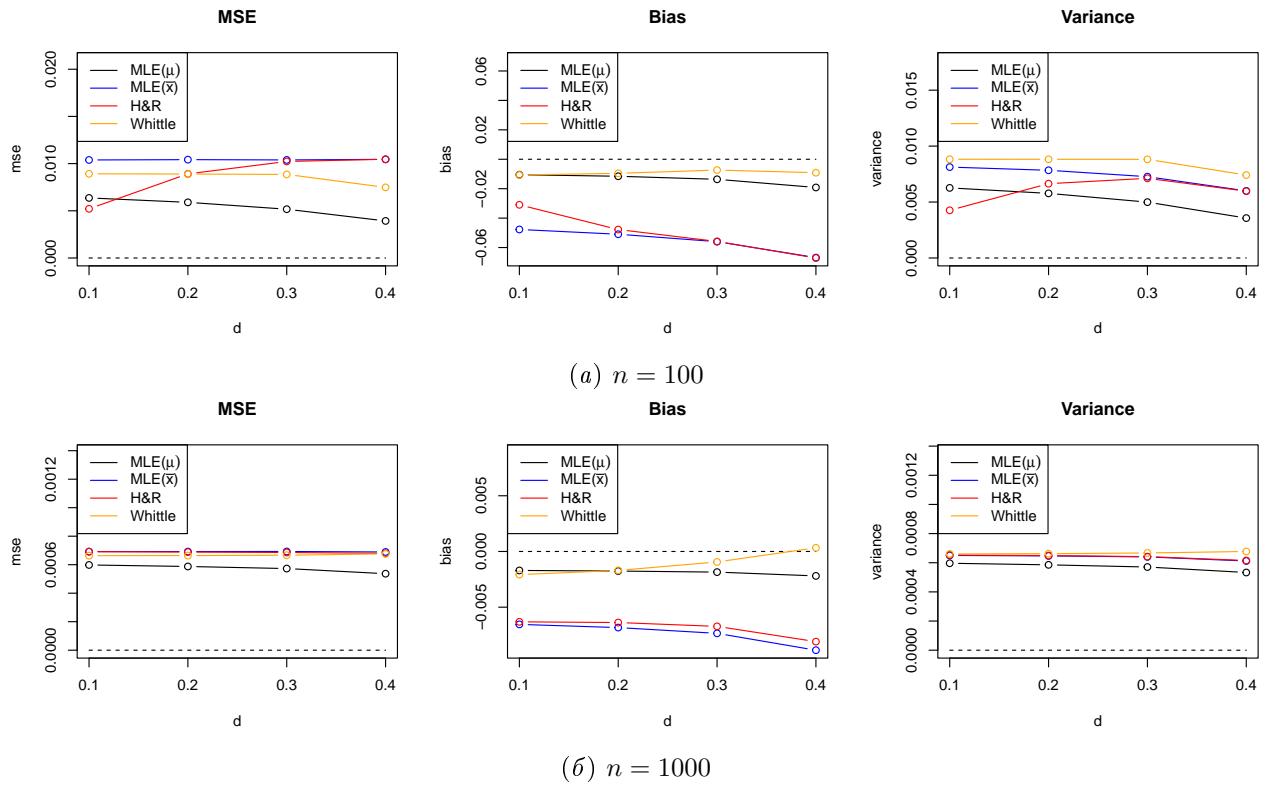


Рис. 1.2. Среднеквадратичное отклонение, смещение и дисперсия оценок параметра  $d$  модели ARFIMA( $0, d, 0$ ) (500 повторений)

### 1.3.4. Сходимость оценок к истинным значениям

Известно [16, Theorem 8.1], что

$$\sqrt{n}(\hat{\varphi}_{\text{ML}} - \varphi_0) \xrightarrow{d} \mathcal{N}_{k+1}(0, \mathcal{I}^{-1}(\varphi_0)), \quad (1.5)$$

где  $\varphi_0$  — истинный вектор параметров,  $\mathcal{I}(\varphi)$  — информационная матрица Фишера. Также известно [16, Proposition 8.3], что вектор  $\hat{\varphi}_W$  имеет такое же асимптотическое распределение, что и  $\hat{\varphi}_{\text{ML}}$ .

Покажем, что методы  $\text{MLE}(\mu)$  и Whittle реализованы корректно, посмотрев на дисперсию оценок для  $n = 10000$ . В таблице 1.3 представлены оценки дисперсий  $\hat{d}$  и  $\hat{\phi}$ , в скобках указана теоретическая дисперсия. Как видим, для разных значений параметров значение дисперсий оценок близки к теоретическим.

Таблица 1.1. Смещение и среднеквадратичное отклонение оценок параметров  $d$  и  $\phi$  модели ARFIMA(1,  $d$ , 0) ( $n = 100$ , 500 повторений)

$d$	$\phi$	MSE						Bias						Whittle			
		MLE( $\bar{x}$ )			H&R			Whittle			MLE( $\bar{x}$ )			H&R			
		MLE( $\mu$ )	$\hat{d}$	$\hat{\phi}$	$\hat{d}$	$\hat{\phi}$	$\hat{d}$	$\hat{\phi}$	$\hat{d}$	$\hat{\phi}$	$\hat{d}$	$\hat{\phi}$	$\hat{d}$	$\hat{\phi}$	$\hat{d}$	$\hat{\phi}$	
0.1	0.1	0.049	0.056	0.119	0.114	<b>0.009</b>	<b>0.018</b>	0.069	0.067	-0.077	0.066	-0.229	0.199	<b>-0.054</b>	<b>0.035</b>	-0.086	0.068
0.2	0.1	0.047	0.055	0.151	0.141	<b>0.025</b>	<b>0.032</b>	0.077	0.073	-0.078	<b>0.067</b>	-0.265	0.232	-0.119	0.099	-0.094	0.074
0.3	0.1	<b>0.041</b>	<b>0.049</b>	0.183	0.165	0.049	0.055	0.084	0.081	-0.076	<b>0.066</b>	-0.301	0.266	-0.179	0.161	-0.109	0.09
0.4	0.1	<b>0.029</b>	<b>0.038</b>	0.211	0.187	0.081	0.089	0.179	0.194	-0.072	<b>0.065</b>	-0.34	0.305	-0.243	0.23	-0.26	0.241
0.1	0.5	0.045	0.041	0.086	0.053	<b>0.010</b>	<b>0.015</b>	0.057	0.054	-0.071	0.034	-0.222	0.151	-0.07	0.034	<b>-0.066</b>	<b>0.024</b>
0.2	0.5	0.042	0.038	0.092	0.055	<b>0.031</b>	<b>0.025</b>	0.074	0.058	-0.081	<b>0.046</b>	-0.244	0.171	-0.154	0.108	-0.153	0.107
0.3	0.5	<b>0.040</b>	<b>0.036</b>	0.1	0.059	0.063	0.043	0.098	0.062	-0.093	<b>0.060</b>	-0.267	0.192	-0.232	0.174	-0.209	0.161
0.4	0.5	<b>0.037</b>	<b>0.033</b>	0.115	0.067	0.104	0.065	0.104	0.066	-0.103	<b>0.073</b>	-0.304	0.226	-0.306	0.235	-0.228	0.177
0.1	0.9	0.029	0.029	0.014	0.01	<b>0.007</b>	<b>0.007</b>	0.034	0.025	0.075	-0.089	0.01	-0.049	<b>0.001</b>	<b>-0.043</b>	0.049	-0.069
0.2	0.9	0.019	0.018	0.011	0.006	<b>0.009</b>	<b>0.004</b>	0.026	0.019	0.046	-0.065	<b>-0.011</b>	-0.035	-0.037	<b>-0.026</b>	0.02	-0.056
0.3	0.9	0.012	0.01	<b>0.009</b>	0.004	0.014	<b>0.003</b>	0.022	0.015	<b>0.016</b>	-0.043	-0.033	-0.023	-0.076	<b>-0.011</b>	-0.024	-0.039
0.4	0.9	<b>0.008</b>	0.006	0.009	<b>0.002</b>	0.025	<b>0.002</b>	0.028	0.01	-0.016	-0.024	-0.061	-0.008	-0.121	<b>0.003</b>	-0.095	-0.016

Таблица 1.2. Смещение и среднеквадратичное отклонение оценок параметров  $d$  и  $\phi$  модели ARFIMA(1,  $d$ , 0) ( $n = 1000$ , 500 повторений)

$d$	$\phi$	MSE · 100						Bias · 100									
		MLE( $\bar{x}$ )			H&R			Whittle			MLE( $\bar{x}$ )			H&R			
		MLE( $\mu$ )	$\hat{d}$	$\hat{\phi}$	MLE( $\bar{x}$ )	$\hat{d}$	$\hat{\phi}$	MLE( $\mu$ )	$\hat{d}$	$\hat{\phi}$	MLE( $\bar{x}$ )	$\hat{d}$	$\hat{\phi}$	MLE( $\mu$ )	$\hat{d}$	$\hat{\phi}$	
0.1	0.1	0.186	0.290	0.27	0.37	0.224	0.317	0.232	0.33	-0.581	0.448	-1.941	1.732	-1.729	1.508	-0.687	0.519
0.2	0.1	0.181	0.287	0.271	0.371	0.267	0.366	0.232	0.329	-0.599	0.465	-2.026	1.824	-1.981	1.773	-0.615	0.468
0.3	0.1	0.174	0.282	0.272	0.372	0.268	0.367	0.232	0.328	-0.639	0.505	-2.177	1.987	-2.176	1.975	-0.476	0.372
0.4	0.1	0.156	0.267	0.272	0.373	0.274	0.371	0.232	0.325	-0.795	0.666	-2.606	2.443	-2.725	2.542	-0.263	0.233
0.1	0.1	0.761	0.8	1.429	1.3	0.504	0.552	1.104	1.05	-2.047	1.588	-5.86	5.102	-3.256	2.741	-2.555	1.904
0.2	0.1	0.710	0.759	1.432	1.302	0.978	0.969	1.213	1.125	-2.018	1.571	-6.08	5.337	-5.157	4.556	-2.721	2.065
0.3	0.1	0.617	0.675	1.462	1.323	1.246	1.175	1.23	1.15	-1.984	1.560	-6.506	5.78	-6.129	5.463	-2.578	1.948
0.4	0.1	0.473	0.539	1.499	1.353	1.507	1.354	1.294	1.18	-2.226	1.861	-7.514	6.838	-7.622	6.905	-2.695	2.099
0.1	0.9	0.338	0.155	0.288	0.122	0.259	0.097	0.387	0.193	0.623	-0.774	-0.095	-0.56	-0.176	-0.504	0.583	-0.92
0.2	0.9	0.273	0.106	0.233	0.077	0.239	0.077	0.326	0.128	0.42	-0.611	-0.388	-0.355	-0.627	-0.303	0.041	-0.652
0.3	0.9	0.241	0.093	0.217	0.068	0.268	0.075	0.326	0.097	0.287	-0.53	-0.667	-0.204	-1.395	-0.029	-1.003	-0.285
0.4	0.9	0.173	0.067	0.182	0.051	0.381	0.076	0.545	0.09	-0.129	-0.295	-1.4	0.178	-2.602	0.359	-3.357	0.389

Метод	$nD\hat{d}$ (1.83167)	$nD\hat{\phi}$ (2.98284)	Метод	$nD\hat{d}$ (1.83167)	$nD\hat{\phi}$ (2.98284)
MLE( $\mu$ )	1.72733	2.92604	MLE( $\mu$ )	1.69348	2.90374
Whittle	1.78350	2.94929	Whittle	1.84265	2.96828

(a)  $d = 0.2, \varphi = 0.1$

Метод	$nD\hat{d}$ (4.91219)	$nD\hat{\phi}$ (6.06018)	Метод	$nD\hat{d}$ (4.91219)	$nD\hat{\phi}$ (6.06018)
MLE( $\mu$ )	5.00869	6.305	MLE( $\mu$ )	4.75975	6.06711
Whittle	5.08405	6.35985	Whittle	5.28104	6.4936

(b)  $d = 0.4, \varphi = 0.1$

Метод	$nD\hat{d}$ (2.49203)	$nD\hat{\phi}$ (0.77885)	Метод	$nD\hat{d}$ (2.49203)	$nD\hat{\phi}$ (0.77885)
MLE( $\mu$ )	2.42011	0.78209	MLE( $\mu$ )	2.26718	0.749318
Whittle	2.44394	0.77549	Whittle	2.57117	0.77311

(c)  $d = 0.2, \varphi = 0.5$

(d)  $d = 0.4, \varphi = 0.5$

Метод	$nD\hat{d}$ (2.49203)	$nD\hat{\phi}$ (0.77885)	Метод	$nD\hat{d}$ (2.49203)	$nD\hat{\phi}$ (0.77885)
MLE( $\mu$ )	2.42011	0.78209	MLE( $\mu$ )	2.26718	0.749318
Whittle	2.44394	0.77549	Whittle	2.57117	0.77311

(e)  $d = 0.2, \varphi = 0.9$

(f)  $d = 0.4, \varphi = 0.9$

Таблица 1.3. Дисперсия оценок  $\hat{d}$  и  $\hat{\phi}$ ,  $n = 10000$ , 100 повторений

## Глава 2

### Метод Monte Carlo SSA

#### 2.1. Проверка статистических гипотез

Рассмотрим некоторый критерий со статистикой  $T$ . Введем обозначения.

**Определение 2.1.** Ошибка первого рода — вероятность отвергнуть нулевую гипотезу, если она верна:  $\alpha_I(\alpha) = \mathsf{P}_{H_0}(T \in A_{\text{крит}}(\alpha))$ .

**Определение 2.2.** Если  $\alpha_I = \alpha$ , то говорят, что критерий точный при уровне значимости  $\alpha$ , иначе говорят, что критерий неточный. При  $\alpha_I < \alpha$  критерий является консервативным, а при  $\alpha_I > \alpha$  — радикальным.

**Определение 2.3.** Мощность критерия против альтернативы  $H_1$  — вероятность отвергнуть нулевую гипотезу, если верна альтернативная:  $\beta(\alpha) = \mathsf{P}_{H_1}(T \in A_{\text{крит}}(\alpha))$ .

#### 2.1.1. Поправка неточных критериев

Зафиксируем некоторый неточный (консервативный или радикальный) критерий и уровень значимости  $\alpha^*$ . Пусть дана зависимость ошибки первого рода от уровня значимости  $\alpha_I(\alpha) = \mathsf{P}_{H_0}(p < \alpha)$ . Тогда критерий с формальным уровнем значимости  $\tilde{\alpha}^* = \alpha_I^{-1}(\alpha^*)$  является точным: ошибка первого рода  $\alpha_I(\tilde{\alpha}^*) = \alpha^*$ .

Если зависимость  $\alpha_I(\alpha)$  неизвестна, она оценивается с помощью моделирования. Приведем алгоритм поправки в этом случае. Помимо критерия и уровня значимости, зафиксируем количество выборок  $M$  для оценки  $\alpha_I(\alpha)$  и их объем  $N$ .

**Алгоритм 1.** Поправка уровня значимости по зависимости  $\alpha_I(\alpha)$  [22]

1. Моделируется  $M$  выборок объема  $N$  при верной  $H_0$ .
2. По моделированным данным строится оценка зависимости ошибки первого рода от уровня значимости  $\alpha_I(\alpha)$ .
3. Рассчитывается формальный уровень значимости:  $\tilde{\alpha}^* = \alpha_I^{-1}(\alpha^*)$ . Критерий с таким уровнем значимости является асимптотически точным при  $M \rightarrow \infty$ .

