Санкт-Петербургский государственный университет
Прикладная математика и информатика

Отчет по учебной практике 1 (проектно-технологическая) (семестр 1)

Выполнил:
Потешкин Егор Павлович
группа 24.М22-мм

Научный руководитель:
к. ф.-м. н., доцент
Голяндина Нина Эдуардовна
Кафедра Статистического Моделирования

Санкт-Петербург
2024

Оглавление

Введение

TODO

Глава 1

Теория случайных процессов

1.1. Вспомогательные определения

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

Определение 1.1.1. Случайный процесс $\{Y_t : t \in \mathbb{Z}\}$ называют стационарным (в широком смысле), если

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

Замечание 1.1.1. Поскольку $\gamma(0) = \mathsf{cov}(Y_t, Y_t) = \mathsf{D}Y_t$, то дисперсия также не меняется со временем.

Замечание 1.1.2. Далее под стационарностью будет подразумеваться именно стационарность в широком смысле.

Определение 1.1.2. Случайный процесс $\{\varepsilon_t\}$ называют белым шумом $\mathrm{WN}(0, \sigma^2)$, если он стационарный, $\mathsf{E}\varepsilon_t = 0$, $\gamma(h) = 0\ \forall h \neq 0$ и $\mathsf{D}\varepsilon_t = \sigma^2$.

Определение 1.1.3. Моделью $\mathrm{ARMA}(p, q)$, 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 $\{\varepsilon_t\} \sim \mathrm{WN}(0, \sigma^2)$.

Определение 1.1.4. Процесс $\{X_t\}$ называют красным шумом с параметрами ϕ и σ^2, если $\{X_t\}$ — модель $\mathrm{ARMA}(p, q)$ с $p = 1$, $q = 0$ и $\phi = \phi_1 > 0$.

Определение 1.1.5. Спектральной плотностью стационарного процесса называется такая функция $f(\omega)$, что

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

Определение 1.1.6. Пусть $\{Y_t\}$ — стационарный процесс. Функцию

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

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

1.2. Процессы с длинной памятью

Определение 1.2.1. Говорят, что стационарный процесс $\{Y_t\}$ обладает длинной памятью, если

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

при $H \to \infty$. 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 [1, Section 3.1]. Там же показано, что они согласованы с определением 1.2.1.

Пример 1.2.1. Процессом с короткой памятью является, например, стационарная модель ARMA(p, q), поскольку $|\gamma(h)| \leqslant CR^h$, где $C > 0$ и $0 < R < 1$ [2].

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. Коэффиенты $\pi_j(d)$ удовлетворяют соотношению

$$\pi_j(d) = (-1)^j \binom{d}{j} = \frac{j-1-d}{j}\pi_{j-1}(d) = \frac{\Gamma(j-d)}{\Gamma(j+1)\Gamma(-d)}, \tag{1.1}$$

где $\Gamma(x)$ — гамма функция. Заметим, что второе равенство в формуле (1.1) 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 1.2.2. 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 (1.1), $\{X_t\}$ — стационарная и обратимая модель ARMA(p,d). Процесс $\{Y_t\}$ называют дробно интегрированной моделью ARMA или ARFIMA(p,d,q).

Предложение 1.2.1. ARFIMA(p,d,q) является стационарным процессом с нулевым средним при $d < 1/2$. Его спектральная плотность определяется выражением

$$\begin{aligned} f_Y(\omega) &= \sigma^2 4^{-d} \sin^{-2d}(\pi\omega)\, f_X(\omega) \\ &= \sigma^2 4^{-d} \sin^{-2d}(\pi\omega) \frac{\left|\theta(e^{-2\pi\omega \mathrm{i}})\right|^2}{\left|\phi(e^{-2\pi\omega \mathrm{i}})\right|^2}, \quad \omega > 0 \\ &\sim \sigma^2 \omega^{-2d} \frac{|\theta(1)|^2}{|\phi(1)|^2}, \quad \omega \to 0, \end{aligned} \tag{1.2}$$

Доказательство. См. [3, Proposition 6.1]. □

Замечание 1.2.1. Из формулы (1.2) видно, что монотонность спектральной плотности процесса $\{Y_t\}$ зависит от поведения спектральной плотности процесса $\{X_t\}$.