### 2.1.2. Сравнение критериев

Точные критерии, проверяющие одну и ту же гипотезу, можно использовать и сравнивать по мощности: чем больше мощность, тем лучше. Если критерий является консервативным, использовать и сравнивать его с другими критериями по мощности также можно, учитывая, что его мощность будет занижена. Радикальный же критерий, без поправки, введенной в разделе 2.1.1, нельзя использовать и сравнивать по мощности с другими критериями. Поэтому введем понятие ROC-кривой, соответствующее мощности критерия, к которому была применена поправка.

**Определение 2.4.** ROC-кривая — это кривая, задаваемая параметрически

$$\begin{cases} x = \alpha_I(\alpha) \\ y = \beta(\alpha) \end{cases}, \quad \alpha \in [0, 1]$$

**Замечание 2.1.** С помощью ROC-кривых можно сравнивать по мощности неточные (в частности, радикальные) критерии. Отметим, что для точного критерия ROC-кривая совпадает с графиком мощности, так как  $\alpha_I(\alpha) = \alpha$ .

## 2.2. Monte Carlo SSA

Метод Monte Carlo SSA (MC-SSA) тесно связан с методом SSA (Singular Spectrum Analysis), состоящим из четырех этапов: *вложение, разложение, группировки и диагонального усреднения*. Поэтому опишем сначала его.

### 2.2.1. Метод SSA

Пусть  $\mathbf{X} = (x_1, \dots, x_N)$  — временной ряд длины  $N$ . Зафиксируем длину окна  $L$ ,  $1 < L < N$ . Рассмотрим  $K = N - L + 1$  векторов вложения  $X_i = (x_i, \dots, x_{i+L-1})$  и составим из столбцов  $X_i$  так называемую траекторную матрицу:

$$\mathbf{X} = [X_1 : \dots : X_K].$$

Далее траекторная матрица  $\mathbf{X}$  разбивается в сумму матриц единичного ранга. В базовом SSA используются собственные векторы матрицы  $\mathbf{X}\mathbf{X}^T$ , в Toeplitz SSA используются собственные векторы матрицы  $\mathbf{T}$  с элементами

$$t_{ij} = \frac{1}{N - |i - j|} \sum_{n=1}^{N - |i - j|} x_n x_{n+|i-j|}, \quad i, j \leqslant L. \quad (2.1)$$

Обозначим за  $P_1, \dots, P_L$  собственные векторы матрицы  $\mathbf{X}\mathbf{X}^T$  либо матрицы  $\mathbf{T}$ . Тогда получаем следующее разложение:

$$\mathbf{X} = \sum_{i=1}^L \sigma_i P_i Q_i^T = \mathbf{X}_1 + \dots + \mathbf{X}_L,$$

где  $S_i = \mathbf{X}^T P_i$ ,  $Q_i = S_i / \|S_i\|$ ,  $\sigma_i = \|S_i\|$ .

После этого полученные матрицы группируются и каждая из группированных матриц преобразовывается обратно во временной ряд. Таким образом, результатом SSA является разложение временного ряда.

### 2.2.2. Постановка задачи

Рассмотрим задачу поиска сигнала (неслучайной составляющей) во временном ряде. Модель выглядит следующим образом:

$$\mathbf{X} = \mathbf{S} + \boldsymbol{\xi},$$

где  $\mathbf{S}$  — сигнал,  $\boldsymbol{\xi}$  — стационарный процесс с нулевым средним. Тогда нулевая гипотеза  $H_0 : \mathbf{S} = 0$  (отсутствие сигнала, ряд состоит из чистого шума) и альтернатива  $H_1 : \mathbf{S} \neq 0$  (ряд содержит сигнал, например, периодическую составляющую).

### 2.2.3. Множественный тест

Зафиксируем длину окна  $L$  и модель шума  $\boldsymbol{\xi}$ . Пусть  $\mathbf{R}_1, \dots, \mathbf{R}_G$  — реализации  $\boldsymbol{\xi}$ , которые в дальнейшем будем называть суррогатными. Обозначим за  $\mathbf{X}$  и  $\boldsymbol{\Xi}_i$ ,  $i = 1, \dots, G$ , траекторные матрицы ряда  $\mathbf{X}$  и каждой суррогатной реализации соответственно. Рассмотрим  $H$  проекционных векторов  $W_1, \dots, W_H$ , каждый из которых соответствует некоторой частоте  $\omega_k$ ,  $\|W_k\| = 1$ ,  $k = 1, \dots, H$ .

**Алгоритм 2.** Multiple MC-SSA [23]

1. Для  $k = 1, \dots, H$  вычисляется статистика  $\widehat{p}_k = \|\mathbf{X}^T W_k\|^2$ , выборка  $P_k = \{p_{ki}\}_{i=1}^G$  с элементами  $p_{ki} = \|\boldsymbol{\Xi}_i^T W_k\|^2$ , ее среднее  $\mu_k$  и стандартное отклонение  $\sigma_k$ .
  2. Вычисляется  $\eta = (\eta_1, \dots, \eta_G)$ , где
- $$\eta_i = \max_{1 \leq k \leq H} (p_{ki} - \mu_k) / \sigma_k, \quad i = 1, \dots, G.$$
3. Находится  $q$  как выборочный  $(1 - \alpha)$ -квантиль  $\eta$ , где  $\alpha$  — уровень значимости.

4. Нулевая гипотеза не отвергается, если

$$t = \max_{1 \leq k \leq H} (\hat{p}_k - \mu_k) / \sigma_k < q.$$

5. Если  $H_0$  отвергнута, вклад  $W_k$  (и соответствующей частоты) значим, если  $\hat{p}_k$  пре-  
восходит  $\mu_k + q\sigma_k$ . Таким образом,  $[0, \mu_k + q\sigma_k]$  считаются скорректированными  
интервалами прогнозирования.

#### 2.2.4. Ограничение на модель шума

Для модели шума  $\xi$  важно, чтобы спектральная плотность процесса была строго монотонной. Это связано с тем, что в таком случае собственные векторы автоковариационной матрицы стационарного процесса ведут себя как синусоиды с равностоящими частотами, а соответствующие им собственные числа примерно равны значению спектральной плотности в этих частотах. Для процессов с короткой памятью это верно, поскольку теплицеву симметричную матрицу можно аппроксимировать циркулянтной матрицей [24], собственные векторы которой равны

$$v_j = \frac{1}{\sqrt{n}} (1, \omega^j, \omega^{2j}, \dots, \omega^{(n-1)j}), \quad j = 0, 1, \dots, n-1,$$

а соответствующие им собственные числа равны

$$\lambda_j = \sum_{k=0}^{n-1} c_k \omega^{kj},$$

где  $\omega = \exp\{2\pi i/n\}$ . Тогда при строгой монотонности  $f_\xi(\omega)$  вклад собственных векторов будет попарно различным, делая их сильно разделимыми [2, Раздел 1.5.4]. Если же опустить требование строгой монотонности, компоненты могут смешаться, что сделает невозможным определение доминирующей частоты значимого вектора.

Для процессов с длинной памятью покажем правдивость этого факта, проведя численный эксперимент. Рассмотрим модель ARFIMA(0,  $d$ , 0) с  $d = 0.4$  и пусть размер автоковариационной матрицы  $\Sigma_n$  равен  $n = 100$ . Частоту векторов будем оценивать с помощью метода ESPRIT [14, Раздел 3.1].

На рис. 2.1 представлены первые 10 собственных векторов матрицы  $\Sigma_n$ , на рис. 2.2 по оси  $Ox$  отложена частота, которой соответствует собственный вектор, а по оси  $Oy$  отложено соответствующее вектору собственное число, дополнительно синей линией проведена спектральная плотность процесса. Как видим, действительно, собственные

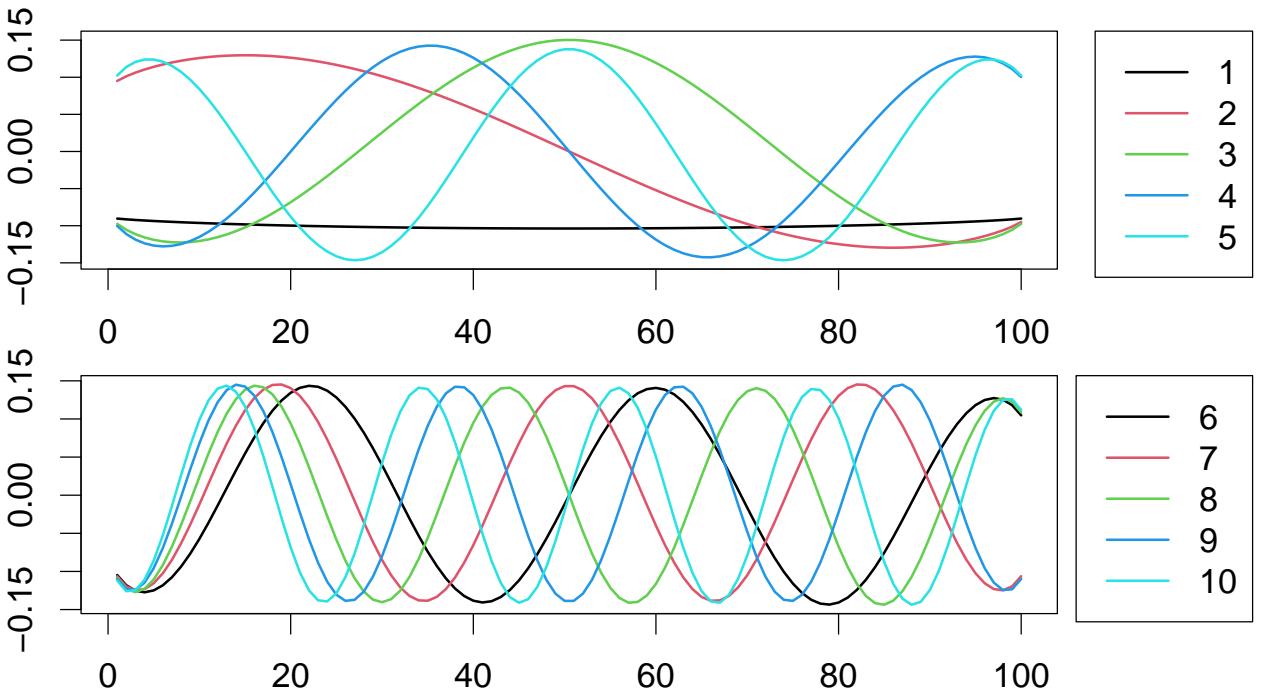


Рис. 2.1. Собственные векторы автоковариационной матрицы модели ARFIMA( $0, d, 0$ )

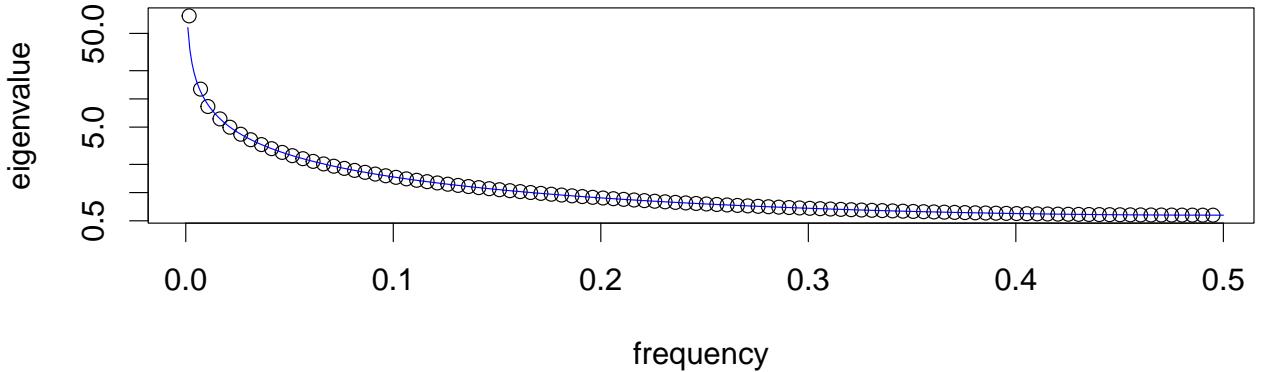


Рис. 2.2. Собственные числа автоковариационной матрицы модели ARFIMA( $0, d, 0$ )

векторы ведут себя как периодики, собственные числа хорошо приближаются значением спектральной плотности в соответствующей частоте, и по рис. 2.3 разница между частотами примерно равна  $1/(2n) = 0.005$ .

### 2.2.5. Используемый вариант MC-SSA

В разделе 2.2.3 предполагалось, что векторы  $W_1, \dots, W_H$  фиксированные и не зависят от исходного ряда. Такой критерий MC-SSA является точным, то есть ошибка первого рода равна заданному уровню значимости. В этой работе будут рассматриваться векторы  $W_k$ , порожденные рядом  $X$ , при этом по-прежнему при вычислении  $p_{ki}$  используются те же  $W_k$ , что и при вычислении  $\hat{p}_k$ .

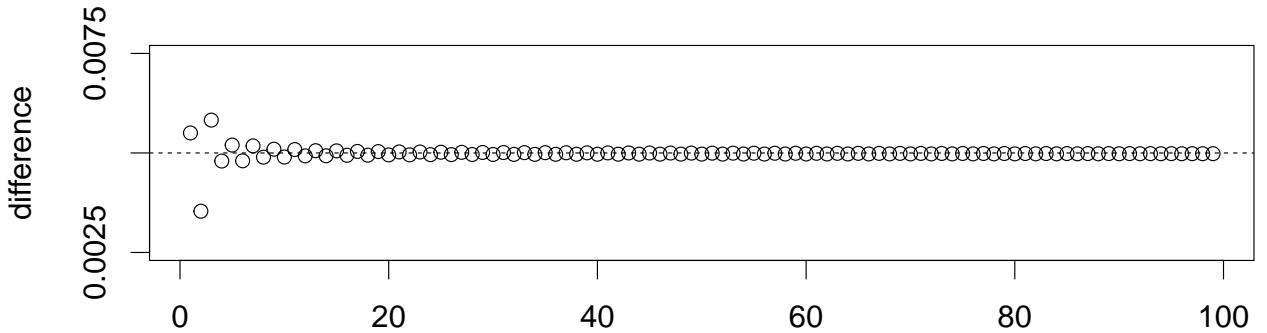


Рис. 2.3. Разница частот между ближайшими собственными векторами

Поскольку в этом варианте векторы  $W_k$  не заданы заранее, а порождены исходным рядом, критерий MC-SSA становится, вообще говоря, радикальным. Бороться с этой проблемой позволяет метод эмпирической поправки критерия, описанный в разделе 2.1.1.

В качестве  $W_1, \dots, W_H$  будем брать собственные векторы матрицы  $\mathbf{X}\mathbf{X}^T$  или  $\mathbf{T}$  (см. формулу (2.1)). Такой способ выбора векторов для проекции самый распространенный, поскольку, если есть значимые векторы, можно восстановить сигнал с помощью SSA на их основе. Будем под MC-SSA подразумевать именно этот вариант критерия. Варианты критерия будут определяться конкретным разложением траекторной матрицы. Заметим, что обычно используется сингулярное разложение.

## 2.2.6. Сравнение MC-SSA по мощности при разных моделях шума

Пусть  $\xi$  — красный шум, а  $\eta$  — модель ARFIMA(0,  $d$ , 0). Будем считать дисперсию белого шума одинаковой для обоих процессов и равной  $\sigma^2$ . Дисперсии  $\xi$  и  $\eta$  соответственно равны

$$D\xi = \frac{\sigma^2}{1 - \phi^2}, \quad D\eta = \sigma^2 \frac{\Gamma(1 - 2d)}{\Gamma(1 - d)^2}.$$

Тогда дисперсии процессов равны тогда и только тогда, когда

$$\phi = \pm \sqrt{1 - \frac{\Gamma(1 - d)^2}{\Gamma(1 - 2d)}}.$$

Пусть  $d = 0.4$ . Тогда при  $\phi \approx 0.719$  процессы  $\xi$  и  $\eta$  имеют одинаковую дисперсию. На рис. 2.4 изображены спектральные плотности процессов. На нем видно, что процесс  $\eta$  имеет меньшее значение плотности для всех значений  $\omega \in (0, 0.2)$ , за исключением близких к нулю.

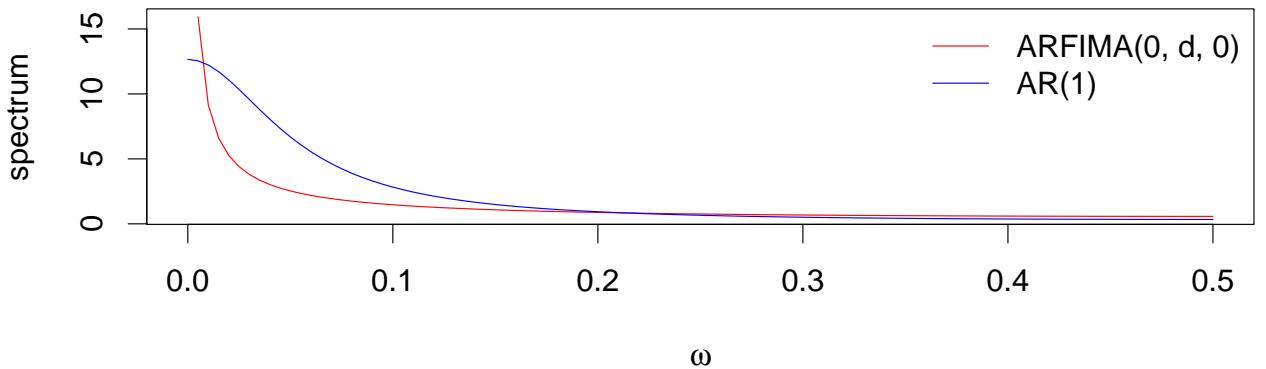


Рис. 2.4. Спектральная плотность процессов с одинаковой дисперсией

Предполагается, что если рассмотреть в качестве альтернативы сигнал с частотой  $\omega : f_{\eta}(\omega) < f_{\xi}(\omega)$ , то мощность критерия MC-SSA против этой альтернативы при модели шума  $\eta$  больше, чем при модели шума  $\xi$ . Убедимся в этом. Пусть длина ряда  $N = 100$ ,  $\sigma^2 = 1$  и

$$S = \{A \cos(2\pi n\omega)\}_{n=1}^N, \quad A = 1, \quad \omega = 0.075.$$

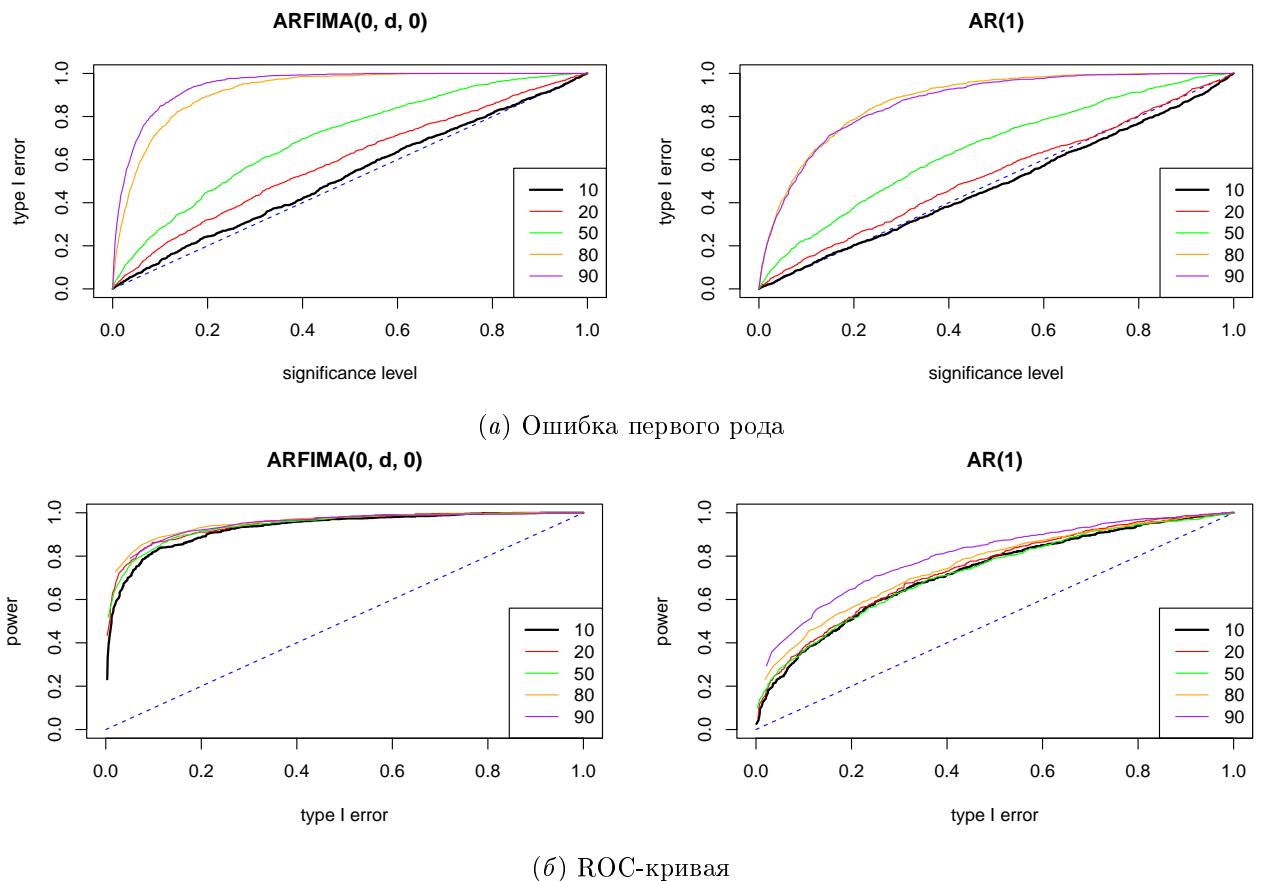


Рис. 2.5. Сравнение мощностей MC-SSA

На рис. 2.5, *a* изображены график ошибок первого рода критериев MC-SSA для

Таблица 2.1. Оценка параметров модели ARFIMA( $0, d, 0$ ) ряда Nile Minima

Метод	$\hat{d}$	$\hat{\sigma}^2$
MLE( $\bar{x}$ )	0.39264	0.48939
H&R	0.39327	0.48934
Whittle	0.40547	0.49026

разных длин окна  $L$ . По нему видно, что рассматриваемые критерии являются радикальными, поэтому сравнивать их по мощности будем с помощью ROC-кривых, которые являются графиками мощности критериев, к которым была применена поправка из раздела 2.1.1. По рис. 2.5,  $b$  видно, что, действительно, мощность критерия против данной альтернативы при модели шума ARFIMA( $0, d, 0$ ) больше, чем при модели шума AR(1) с такой же дисперсией.

## 2.3. Применение MC-SSA на реальных временных рядах с длинной памятью

Рассмотрим несколько примеров реальных временных рядов с длинной памятью и применим к ним MC-SSA. Оценивать параметры будем теми же методами, что и в разделе 1.3.3.

### 2.3.1. Nile Minima

На рис. 2.6,  $a$  изображен ежегодный минимальный уровень воды реки Нил за период с 622 по 1284 год (663 наблюдения), данные были взяты из [25]. Нерегулярные циклы или тенденции в этом временном ряду, обусловленные длинной памятью, впервые были обнаружены и обсуждены Хёрстом, британским инженером, который работал гидрологом на реке Нил. Подтверждает присутствие длинной памяти график медленно угасающей автокорреляционной функции на рис. 2.6,  $b$ .

Оценим параметры модели ARFIMA( $0, d, 0$ ). В таблице 2.1 представлены оценки параметров  $d$  и  $\sigma^2$ . Поскольку истинное среднее неизвестно и оценка  $d$  по Whittle дает наименьшее смещение (см. рис. 1.2), в качестве нулевой гипотезы MC-SSA выберем модель ARFIMA( $0, d, 0$ ) с  $d = 0.40547$  и  $\sigma^2 = 0.48971$ , на рис. 2.7 изображена периодограмма ряда вместе с оцененной спектральной плотностью.

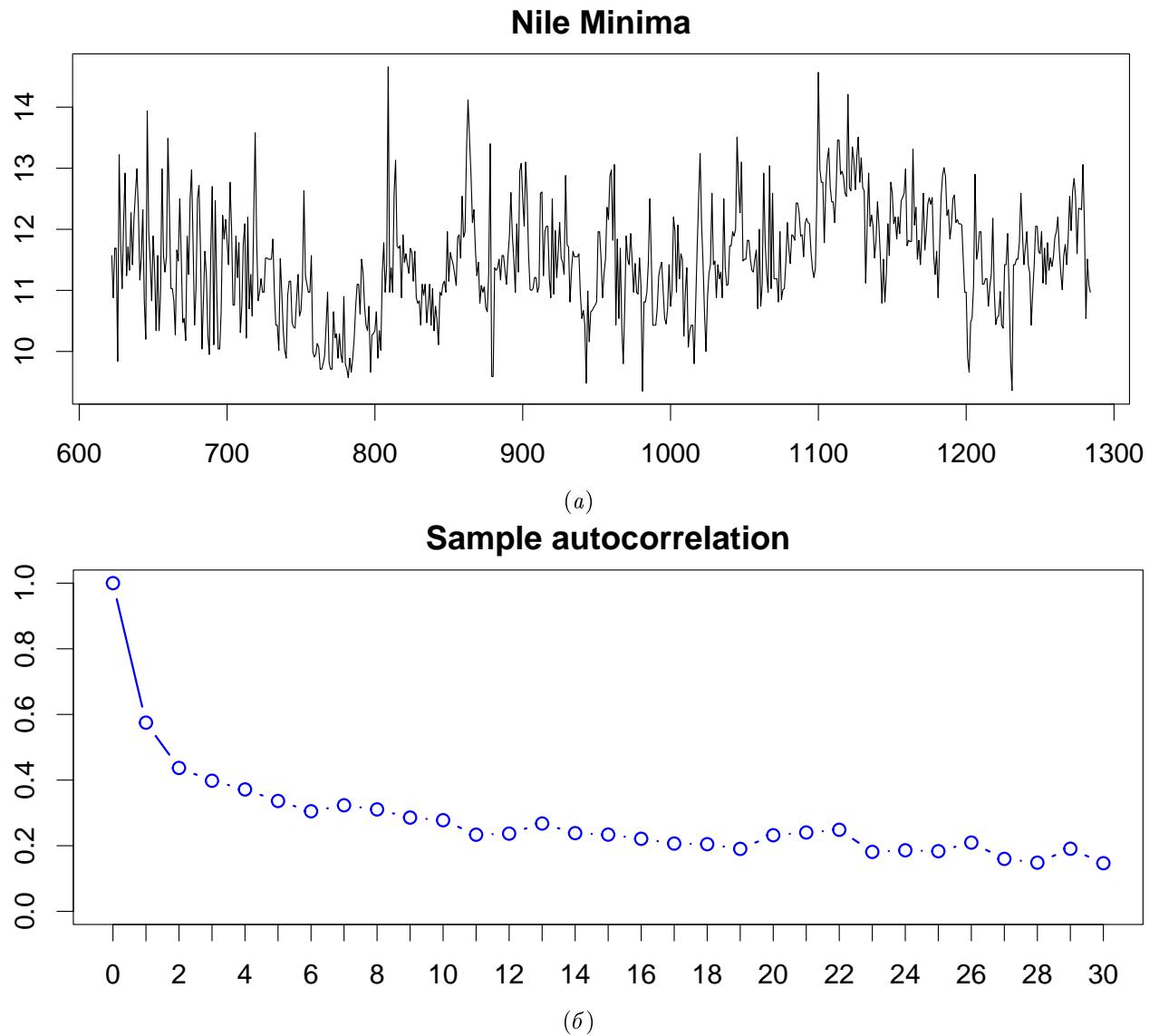


Рис. 2.6. Ежегодный минимальный уровень воды реки Нил

Применим MC-SSA с длиной окна  $L = 330 \approx N/2$ . На рис. 2.8 изображены 95%-ные доверительные интервалы статистик  $\hat{p}_k$ ,  $k = 1, \dots, L$  (см. алгоритм 2). Ни одна из статистик не является значимой, это означает, что нет оснований полагать, что в этом временном ряде присутствует неслучайный сигнал.

### 2.3.2. Ireland Wind

На рис. 2.9 изображены среднесуточные данные о скорости ветра (в узлах) за период с 1961 по 1978 год (6574 дней) на станции Roche's Point в Республике Ирландия [5].

В таблице 2.2 представлены оценки параметров. Полученные оценки примерно одинаковые, но поскольку Whittle дает менее смещенную оценку (см. рис. 1.2 и таблицу 1.2),

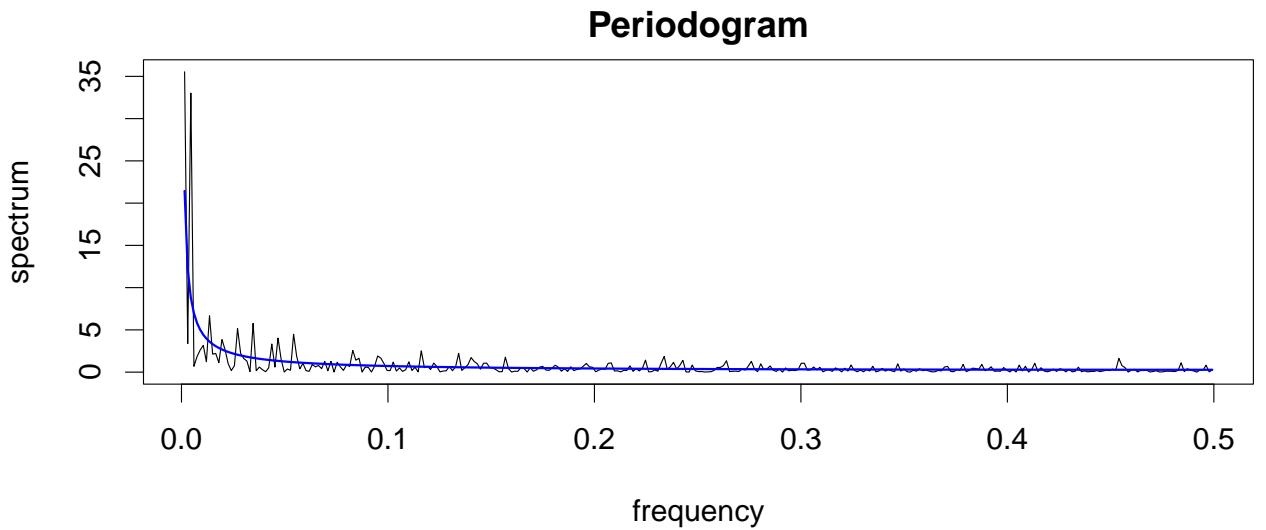


Рис. 2.7. Периодограмма ряда Nile Minima

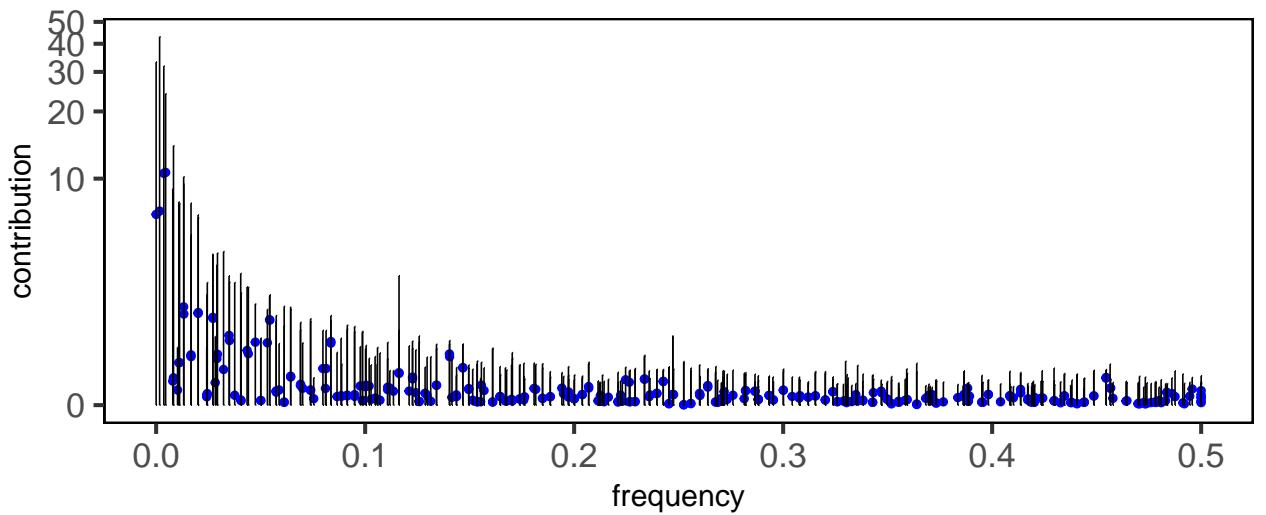


Рис. 2.8. Результат работы MC-SSA для ряда Nile Minima

будем использовать именно ее.