Следствие 1.2.1. В условиях предложения 1.2.1 при $0 < d < 1/2$

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

где

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

Доказательство. См. [3, Corollary 6.1]. □

Замечание 1.2.2. Из следствия 1.2.1 сразу следует, что ARFIMA(p,d,q) с $d \in (0, 1/2)$ обладает длинной памятью.

1.2.1. Возникновение процессов с длинной памятью

Нас интересуют процессы с монотонной спектральной плотностью, поскольку они довольно распространены в реальном мире. Такими процессами являются процессы со степенной спектральной плотностью $f(\omega) \sim \omega^{-\alpha}$, имеющие большое применение в различных областях, например, в физике, биологии, астрофизике, геофизике и экономике.

Процессы с длинной памятью, являющиеся частным случаем процессов со степенной спектральной плотностью, довольно распространены. Например, в работе [4] обнаружена длинная память в таких среднегодовых гидрологических временных рядах, как количество осадков, температура и данных 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].

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 его спектральная плотность $f_X(\omega) = f_X(\omega; \boldsymbol{\psi}, \sigma)$, где

$$\boldsymbol{\psi} = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)^{\mathrm{T}}.$$

Поставим задачу оценить параметры $\boldsymbol{\varphi}^{\mathrm{T}} = \left(d, \boldsymbol{\psi}^{\mathrm{T}}\right)$ и σ^2.

1.3.1. Maximum likelihood estimation (MLE)

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

$$\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}} \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{\Sigma}_n),$$

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

$$(2\pi)^{-n/2} \left|\boldsymbol{\Sigma}_n\right|^{-1/2} \exp\left\{-\frac{1}{2} Y^{\mathrm{T}} \boldsymbol{\Sigma}_n^{-1} Y\right\}.$$

Рассмотрим логарифм функции правдоподобия. Отбрасывая аддитивные константы, получаем

$$\ell(\boldsymbol{\varphi}, \sigma^2) = -\frac{1}{2} \ln\left|\boldsymbol{\Sigma}_n\right| - \frac{1}{2} Y^{\mathrm{T}} \boldsymbol{\Sigma}_n^{-1} Y.$$

Положим $\boldsymbol{\Gamma}_n = \boldsymbol{\Sigma}_n / \sigma^2$ и, максимизируя ℓ по σ^2, получаем

$$\ell_c(\boldsymbol{\varphi}) = -\frac{n}{2} \ln\left(S(\boldsymbol{\varphi})/n\right) - \frac{1}{2} \ln g_n(\boldsymbol{\varphi}),$$

где $S(\boldsymbol{\varphi}) = \boldsymbol{Y}^{\mathrm{T}} \boldsymbol{\Gamma}_n \boldsymbol{Y}$, $g_n(\boldsymbol{\varphi}) = |\boldsymbol{\Gamma}_n|$. Тогда

$$\widehat{\boldsymbol{\varphi}}_{\mathrm{ML}} = \operatorname*{argmax}_{\boldsymbol{\varphi}} \ell_c(\boldsymbol{\varphi}), \quad \widehat{\sigma}^2_{\mathrm{ML}} = S(\widehat{\boldsymbol{\varphi}}_{\mathrm{ML}}).$$

Замечание 1.3.1. В случае ненулевого матожидания $\mathsf{E}Y_t = \mu$, для получения оценок максимального правдоподобия нужно вместо $\boldsymbol{Y}$ рассматривать $\boldsymbol{Y} - \mu$.

Замечание 1.3.2. Для вычисления ℓ_c можно использовать алгоритм Левинсона-Дурбина, имеющий временную трудоемкость $O(n^2)$.

1.3.2. Whittle estimation

Метод максимального правдоподобия применим, когда известно матожидание μ. При неизвестном μ можно использовать его оценку $\overline{\boldsymbol{Y}}$, 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) [9]. Пусть $f(\omega, \boldsymbol{\varphi}, \sigma^2)$ — спектральная плотность $\{Y_t\}$, $I(\omega)$ — периодограмма $\boldsymbol{Y}$, тогда

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

где $m = \lfloor (n-1)/2 \rfloor$, $\omega_j = j/n$, $j = 1, 2, \ldots, m$. Заметим, что $f(\omega; \boldsymbol{\varphi}, \sigma^2) = \sigma^2 g(\omega; \boldsymbol{\varphi})$. Тогда, максимизируя ℓ_W по σ^2, получаем

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

где

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

Замечание 1.3.3. Такой метод оценки параметров можно использовать при неизвестном среднем, поскольку при ее вычислении не используется значение периодограммы в нуле.