Поскольку ряд достаточно длинный, чтобы не делать поправку, рассмотрим в качестве векторов для проекции косинусы с равнотоящими частотами с шагом  $1/(2L)$ . Также выберем длину окна не слишком большой, чтобы метод MC-SSA считался за адекватное время, скажем,  $L = 365$ . На рис. 2.10 представлен результат работы MC-SSA для обоих моделей, уровень значимости, как и в прошлом примере, равен  $\alpha = 0.05$ . Для модели ARFIMA(0,  $d$ , 0) значимых векторов два, их периоды равны  $365/34 \approx 10.74$  и  $365/60 \approx 6.08$  дней соответственно, что сложно как-то интерпретировать. Однако стоит заметить, что вклады значимых векторов не слишком превосходят верхние границы соответствующих предсказательных интервалов, поэтому, скорее всего, векторы значи-

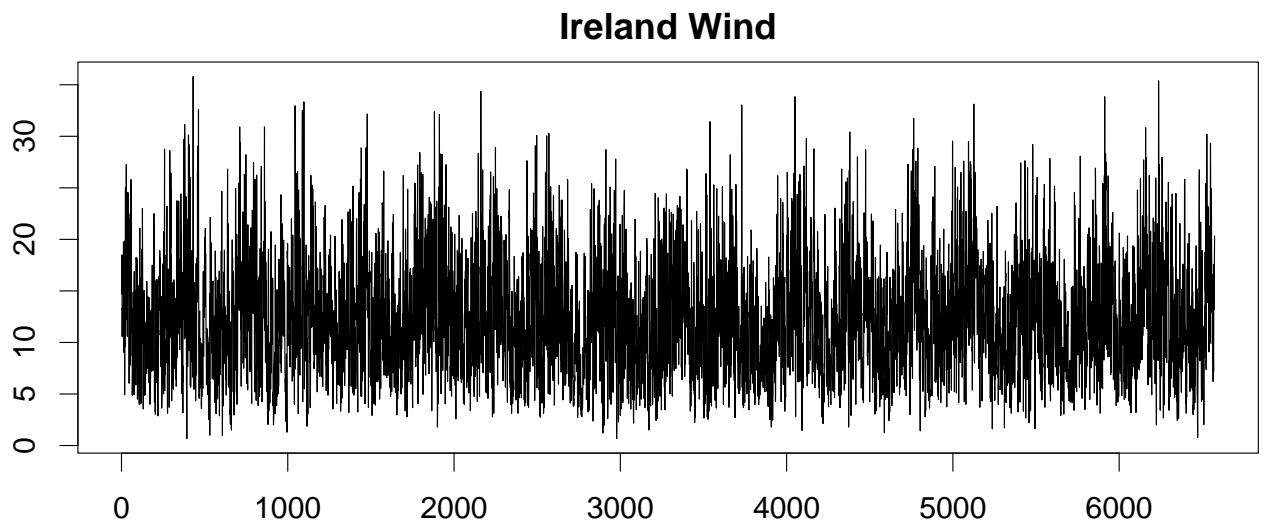


Рис. 2.9. Среднесуточные данные о скорости ветра в Республике Ирландия

Таблица 2.2. Оценка параметров ряда Ireland Wind

Метод	ARFIMA(0, d, 0)		ARFIMA(1, d, 0)		
	$\hat{d}$	$\hat{\sigma}^2$	$\hat{d}$	$\hat{\phi}$	$\hat{\sigma}^2$
MLE( $\bar{x}$ )	0.37117	24.39916	0.17306	0.28403	23.7581
H&R	0.36891	24.38116	0.17245	0.28309	23.73458
Whittle	0.37287	24.40285	0.17598	0.28105	23.75983

мы случайно (количество случайно значимых векторов в среднем равно  $\alpha \cdot L = 18.25$ ). В случае модели ARFIMA(1, d, 0) значим всего один вектор с периодом 365 дней, что интерпретируется как наличие во временном ряде годовой периодичности.

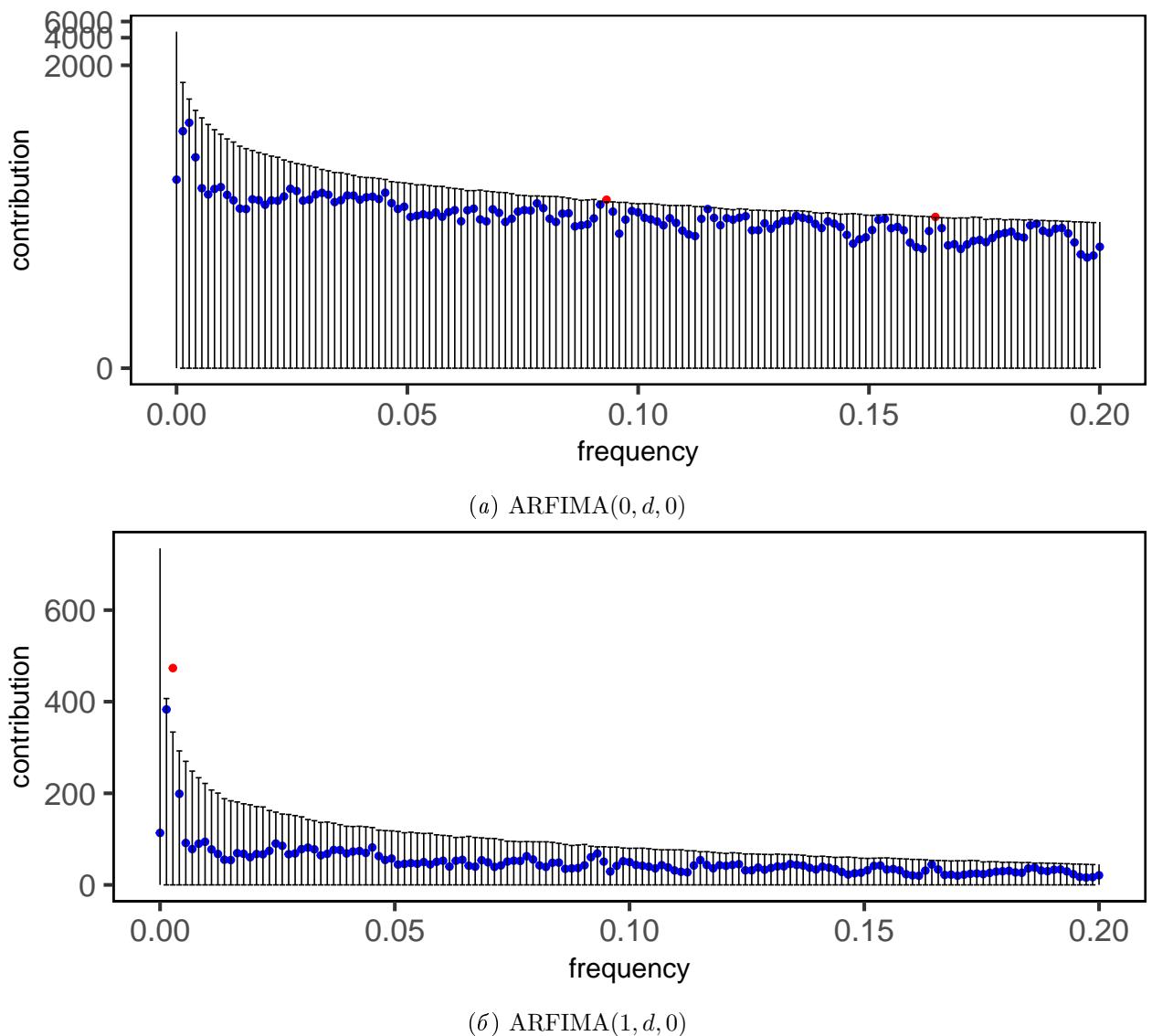


Рис. 2.10. Результат работы MC-SSA для ряда Ireland Wind

## Глава 3

### Метод autoMCSSA

#### 3.1. Сравнение способов задания проекционных векторов

Сравним два способа задания векторов для проекции  $W_k$  критерия MC-SSA (см. алгоритм 2):

1. Собственные векторы матрицы  $\mathbf{XX}^T$  или  $\mathbf{T}$ .
2. Косинусы с равноотстоящими частотами:

$$V_k = \begin{pmatrix} \cos(2\pi\omega_k 1) \\ \dots \\ \cos(2\pi\omega_k L) \end{pmatrix}, \quad W_k = \frac{V_k}{\|V_k\|}, \quad k = 1, \dots, L,$$

где  $\omega_k = k/(2L)$ .

Для краткости будем называть соответствующие им критерии MC-SSA «ev» и «cos» соответственно.

Введем понятие соотношения сигнал-шум. Обычно под ним понимают соотношение

$$\frac{\sum_{n=1}^N s_n^2 / N}{D\xi}, \quad (3.1)$$

где  $S = (s_1, \dots, s_N)$  — сигнал,  $\xi$  — шум. Но, поскольку это определение не учитывает поведение спектральной плотности шума, обобщим его.

**Определение 3.1.** Будем называть

$$\text{SNR}(S, \xi) = \frac{1}{N} \sum_{j=0}^{N-1} \frac{I(j/N)}{f_\xi(j/N)} \quad (3.2)$$

соотношением сигнал-шум, где  $I(\omega)$  — периодограмма (1.1)  $S$ ,  $f_\xi(\omega)$  — спектральная плотность  $\xi$ .

**Замечание 3.1.** Для белого шума с плотностью  $f_\xi(\omega) = \sigma^2$  формулы (3.1) и (3.2) совпадают.

Будем рассматривать ряды длины  $N = 100$  и  $L \in \{10, 20, 50, 80, 90\}$ . В качестве альтернативы рассмотрим

$$\mathbf{S} = \{Ae^{an} \cos(2\pi\omega n)\}_{n=1}^N.$$

Нас интересуют два конкретных случая:

1.  $a = 0$  — гармонический ряд;
2.  $a \neq 0$  — экспоненциально-модулированный гармонический ряд.

Помимо этого, будем рассматривать два возможных случая  $\omega$ : когда для рассматриваемых  $L$  значение  $L\omega$  (а значит и  $2L\omega$ ) является и не является целым. Таким образом, рассмотрим 4 альтернативы:

1.  $a = 0, L\omega$  — целые;
2.  $a = 0, L\omega$  — нецелые;
3.  $a \neq 0, L\omega$  — целые;
4.  $a \neq 0, L\omega$  — нецелые.

Пусть  $\omega_1 = \omega_3 = 0.1$ ,  $\omega_2 = \omega_4 = 0.085$ ,  $A_1 = 0.9$ ,  $A_3 = 0.02$ ,  $a_3 = a_4 = 0.05$ ,  $A_2$  и  $A_4$  выбираются таким образом, чтобы  $\text{SNR}(\mathbf{S}_1, \boldsymbol{\xi}) \approx \text{SNR}(\mathbf{S}_2, \boldsymbol{\xi})$  и  $\text{SNR}(\mathbf{S}_3, \boldsymbol{\xi}) \approx \text{SNR}(\mathbf{S}_4, \boldsymbol{\xi})$ . В качестве шума рассмотрим красный шум с  $\phi = 0.7$  и  $\sigma^2 = 1$ .

На рис. 3.1 изображены графики ошибок первого рода критериев «ev» и «cos» для рассматриваемых длин окна. Как видно по графикам, критерий «ev» радикальный для всех  $L > 10$ , а «cos», в свою очередь, является точным критерием для любой длины окна.

В таблице 3.1 представлены значения мощности критериев после поправки, описанной в разделе 2.1.1 для рассмотренных  $L$  при уровне значимости  $\alpha^* = 0.1$ . Прочерки в таблице указывают на то, что критерий при данной длине окна слишком радикальный и построить поправку невозможно. Стоит отметить, что при  $a \neq 0$  ряд становится нестационарным, поэтому приходится использовать собственные векторы  $\mathbf{X}\mathbf{X}^T$ , которые дают для больших  $L$  слишком радикальный критерий и тем самым имеют потенциально меньшую мощность, чем при использовании собственных векторов  $\mathbf{T}$ .

По таблице видно, что в оптимальной длиной окна при постоянной амплитуде сигнала ( $a \neq 0$ ) является  $L = 90$ . В случае же модуляции оптимальной является  $L = 10$ .

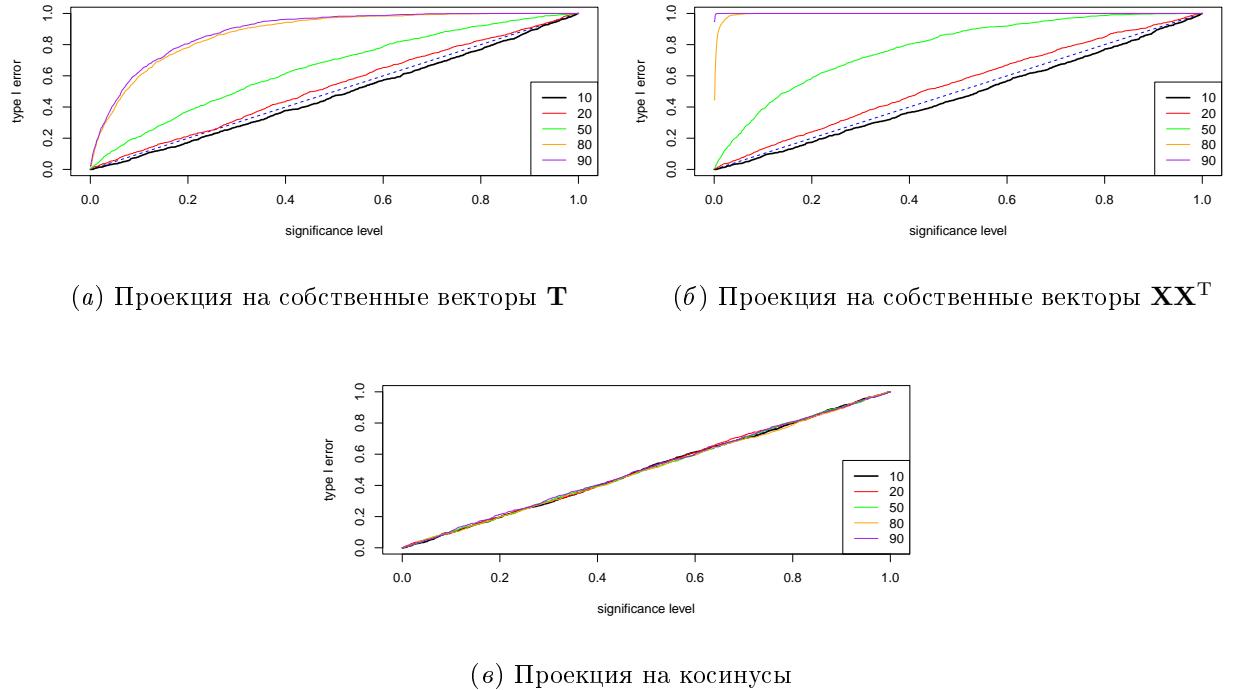


Рис. 3.1. Ошибка первого рода

Таблица 3.1. Мощность поправленных критериев при уровне значимости  $\alpha^* = 0.1$ 

$a = 0, L\omega$ — целые	$L = 10$	$L = 20$	$L = 50$	$L = 80$	$L = 90$
Проекция на собственные векторы $\mathbf{T}$	0.424	0.484	0.525	0.582	0.646
Проекция на косинусы	0.543	0.546	0.546	0.614	0.646
$a = 0, L\omega$ — нецелые	$L = 10$	$L = 20$	$L = 50$	$L = 80$	$L = 90$
Проекция на собственные векторы $\mathbf{T}$	0.349	0.432	0.425	0.497	0.537
Проекция на косинусы	0.475	0.431	0.403	0.505	0.488
$a \neq 0, L\omega$ — целые	$L = 10$	$L = 20$	$L = 50$	$L = 80$	$L = 90$
Проекция на собственные векторы $\mathbf{XX}^T$	0.329	0.303	0.307	—	—
Проекция на косинусы	0.480	0.347	0.163	0.228	0.299
$a \neq 0, L\omega$ — нецелые	$L = 10$	$L = 20$	$L = 50$	$L = 80$	$L = 90$
Проекция на собственные векторы $\mathbf{XX}^T$	0.298	0.306	0.282	—	—
Проекция на косинусы	0.469	0.302	0.157	0.209	0.249

Связано такое поведение мощности с тем, что при  $a \neq 0$  частота  $\omega$ , соответствующая сигналу, растекается по всему спектру частот и чем больше разрешение спектра (в

данном случае  $L$ ), тем сильнее это растекание. Если сравнивать по оптимальным  $L$ , то критерий «cos» дает наиболее мощный критерий в общем случае, когда модуляция непостоянная.

Учитывая огромную трудоемкость критерия «ev», и результат его численного сравнения с критерием «cos», рекомендуется использовать косинусы в качестве проекционных векторов, поскольку критерий не является радикальным (следовательно не требует поправки, что заметно уменьшает трудоемкость MC-SSA) и по мощности не уступает, а в некоторых случаях даже превосходит вариант с векторами, порожденными исходным рядом. Далее будем рассматривать именно такой способ выборка векторов  $W_k$  в алгоритме MC-SSA.

### 3.2. SSA с проекцией

Базовый вариант SSA использует адаптивный базис для оценки подпространства сигнала, но существует возможность зафиксировать некоторые компоненты разложения. Пусть  $\mathbf{D} \in \mathbb{R}^{L \times m}$  — матрица, проекцию на столбцы которой мы хотим зафиксировать в разложении  $\mathbf{X}$ . Тогда SSA с проекцией отличается от базового алгоритма только шагом разложения:

1. В случае, если столбцы матрицы  $\mathbf{D}$  не ортонормированы,  $\mathbf{D}$  приводится к нужному виду путем ортогонализации Грамма-Шмидта.
2. Вычисляется матрица  $\mathbf{C} = \mathbf{DD}^T \mathbf{X}$ .
3. Вычисляется матрица  $\mathbf{X}^* = \mathbf{X} - \mathbf{C}$ .
4. Матрица  $\mathbf{X}^*$  раскладывается в сумму матриц ранга 1.

**Замечание 3.2.** Таким же образом можно определить матрицу-проектор на подпространство строк матрицы  $\mathbf{X}$ .

### 3.3. Автоматическая группировка в SSA

Для ряда  $\mathbf{X}$  длины  $N$  и  $0 \leq \omega_1 \leq \omega_2 \leq 0.5$  определим меру, следуя [9]:

$$T(\mathbf{X}; \omega_1, \omega_2) = \frac{1}{\|\mathbf{X}\|^2} \sum_{k: \omega_1 \leq k/N \leq \omega_2} I'(k/N), \quad (3.3)$$

где

$$I'(k/N) = \begin{cases} I(k/N) + I(1 - k/N), & \text{для } 0 < k < N/2, \\ I(k/N), & \text{иначе,} \end{cases} \quad (3.4)$$

$I(\omega)$  — периодограмма (1.1) ряда  $\mathbf{X}$ . Величину  $T(\mathbf{X}, \omega_1, \omega_2)$  можно рассматривать как долю вклада частот, содержащегося в интервале  $[\omega_1, \omega_2]$ .

Зафиксируем длину окна  $L$ , порог  $T_0 \in [0, 1]$  и  $\delta > 0$ . Пусть

$$r(\omega) = \begin{cases} 2, & \text{для } \omega \in (0, 0.5), \\ 1, & \text{иначе.} \end{cases}$$

Будем выделять сигнал  $\mathbf{S}$  по частоте  $\omega$  следующим образом:

1. Применить SSA с длиной окна  $L$ .
2. Выбрать первые  $r(\omega)$  компонент разложения с мерой  $T$  на интервале  $[\omega - \delta, \omega + \delta]$ , превышающей  $T_0$ .

### 3.4. Метод autoMCSSA

На рис. 3.2 изображена блок-схема алгоритма autoMCSSA. Поскольку при оценке неизвестных параметрах шума метод MC-SSA может обнаружить не все частоты, принадлежащие сингалу, предлагается оценивать сигнал последовательно, применяя критерий MC-SSA к остатку ряда до тех пор, пока гипотеза  $H_0 : \mathbf{S} = 0$  не перестанет отвергаться. Если на очередной итерации алгоритма гипотеза отвергается, определяется максимально значимая частота  $\omega^*$  и вычисляется новая оценка сигнала  $\tilde{\mathbf{S}}$  с помощью подхода, описанного в разделе 3.3. Как только гипотеза перестает отвергаться, алгоритм завершает свою работу и тогда  $\tilde{\mathbf{S}}$  — итоговая оценка сигнала.

Остановимся поподробнее на блоке, в котором определяется максимально значимая частота  $\omega^*$ . Оценивать ее будем с помощью MC-SSA:

1. Найти индекс наиболее значимой частоты, т.е.  $k = \operatorname{argmax}_i (\hat{p}_i - c_i)$ , где  $c_i$  — верхняя граница доверительного интервала для  $\hat{p}_i$ ;
2. Вычислить значение  $\widehat{\omega^*}$  как взвешенное среднее частот  $\omega_{k-1}, \omega_k, \omega_{k+1}$  с весами  $w_i = \max(0, \hat{p}_i - c_i)$ ;

Такой способ оценки позволяет получить более точную оценку  $\omega^*$  в случае, когда она не попадает в решетку  $k/(2L)$ .

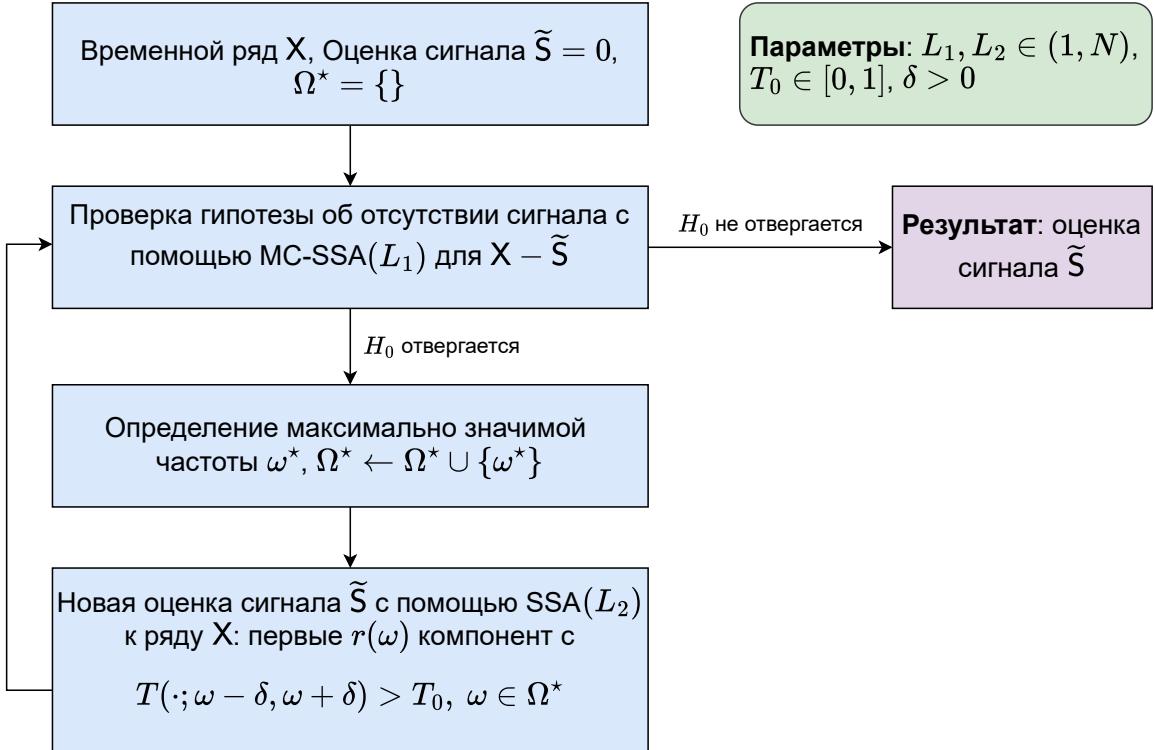


Рис. 3.2. Блок-схема алгоритма autoMCSSA

### 3.4.1. Пример работы алгоритма

Рассмотрим пример работы autoMCSSA. Пусть  $X = S + \xi$ , где  $\xi$  — красный шум с параметрами  $\phi = 0.7$  и  $\sigma^2 = 1$ ,  $N = 200$ ,  $S = (s_1, \dots, s_N)$ ,

$$s_n = 0.075 e^{0.02n} \cos(2\pi n/8) + \cos(2\pi n/4) + 0.2 \cdot (-1)^n.$$

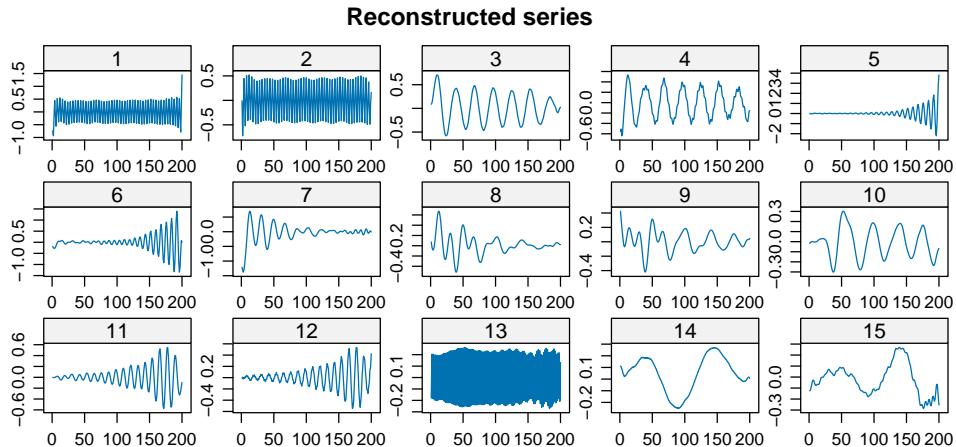


Рис. 3.3. Элементарные восстановленные компоненты

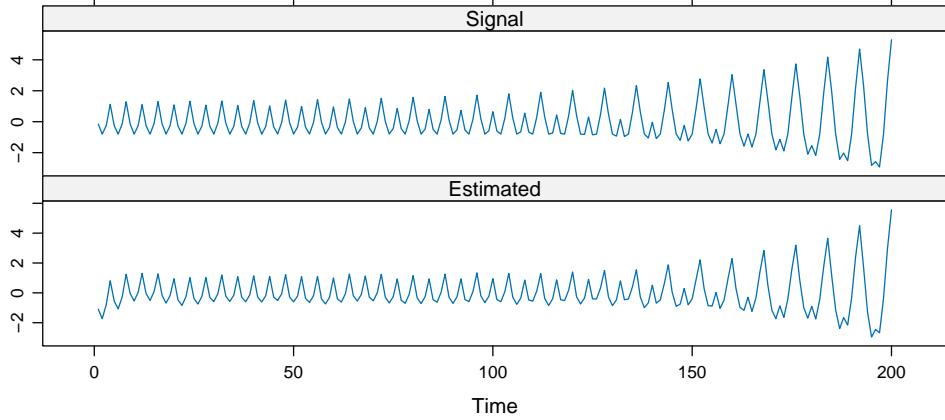


Рис. 3.4. Результат autoMCSSA ( $L_1 = 50$ ,  $L_2 = 100$ ,  $\delta = 1/80$ ,  $T_0 = 0.5$ )

На рис. 3.3 представлены первые 15 элементарных восстановленных с помощью SSA компонент. Сигналу соответствуют компоненты с индексами 1, 2, 5, 6 и 13. Не видя формулы, по которой этот сигнал задается, сказать наверняка, какие компоненты неслучайные, проблематично, поскольку компоненты 3, 4 и 11, 12 похожи на пары гармоник. Мы применили алгоритм autoMCSSA к этому ряду и получили, что разработанный метод правильно идентифицировал компоненты, соответствующие сигналу, на рис. 3.4 представлены истинная форма сигнала  $S$  и его оценка методом autoMCSSA.

**Замечание 3.3.** Для демонстрации работы алгоритма параметры  $\delta$  и  $T_0$  были подобраны так, чтобы autoMCSSA выделил весь сигнал. В общем случае выбор оптимальных  $\delta$  и  $T_0$  проблематичен. Эта проблема будет исследована в дальнейшем.

### 3.4.2. Подходы к выделению сигнала

При описании алгоритма autoMCSSA для выделения сигнала  $S$  использовался базовый вариант SSA. Можно обобщить этот подход на SSA с проекцией, зафиксировав подходящий базис.

В данном разделе будем считать, что сигнал представляет из себя экспоненциально-модулированную гармонику:

$$S = \{Ae^{an} \cos(2\pi\omega n)\}_{n=1}^N,$$

где  $\omega \in (0, 0.5)$ . Пусть  $\hat{\omega}$  — оценка  $\omega$ . Обозначим

$$D_1 = \begin{pmatrix} \cos(2\pi\hat{\omega}1) \\ \dots \\ \cos(2\pi\hat{\omega}L_2) \end{pmatrix}, D_2 = \begin{pmatrix} \sin(2\pi\hat{\omega}1) \\ \dots \\ \sin(2\pi\hat{\omega}L_2) \end{pmatrix} \in \mathbb{R}^{L_2}.$$

Рассмотрим следующие варианты выделения сигнала  $S$  по частоте  $\hat{\omega}$ :

1. «adaptive»: применить SSA и выбрать первые две компоненты разложения, у которых мера  $T$  (3.3) на интервале  $[\hat{\omega} - \delta, \hat{\omega} + \delta]$ ,  $\delta > 0$ , больше некоторого порога  $T_0 \in [0, 1]$ ;
2. «semi-adaptive»: применить SSA с проекцией с  $\mathbf{D} = D_1 \in \mathbb{R}^{L_2 \times 1}$  и выбрать, помимо компоненты, соответствующей вектору  $D_1$ , первую компоненту разложения, у которой мера  $T$  (3.3) на интервале  $[\hat{\omega} - \delta, \hat{\omega} + \delta]$ ,  $\delta > 0$ , больше некоторого порога  $T_0 \in [0, 1]$ ;
3. «fixed»: применить SSA с проекцией с  $\mathbf{D} = [D_1 : D_2] \in \mathbb{R}^{L_2 \times 2}$  и выбрать компоненты разложения, соответствующие векторам  $D_1, D_2$ .