Замечание 1.3.4. Периодограмму временного ряда можно вычислить за $O(n \log n)$ с помощью быстрого преобразования Фурье (FFT), что делает этот метод значительно быстрее MLE для больших n.

1.3.3. Численное сравнение методов оценки параметров

Сравним качество оценки параметров следующих моделей:

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

2. $p = q = 0$ — модель ARFIMA$(0, \mathrm{d}, 0)$.

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

Будем оценивать параметры этих моделей методами, реализация которых есть на языке программирования R [10]:

1. Функция arima из пакета stats, соответствующая MLE модели ARMA;

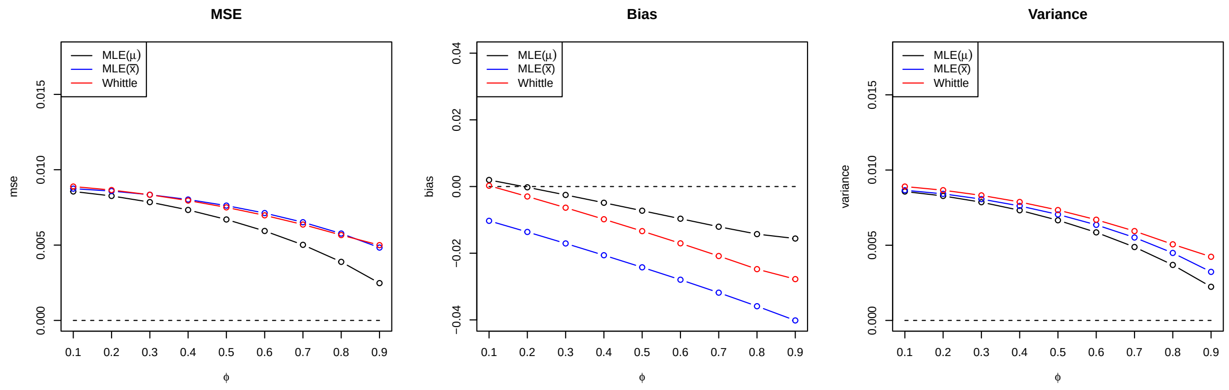
2. Функция arfima из пакета arfima [11], соответствующая MLE модели ARFIMA;

3. Функция fracdiff из пакета fracdiff [12], соответствующая апроксимации MLE модели ARFIMA, описанной в работе [5] (обозначим за H&R).

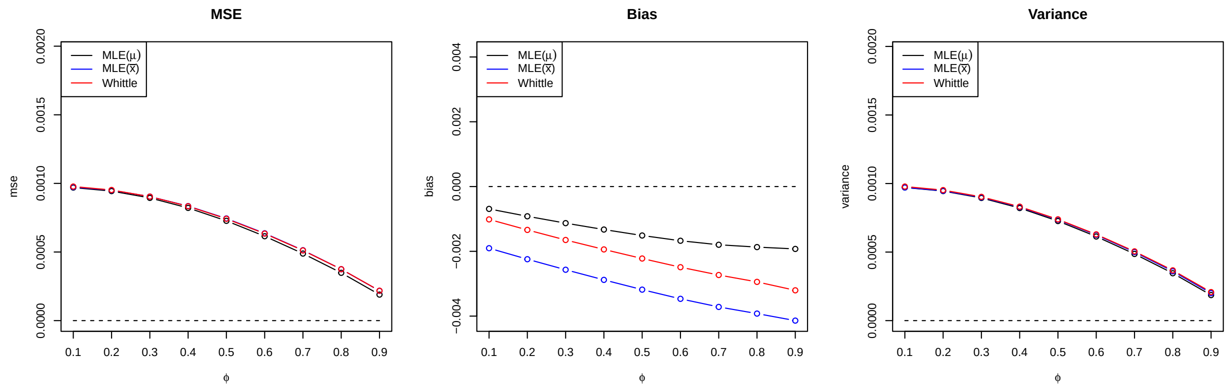
Качественная реализации оценки Уиттла не была найдена, поэтому для сравнения с MLE она была реализована самостоятельно. Поскольку для реальных временных рядов матожидание μ неизвестно, будем рассматривать MLE с известным средним и с его оценкой — выборочным средним (будем обозначать их MLE(μ) и MLE$(\bar{x})$ соответственно). Не умаляя общности, пусть $\mu = 0$.