Заметим, что вариант «adaptive» соответствует методу, который использовался в алгоритме autoMCSSA.

### Численное сравнение подходов

Проведем численный эксперимент с целью понять, какой из предложенных способов восстановления сигнала наиболее точен. Пусть  $N = 99$ , процесс  $\xi$  — красный шум с параметрами  $\phi = 0.7$ ,  $\sigma^2 \in \{0.2, 0.4, 0.6, 0.8, 1\}$ . Для SSA  $L_2 = 50$ , для MC-SSA  $L_1 = 40$  (выводы устойчивы к выбору длины окна). В вариантах «adaptive» и «semi-adaptive»  $\delta = 0.025$  и  $T_0 = 0.5$ . Рассмотрим два типа сигнала  $S$ , один из которых является частным случаем другого:

1.  $a = 0, A = 1$  — гармоника с постоянной амплитудой.
2.  $a = 0.05, A = 0.025$  — экспоненциально-модулированная гармоника.

Возьмем  $\omega = 0.115$ . Заметим, что при таком выборе частоты сигнала  $L_1\omega$  нецелое, а значит  $\omega$  не попадает в решетку  $k/(2L_1)$ . Оценивать частоту  $\omega$  будем также, как в разделе 3.4.

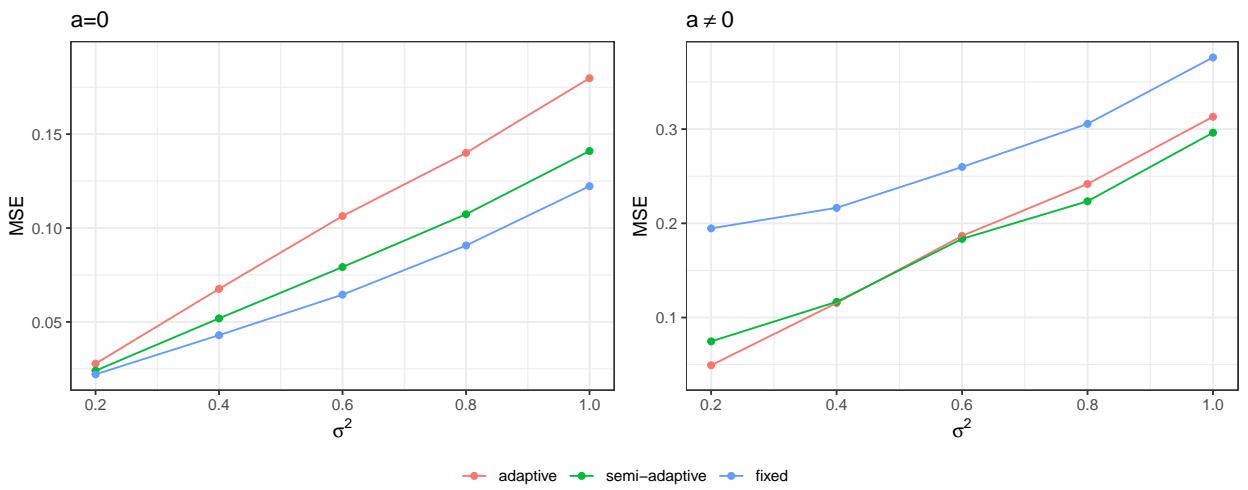
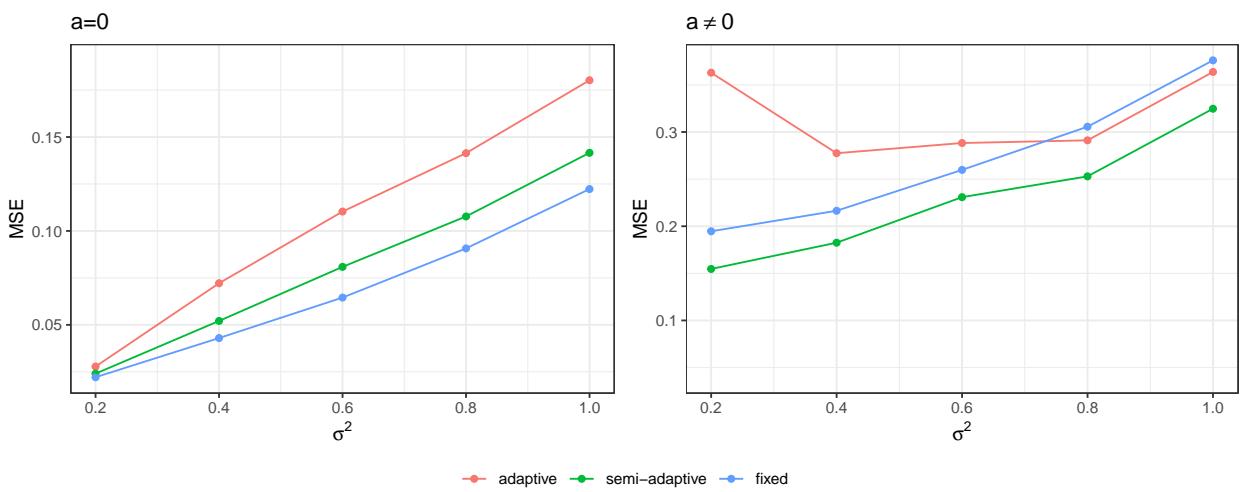
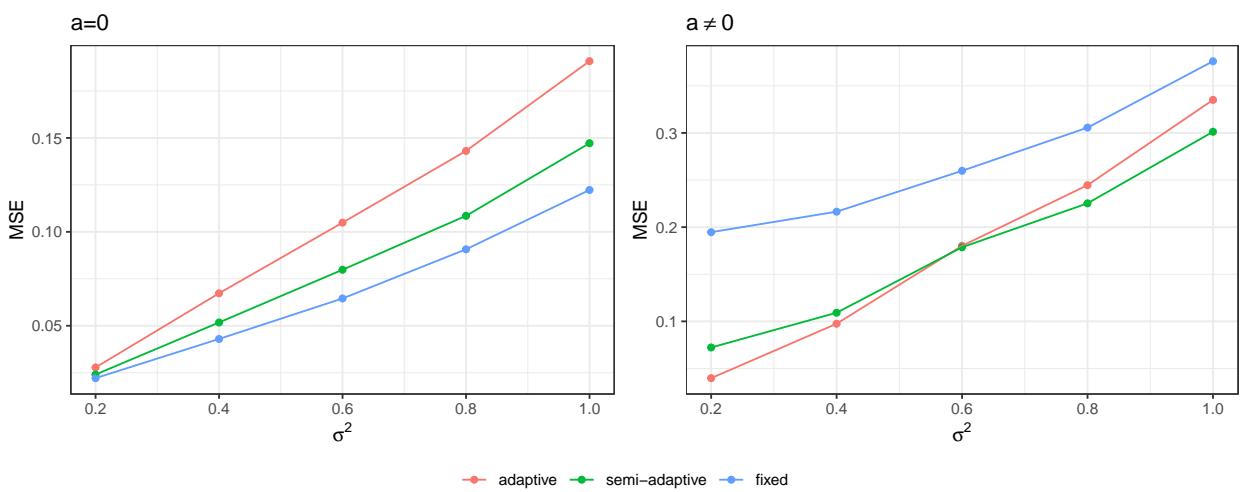
На рис. 3.5 изображена зависимость MSE восстановления сигнала от дисперсии белого шума  $\sigma^2$ . По графикам видно, что в случае постоянной амплитуды ( $a = 0$ ) выигрывает вариант «fixed», однако в случае непостоянной амплитуды фиксированный базис оказывается наихудшим. Полуадаптивный базис, являясь неким компромиссом между адаптивным и фиксированным базисами, оказывается вторым по точности в случае  $a = 0$  и сравнимым с адаптивным в рассмотренном случае  $a \neq 0$ . При увеличении  $|a|$ , начиная с какого-то момента, фиксированная половина базиса ухудшает восстановление сигнала.

Теперь посмотрим, как будут изменяться ошибки при уменьшении/увеличении  $\delta$  для фиксированного  $L_1$ . На рис. 3.6  $\delta$  уменьшена, а на рис. 3.7 увеличена в два раза ( $\delta = 0.0125$  и  $0.05$  соответственно). Из этих графиков видно, что слишком маленькое  $\delta$  приводит к ухудшению точности адаптивного и полуадаптивного вариантов в случае  $a \neq 0$ . Связано это с тем, что частота экспоненциально-модулированной гармоники, в отличие от гармоники с постоянной амплитудой, всегда растекается по спектру и чем больше абсолютное значение показателя экспоненты  $a$ , тем сильнее это растекание. Увеличение  $\delta$  в два раза не привело к значительному изменению точности методов, однако, если и дальше увеличивать  $\delta$ , ошибки, как и в случае слишком маленького  $\delta$ , опять возрастут.

Исходя из полученных результатов, можно сделать следующие выводы. Во-первых, параметр  $\delta$ , определяющий длину частотного интервала при отборе компонент, не следует выбирать слишком маленьким, если есть подозрения на амплитудную модуляцию сигнала. Во-вторых, при отсутствии или наличии умеренной амплитудной модуляции следует выбирать полуадаптивный метод, имеющий наименьшее MSE восстановления сигнала среди рассмотренных подходов. Если применить autoMCSSA к рассмотренному в разделе 3.4.1 примеру, выделяя сигнал полуадаптивным подходом, то полученная оценка сигнала сравнима с результатом адаптивного подхода.

### 3.5. Выбор оптимальных параметров autoMCSSA

Как было показано в разделе 3.4.2, оптимальная величина частотного интервала зависит от величины  $|\alpha|$ . Обозначим за  $C = \alpha N$  величину амплитудной модуляции. Величина  $C$  имеет следующий смысл: значение гармоники в конце временного промежутка увеличилось в  $e^C$  раз.

Рис. 3.5. MSE восстановления сигнала ( $\delta = 0.025$ )Рис. 3.6. MSE восстановления сигнала ( $\delta = 0.0125$ )Рис. 3.7. MSE восстановления сигнала ( $\delta = 0.05$ )

Пусть  $\Pi(k/N) = I'(k/N)/\sum_k I'(k/N)$  — нормированная периодограмма с  $I'(k/N)$  из формулы (3.4). Обозначим за  $\Pi^{C,\omega}$  нормированную периодограмму для сигнала с частотой  $\omega$  и амплитудной модуляцией  $C$ . Отсортируем частоты  $k/N$  по убыванию расстояния до частоты  $\omega$ , обозначим их как  $\omega_{(m)}$ . Рассмотрим следующие  $\delta$ :

1. Если  $\omega \leq \omega_{(1)}$ , то

$$\Delta = \{\omega - \omega_{(m)} \mid \omega_{(m)} \leq \omega, m = 1, \dots, \lfloor N/2 \rfloor\}.$$

2. В противном случае

$$\Delta = \{\omega_{(m)} - \omega \mid \omega_{(m)} \geq \omega, m = 1, \dots, \lfloor N/2 \rfloor\}.$$

Упорядочим  $\delta \in \Delta$  по возрастанию и обозначим их как  $\delta_{(n)}$ . Введем частичные суммы вокруг главной частоты:

$$\Sigma_n^{C,\omega} = \sum_{m: |\omega_{(k)} - \omega| \leq \delta_{(n)}} \Pi^{C,\omega}(\omega_{(m)}).$$

Тогда рассмотрим следующий выбор  $\delta$  в зависимости от величины  $C$ : берем такое  $\delta^* = \delta_{(n)}$ , чтобы  $\Sigma_n^{C,\omega} > \Sigma_0$  и  $\Sigma_n^{C,\omega} - \Sigma_{n-1}^{C,\omega} < \varepsilon$ . То есть выбираем такое  $\delta$ , чтобы захватывать большую часть спектра и чтобы рост вклада при дальнейшем увеличении  $\delta$  был относительно незначительным.

Рассмотрим пример. Пусть  $N = 200$ ,  $C = 2$ ,  $\omega = 0.253$ . На рис. 3.8 изображена нормированная периодограмма такого сигнала, красной пунктирной линией обозначена главная частота. Одним цветом покрашены значения  $\Sigma_n^{C,\omega} - \Sigma_{n-1}^{C,\omega}$  для  $n > 1$  и  $\Sigma_n^{C,\omega}$  для  $n = 1$ , то есть при переходе от  $n$  к  $n + 1$  количество частот, участвующих в сумме, увеличивается на 2. Пусть  $\Sigma_0 = 0.9$ ,  $\varepsilon = 0.05$  (рост вклада ниже 5% считается незначительным), тогда  $\delta^*(\omega, C) = 0.013$ , границы частотного интервала обозначены на графике черными пунктирными линиями.

Для проверки, насколько такой выбор  $\delta$  оптimalен, сравним его с другим (итеративным) подходом к выбору  $\delta$ : будем увеличивать  $n$  до тех пор, пока для  $\delta = \delta_n$  не обнаружится  $r(\omega)$  компонент.

### 3.5.1. Численное сравнение

Проведем численный эксперимент и посчитаем ошибку восстановления сигнала для обоих способов выбора  $\delta$ . Не ограничивая общности, будем считать, что  $C \geq 0$ ,

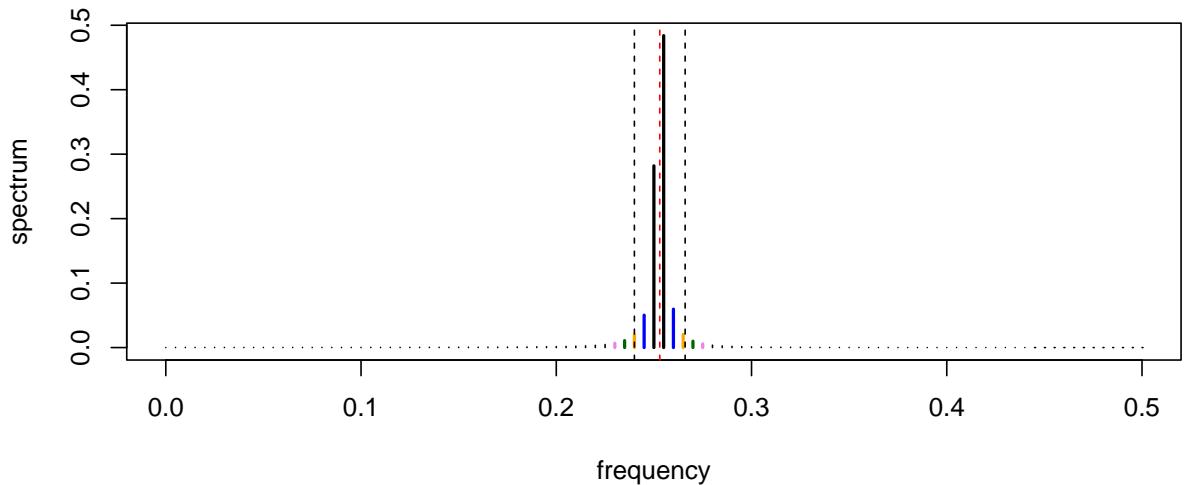


Рис. 3.8.  $\Pi^{C,\omega}(k/N)$  и границы частотного интервала при  $\delta = \delta^*$  ( $N = 200$ ,  $C = 2$ ,  $\omega = 0.253$ ,  $\Sigma_0 = 0.9$ ,  $\varepsilon = 0.05$ )

поскольку форма периодограммы для  $C$  и  $-C$  совпадают. Для исследования были выбраны следующие  $C$ : 0 (отсутствие модуляции), 1 (слабая модуляция), 2 (умеренная модуляция), 5 (сильная модуляция). Учитывая, что, помимо  $\delta$ , существует параметр  $T_0$  (см. раздел 3.3), выбор которого тоже затруднителен, рассмотрим сетку из возможных значений порога.

Пусть  $N = 200$ ,  $\omega = 0.25$ . Параметры шума  $\phi = 0.7$ ,  $\sigma^2 = 4$ . Амплитуды  $A$  для гармоник выбирать будем так, чтобы их SNR по формуле (3.2) совпадали и равнялись 0.25. Для вычисления  $\delta^*$   $\varepsilon = 0.05$ .

**Замечание 3.4.** Поскольку основное преимущество autoMCSSA — умение выделять сигнал с недоминирующими компонентами, частота сигнала и дисперсия белого шума были выбраны так, чтобы сигнал был отделен от шума и чтобы компоненты сигнала в SSA не были доминирующими. Далее в примерах будет рассмотрен сигнал с меньшей частотой, однако стоит учитывать, что в таком случае при корректировке SNR компоненты сигнала выйдут на первые позиции из-за вида спектральной плотности красного шума (невозрастающая функция).

На рис. 3.9 представлены MSE восстановления сигнала для двух подходов к выбору  $\delta$  для разных значений порога  $T_0$ . Стоит отметить следующую зависимость ошибки от порога: при увеличении  $T_0$  ошибки сначала уменьшаются, достигая минимума, а затем

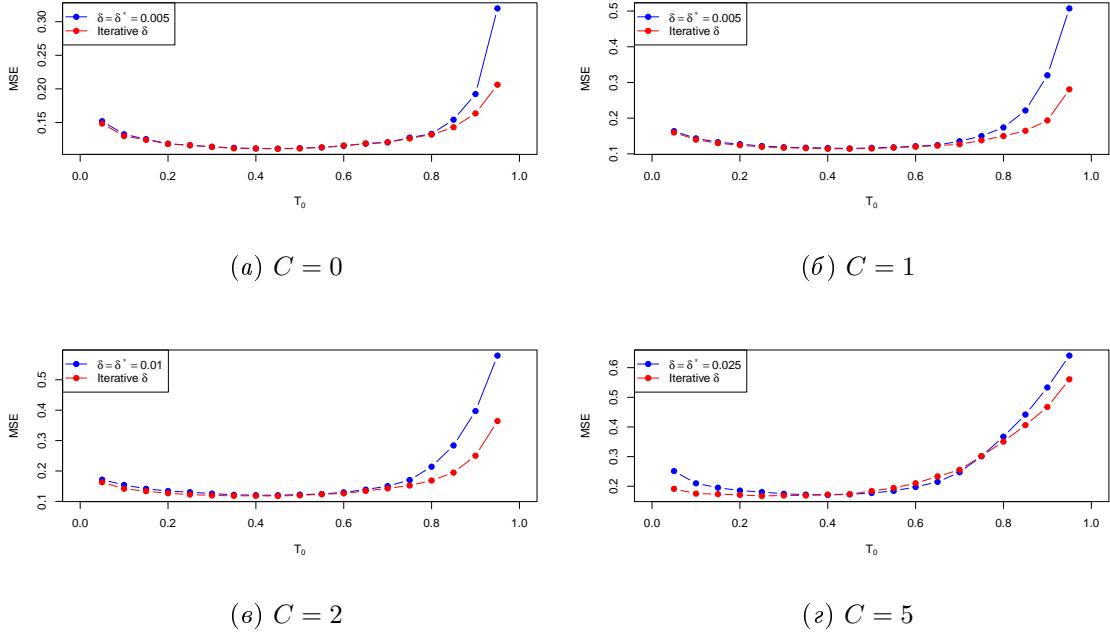


Рис. 3.9. Ошибка восстановления сигнала для разных  $C$  и  $T_0$  ( $\omega = 0.25$ )

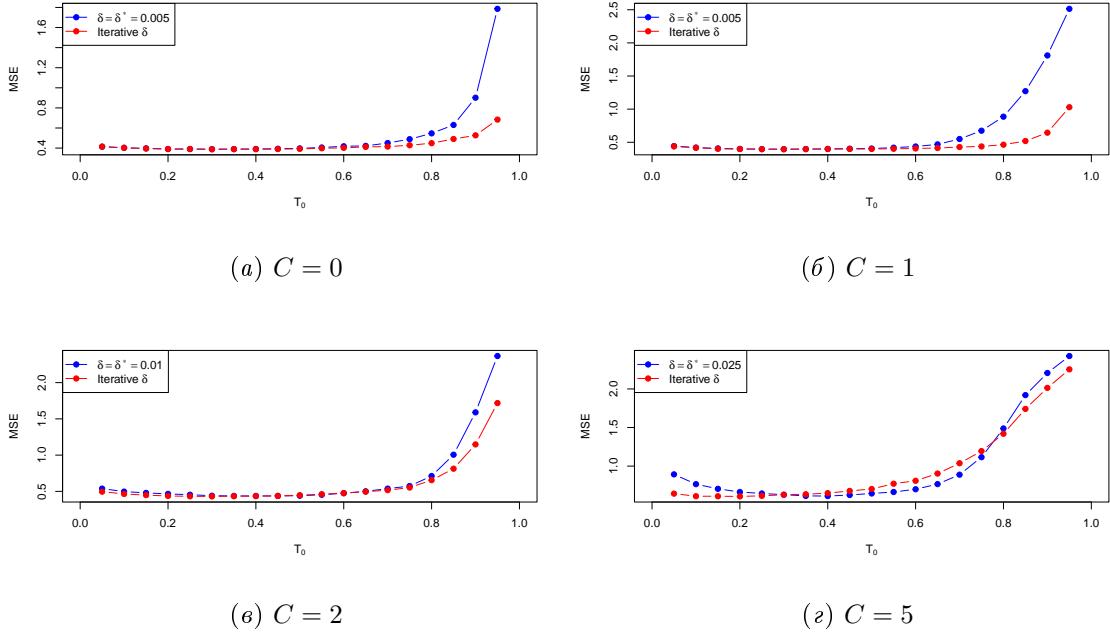


Рис. 3.10. Ошибка восстановления сигнала для разных  $C$  и  $T_0$  ( $\omega = 0.1$ )

увеличиваются. Для  $\delta = \delta^*$  этот минимум находится в интервале  $[0.4, 0.5]$ , но разница в ошибках внутри этого интервала незначительна. Можно заметить, что для  $C \geq 2$  графики для  $\delta = \delta^*$  и итеративного  $\delta$  практически совпадают, за исключением больших значений порога. Для сильной модуляции итеративный вариант дает меньшую ошибку

для малых значений  $T_0$ , но наименьшее значение MSE примерно такое же, что и у первого варианта.

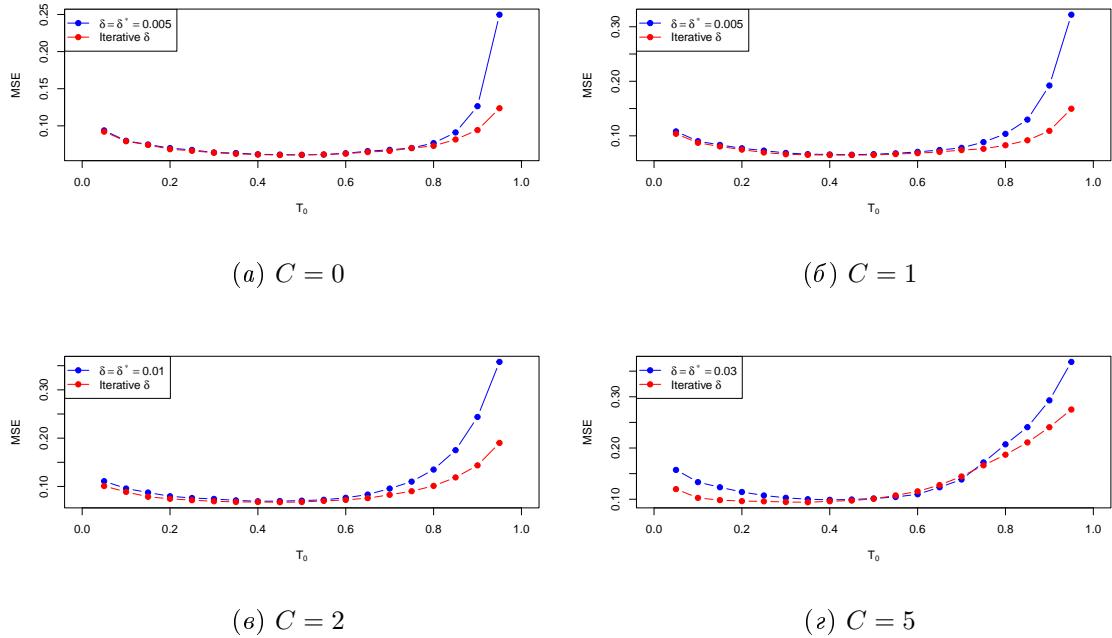


Рис. 3.11. Ошибка восстановления сигнала для разных  $C$  и  $T_0$  ( $\omega = 0.4$ )

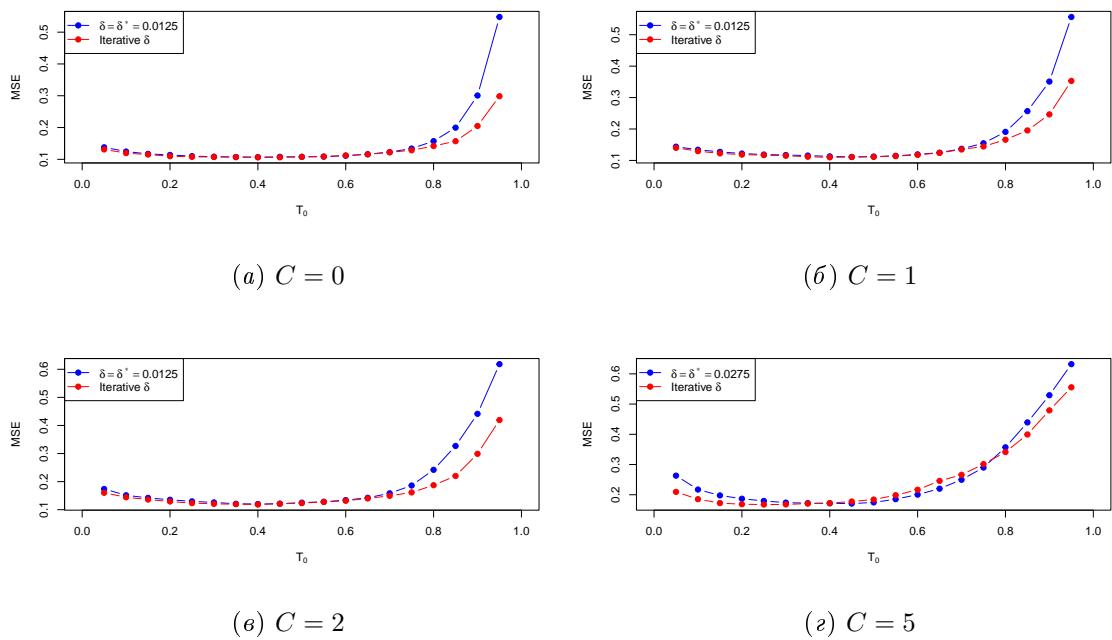


Рис. 3.12. Ошибка восстановления сигнала для разных  $C$  и  $T_0$  ( $\omega = 0.2525$ )

Увеличим/уменьшим главную частоту сигнала и проверим, можно ли сделать те же выводы, что и в случае  $\omega = 0.25$ . На рис. 3.10 и 3.11 представлены результаты

для  $\omega = 0.1$  и  $\omega = 0.4$  соответственно. Как видим, поведение графиков практически идентично случаю  $\omega = 0.25$ .

Теперь пусть частота не попадает в решетку периодограммы:  $\omega = 0.2525$  (наихудший случай, частота попала ровно посередине между частотами). Результаты моделирования представлены на рис. 3.12. Как видно, по данным графикам можно сделать аналогичные предыдущим случаям выводы.

## Выводы

По результатам численного исследования можно сделать следующие выводы:

1. Оба способа выбора  $\delta$  дают примерно одинаковую минимальную ошибку.
2. Минимальная ошибка достигается при  $T_0 \in [0.4, 0.5]$ .

Учитывая это и некоторые вычислительные сложности итеративного метода, предпочтение отдается варианту с  $\delta = \delta^*$ .

### 3.5.2. Случай неизвестной амплитудной модуляции

В предыдущем разделе предполагалось, что истинная модуляция  $C$  известна, что, конечно же, не верно на практике. Поэтому можно допустить максимально возможную (но не слишком большую) в ряде модуляцию  $C_{\max}$  и брать  $\delta = \delta^*(\omega, C_{\max})$ .

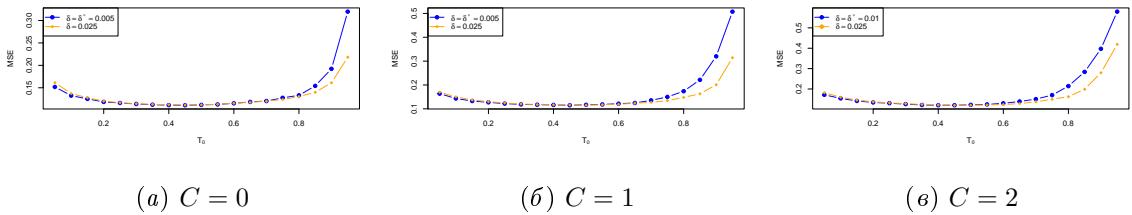


Рис. 3.13. Ошибка восстановления сигнала для разных  $C$  и  $T_0$  ( $\omega = 0.25$ ,  $C_{\max} = 5$ )

Пусть  $C_{\max} = 5$ . На примерах из раздела 3.5.1 посмотрим на применимость данного подхода. На рис. 3.13 и рис. 3.14 представлены результаты моделирования. Видно, что оптимальный порог  $T_0$  совпадает со случаем истинного  $C$ . Можно сделать вывод, что выбор  $\delta$  по верхней границе для амплитудной модуляции не портит оценку сигнала с  $\delta = \delta^*$ .

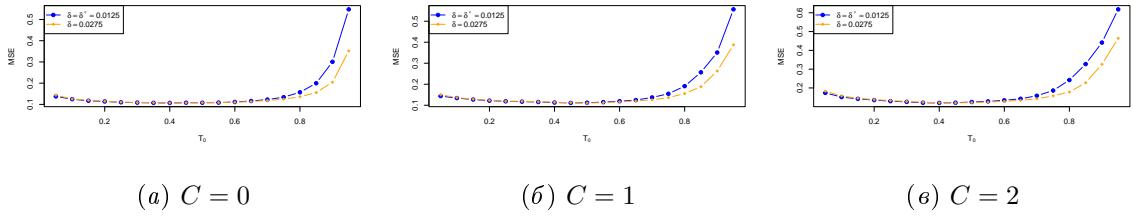


Рис. 3.14. Ошибка восстановления сигнала для разных  $C$  и  $T_0$  ( $\omega = 0.2525$ ,  $C_{\max} = 5$ )

**Замечание 3.5.** Чтобы окончательно удостовериться, что данных подход имеет место быть, он был применен на примере из раздела 3.4.1. В результате при  $C_{\max} = 5$  весь сигнал был успешно выделен.

## Заключение

В ходе данной работы был реализован метод Monte Carlo SSA, когда в качестве модели шума рассматривается процесс с длинной памятью, а также его численное сравнение с Monte Carlo SSA с моделью красного шума, обладающей такой же дисперсией. Было получено, что если в качестве альтернативы рассмотреть сигнал с некоторой частотой, большую мощность против этой альтернативы дает та модель шума, спектральная плотность которой в этой частоте наименьшая.