На рис. 1.1 и 1.2 изображены среднеквадратичное отклонение, смещение и дисперсия оценок параметров ϕ 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 сренее, оценки становятся сильно смещенными. Whittle же, в свою очередь, дает менее смещенную оценку, чем MLE$(\bar{x})$, а в случае оценки d имеет смещение даже меньше, чем у MLE(μ). Однако оценки Whittle обладают наибольшей дисперсией среди всех рассмотренных методов, но разница не такая значительная, как в случае смещения.

В таблице 1.1 представлены значения среднеквадратичной ошибки и смещения оценок параметров d и ϕ модели ARFIMA$(1, d)$. Заметим, что в оценках присутствует смещение, и оно уменьшается с увеличением ϕ. Также для $\phi = 0.1$ и $\phi = 0.5$ оценки d имеют отрицательное смещение, а оценки ϕ, наоборот, положительное. Также в таблице синим цветом выделена лучшая по строке оценка d, а красным — лучшая оценка ϕ. Видно, что метод H&R в большинстве случаев дает оценки с наименьшим MSE, однако оценки имеют большое смещение при $\phi = 0.1$ и $\phi = 0.5$. Наиболее смещенные оценки



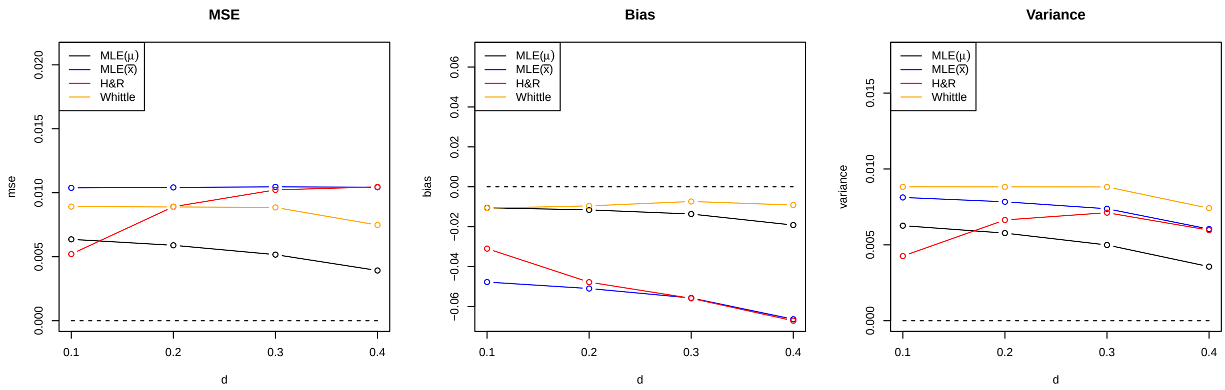
$(a)\ n = 100$



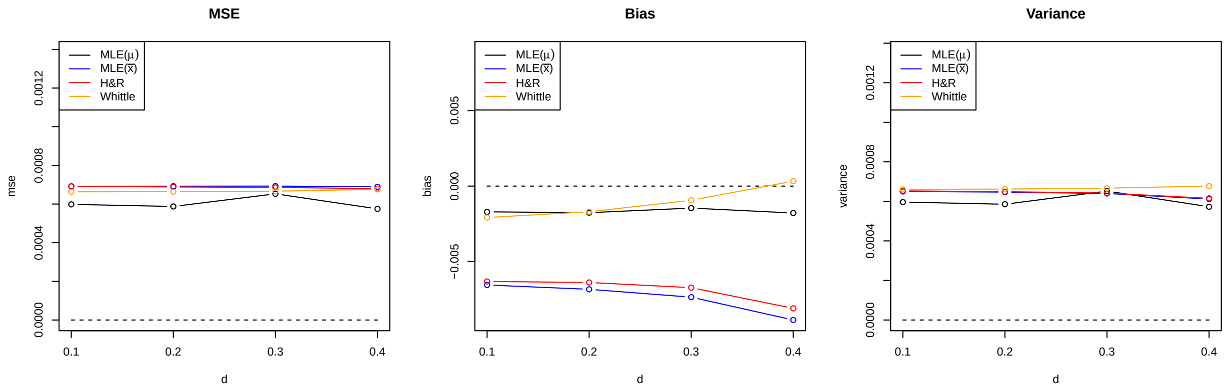
$(б)\ n = 1000$

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

получаются в методе MLE($\bar{x}$), это перестает быть таковым при $\phi = 0.9$. Оценки методом Whittle по смещению находятся посередине, однако имеют наибольшую дисперсию среди остальных.



$(a)\ n = 100$



$(б)\ n = 1000$

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

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

		MSE								Bias							
		MLE(μ)		MLE($\bar{x}$)		H&R		Whittle		MLE(μ)		MLE($\bar{x}$)		H&R		Whittle	
d	ϕ	$\hat{d}$	$\hat{\phi}$	$\hat{d}$	$\hat{\phi}$	$\hat{d}$	$\hat{\phi}$	$\hat{d}$	$\hat{\phi}$	$\hat{d}$	$\hat{\phi}$	$\hat{d}$	$\hat{\phi}$	$\hat{d}$	$\hat{\phi}$	$\hat{d}$	$\hat{\phi}$