Помимо этого, было проведено численное сравнение различных методов оценки параметров модели ARFIMA( $p, d, q$ ). Для этого были реализованы методы максимального правдоподобия и Whittle, а также были взяты функции из пакетов языка программирования R. Получено, что при известном среднем метод максимального правдоподобия является наилучшим методом, дающим наименьшую среднеквадратичную ошибку и смещение параметров. Если же среднее неизвестно, наиболее предпочтительным методом является Whittle. Отметим, что реализованный метод максимального правдоподобия оказался лучше, чем в пакете `arfima`, а качественной реализации метода Whittle на момент написания работы обнаружено не было.

Особое внимание было уделено сравнению способов задания проекционных векторов в модели красного шума: порожденных исходным рядом и представляющих собой косинусы с равноотстоящими частотами. Численные эксперименты показали, что проекция на косинусы дает более мощный критерий в большинстве случаев, особенно при наличии экспоненциальной модуляции сигнала.

Основываясь на полученных результатах сравнения способов задания проекционных векторов, был разработан метод autoMCSSA, позволяющий автоматически идентифицировать значимые компоненты. Проведенные эксперименты подтвердили, что autoMCSSA способен корректно идентифицировать значимые компоненты, которые не обязательно доминируют по вкладу. В дальнейшем планируется заняться формулированием подхода к выбору параметров метода.

## Список литературы

1. Broomhead D. S., King G. P. Extracting qualitative dynamics from experimental data // *Physica D: Nonlinear Phenomena*. — 1986. — Vol. 20, no. 2–3. — P. 217–236.
2. Golyandina N., Nekrutkin V., Zhigljavsky A. *Analysis of Time Series Structure*. — Chapman and Hall/CRC, 2001.
3. Allen M. R., Smith L. A. Monte Carlo SSA: Detecting irregular oscillations in the Presence of Colored Noise // *Journal of Climate*. — 1996. — Vol. 9, no. 12. — P. 3373–3404.
4. Hipel K. W., McLeod A. I. *Time series modelling of water resources and environmental systems*. — Elsevier, 1994.
5. Haslett J., Raftery A. E. Space-Time Modelling with Long-Memory Dependence: Assessing Ireland’s Wind Power Resource // *Journal of the Royal Statistical Society. Series C (Applied Statistics)*. — 1989. — Vol. 38, no. 1. — P. 1–50.
6. Long memory effects and forecasting of earthquake and volcano seismic data / Mariani M. C., Bhuiyan M. A. M., Tweneboah O. K. and Gonzalez-Huizar H. // *Physica A: Statistical Mechanics and its Applications*. — 2020. — Vol. 559. — P. 125049.
7. Barkoulas J., Labys W. C., Onochie J. I. Fractional dynamics in international commodity prices // *Journal of Futures Markets*. — 1997. — Vol. 17. — P. 161–189.
8. Caporale G. M., Gil-Alana L., Plastun A. Long memory and data frequency in financial markets // *Journal of Statistical Computation and Simulation*. — 2019. — Vol. 89, no. 10. — P. 1763–1779.
9. Alexandrov Th. A Method of Trend Extraction Using Singular Spectrum Analysis // *RevStat*. — 2009. — Vol. 7, no. 1. — P. 1–22.
10. Kalantari M., Hassani H. Automatic Grouping in Singular Spectrum Analysis // *Forecasting*. — 2019. — Vol. 1, no. 1. — P. 189–204.
11. Bogalo J., Poncela P., Senra E. Circulant singular spectrum analysis: A new automated procedure for signal extraction // *Signal Processing*. — 2021. — Vol. 179. — P. 107824.
12. Golyandina N., Dudnik P., Shlemov A. Intelligent Identification of Trend Components in Singular Spectrum Analysis // *Algorithms*. — 2023. — Vol. 16, no. 7.
13. Team R Core. — R: A Language and Environment for Statistical Computing. — R Foundation for Statistical Computing, Vienna, Austria, 2024. — Access mode: <https://www.R-project.org/>.

14. Golyandina N., Korobeynikov A., Zhigljavsky A. Singular Spectrum Analysis with R. Use R! — Springer Berlin Heidelberg, 2018.
15. Time Series Analysis: Forecasting and Control / Box G., Jenkins G., Reinsel G., and Ljung G. — Fifth ed. — 2016.
16. Hassler U. Time Series Analysis with Long Memory in View. — Wiley, 2018.
17. Palma W. Long-Memory Time Series: Theory and Methods. — Wiley, 2006.
18. McLeod A. I., Yu H., Krougly Z. Algorithms for Linear Time Series Analysis: With R Package // Journal of Statistical Software. — 2007. — Vol. 23, no. 5. — Access mode: <https://www.jstatsoft.org/v23/i05/>.
19. Whittle P. The Analysis of Multiple Stationary Time Series // Journal of the Royal Statistical Society. Series B (Methodological). — 1953. — P. 125–139.
20. Veenstra J. Q. Persistence and Anti-persistence: Theory and Software : Ph. D. thesis ; Western University. — 2012.
21. Maechler M. — fracdiff: Fractionally Differenced ARIMA aka ARFIMA(P,d,q) Models : 2024. — R package version 1.5-3. Access mode: <https://CRAN.R-project.org/package=fracdiff>.
22. Ларин Е. С. Метод SSA для проверки гипотезы о существовании сигнала во времени ряде : квалификационная работа магистра ; СПбГУ. — 2022.
23. Golyandina N. Detection of signals by Monte Carlo singular spectrum analysis: multiple testing // Statistics and Its Interface. — 2023. — Vol. 16, no. 1. — P. 147–157.
24. Gray R. M. Toeplitz and Circulant Matrices: A Review // Foundations and Trends® in Communications and Information Theory. — 2005. — Vol. 2, no. 3. — P. 155–239.
25. Beran J. Statistics for Long-Memory Processes. — Chapman & Hall/CRC, 1994.

## Приложение А

### Графики

#### A.1. Сравнение `arfima_mle` и `arfima`

На рис. A.1, A.2 и A.3 представлены среднеквадратичное отклонение, смещение и дисперсия оценок параметров  $\phi$  и  $d$  модели ARFIMA(1,  $d$ , 0). По ним, видно, что при  $\phi = 0.1$  на рис. A.1 оценки функцией `arfima` имеют скачок в  $d = 0.4$ , что может говорить о некоторой вычислительной неустойчивости функции для больших  $d$ . Функция `arfima_mle` не только не имеет подобной проблемы, но и дает более точные оценки, например, при  $\phi = 0.5$  на рис. A.2.

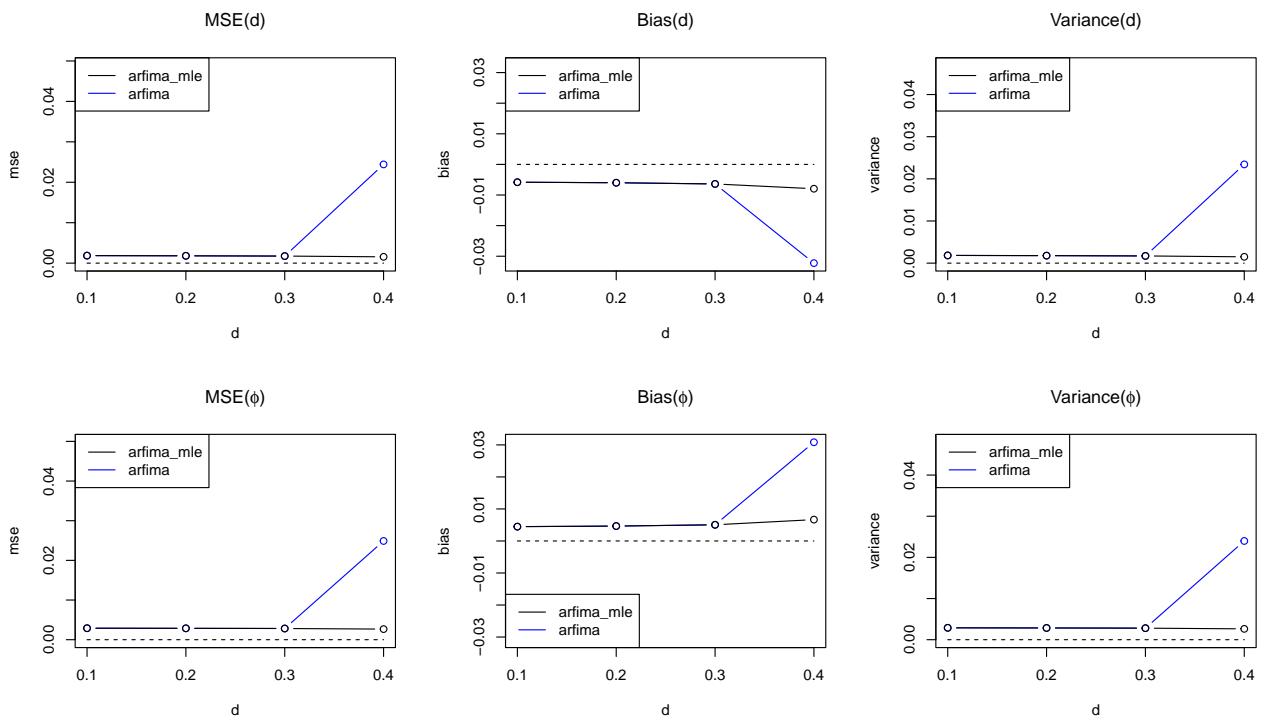


Рис. A.1. Сравнение `arfima_mle` и `arfima` при  $\phi = 0.1$  ( $n = 1000$ , 500 повторений)

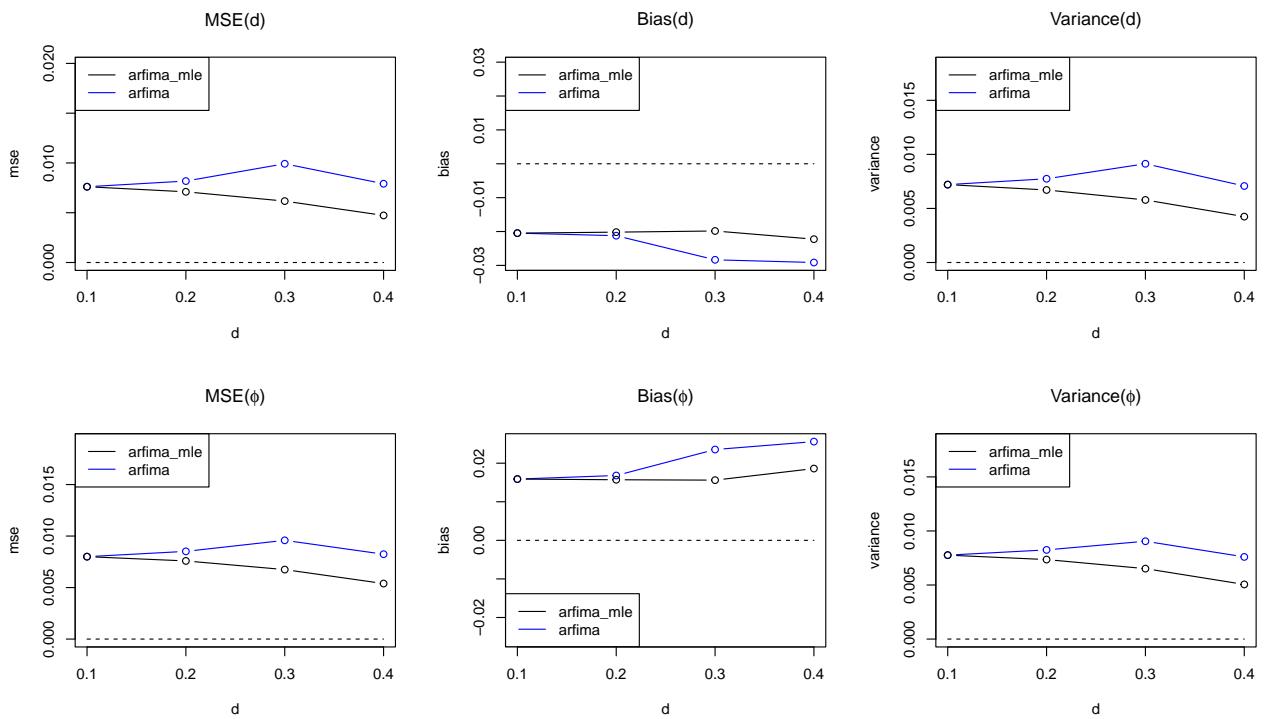


Рис. A.2. Сравнение `arfima_mle` и `arfima` при  $\phi = 0.5$  ( $n = 1000$ , 500 повторений)

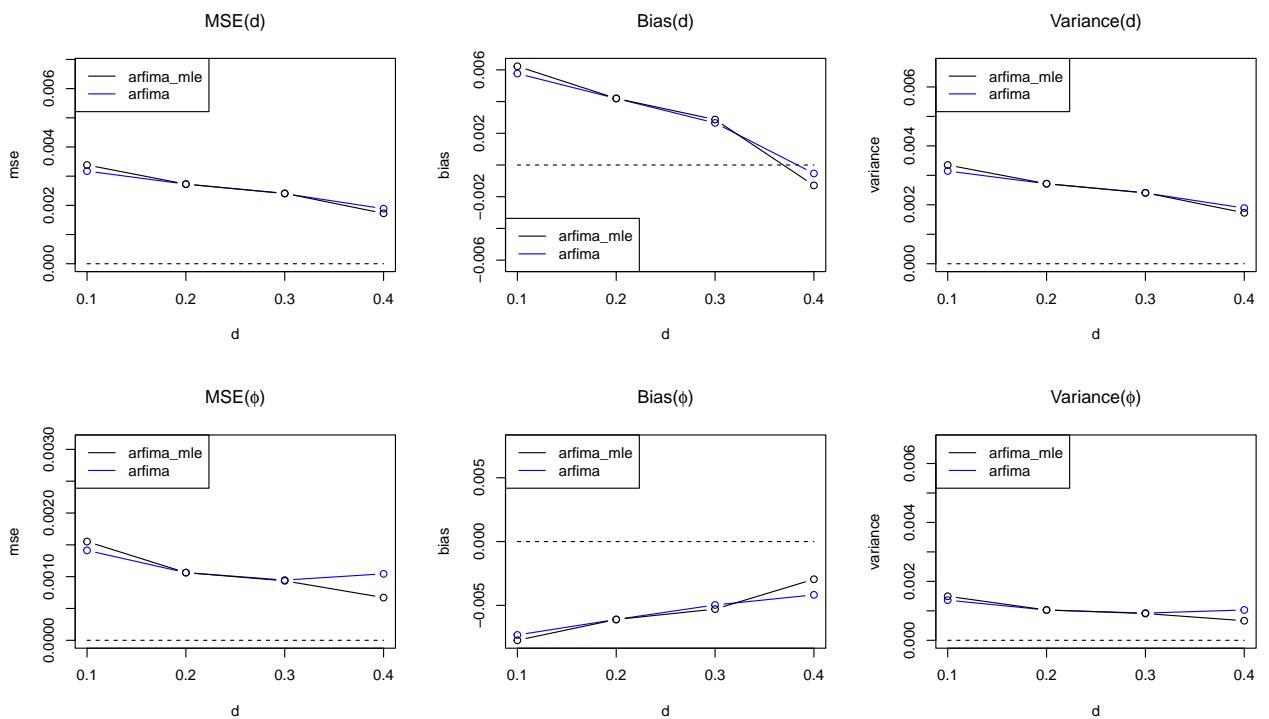


Рис. A.3. Сравнение `arfima_mle` и `arfima` при  $\phi = 0.9$  ( $n = 1000$ , 500 повторений)