0.1	0.1	0.071	0.075	0.205	0.179	0.009	0.018	0.069	0.067	-0.096	0.083	-0.301	0.26	-0.054	0.035	-0.086	0.068
0.2	0.1	0.064	0.07	0.224	0.201	0.025	0.032	0.077	0.073	-0.095	0.083	-0.336	0.296	-0.119	0.099	-0.094	0.074
0.3	0.1	0.081	0.086	0.303	0.272	0.049	0.055	0.084	0.081	-0.12	0.108	-0.427	0.386	-0.179	0.161	-0.109	0.09
0.4	0.1	0.072	0.08	0.367	0.328	0.081	0.089	0.179	0.194	-0.119	0.111	-0.499	0.456	-0.243	0.23	-0.26	0.241
0.1	0.5	0.057	0.048	0.099	0.059	0.01	0.015	0.057	0.054	-0.102	0.065	-0.258	0.185	-0.07	0.033	-0.066	0.024
0.2	0.5	0.055	0.045	0.107	0.062	0.031	0.025	0.074	0.058	-0.108	0.071	-0.277	0.201	-0.154	0.108	-0.153	0.107
0.3	0.5	0.051	0.041	0.119	0.068	0.063	0.043	0.098	0.062	-0.114	0.079	-0.304	0.225	-0.232	0.174	-0.209	0.161
0.4	0.5	0.052	0.043	0.129	0.074	0.104	0.065	0.104	0.066	-0.131	0.098	-0.326	0.244	-0.306	0.235	-0.228	0.177
0.1	0.9	0.029	0.028	0.013	0.008	0.007	0.007	0.034	0.025	0.07	-0.085	0.003	-0.044	0.001	-0.043	0.049	-0.069
0.2	0.9	0.019	0.021	0.011	0.006	0.009	0.004	0.026	0.019	0.038	-0.063	-0.013	-0.034	-0.037	-0.026	0.02	-0.056
0.3	0.9	0.013	0.016	0.009	0.004	0.014	0.003	0.022	0.015	0.007	-0.043	-0.033	-0.022	-0.076	-0.011	-0.024	-0.039
0.4	0.9	0.009	0.018	0.009	0.004	0.025	0.002	0.028	0.01	-0.019	-0.036	-0.059	-0.011	-0.121	0.003	-0.095	-0.016

Глава 2

Метод Monte Carlo SSA

2.1. Проверка статистических гипотез

Рассмотрим некоторый критерий со статистикой T. Введем обозначения.

Определение 2.1.1. Ошибка первого рода — вероятность отвергнуть нулевую гипотезу, если она верна: $\alpha_I(\alpha) = \mathsf{P}_{H_0}(T \in A_{\text{крит}}(\alpha))$.

Определение 2.1.2. Если $\alpha_I = \alpha$, то говорят, что критерий точный при уровне значимости α, иначе говорят, что критерий неточный. При $\alpha_I < \alpha$ критерий является консервативным, а при $\alpha_I > \alpha$ — радикальным.

Определение 2.1.3. Мощность критерия — вероятность отвергнуть нулевую гипотезу, если верна альтернативная: $\beta(\alpha) = \mathsf{P}_{H_1}(T \in A_{\text{крит}}(\alpha))$.

2.1.1. Сравнение критериев

Точные критерии, проверяющие одну и ту же гипотезу, можно сравнивать по мощности: чем больше мощность, тем лучше. Если критерий является консервативным, сравнивать его с другими критерии по мощности также можно, учитывая, что мощность консервативного критерия будет занижена. Если же критерий является радикальным, его нельзя сравнивать по мощности с другими критериями. Поэтому введем понятие ROC-кривой, которая является аналогом мощности в случае неточных критериев.

Определение 2.1.4. ROC-кривая — это кривая, задаваемая параметрически

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

Замечание 2.1.1. 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 $\alpha_I(\alpha) = \alpha$.

2.1.2. Поправка неточных критериев

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

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 $\alpha_I(\alpha)$ [13]

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

2.2. Monte Carlo SSA

Метод Monte Carlo SSA (MC-SSA) тестно связан с методом SSA (Singular Spectrum Analysis), стостоящий из четырех этапов: *вложения*, *разложения*, *группировки* и *диагонального усреднения*. Поэтому опишем сначала его.

2.2.1. Метод SSA

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

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

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

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

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

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

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

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:

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

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

2.2.3. Множественный тест

Зафиксируем длину окна L. Пусть $R_1, \ldots, R_G$ — реализации случайного процесса $\boldsymbol{\xi}$, которые в дальнейшем будем называть суррогатными. Обозначим за $\boldsymbol{\Xi}_k$ и $\boldsymbol{\Xi}_k$, $k = 1, \ldots, H$ траекторные матрицы ряда X и каждой суррогатной реализации соответственно. Рассмотрим H проекционных векторов $W_1, \ldots, W_H$, соответствующих некоторой частоте ω_k, $\|W_k\| = 1$, $k = 1, \ldots, H$.

Алгоритм 2. Multiple MC-SSA [14]

1. Для $k = 1, \ldots, H$ вычисляется статистика $\widehat{p}_k = \|\mathbf{X}^{\mathrm{T}} W_k\|^2$, выборка $P_k = \{p_{ki}\}_{i=1}^{G}$ с элементами $p_{ki} = \|\boldsymbol{\Xi}^{\mathrm{T}} W_k\|^2$, ее среднее μ_k и стандартное отклонение σ_k.

2. Вычисляется $\eta = (\eta_1, \ldots, \eta_G)$, где

$$\eta_i = \max_{1 \leqslant k \leqslant H} (p_{ki} - \mu_k)/\sigma_k, \quad i = 1, \ldots, G.$$

3. Находится q как выборочный $(1 - \alpha)$-квантиль η, где α — уровень значимости.

4. Нулевая гипотеза не отвергается, если
$$t = \max_{1\leqslant k\leqslant H} (\widehat{p}_k - \mu_k)/\sigma_k < q.$$

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

Замечание 2.2.1. При выборе модели шума $\boldsymbol{\xi}$ важно, чтобы спектральная плотность процесса была строго монотонной. Это связано с тем, что собственные векторы автоковариационной матрицы стационарного процесса ведут себя как синусоиды с равностоящими частотами, и при строгой монотонности спектральной плотности вклад этих векторов будет попарно различным, делая их сильно разделимыми. Если же опустить требование строгой монотонности, компоненты могут смешаться, что делает невозможным определение доминирущей частоты значимого вектора.

Замечание 2.2.2. Примерами процессов со строго монотонной спектральной плотностью являются красный шум и модель $\mathrm{ARFIMA}(0, d, 0)$.

2.2.4. Используемый вариант MC-SSA

В разделе 2.2.3 предполагалось, что векторы $W_1, \ldots, W_H$ фиксированные и не зависят от исходного ряда. Такой критерий MC-SSA является точным, то есть ошибка первого рода не превосходит заданный уровень 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 $\widehat{p}_k$.

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 2.1.2.

В качестве $W_1, \ldots, W_H$ будем брать собственные векторы матрицы $\mathbf{X}\mathbf{X}^{\mathrm{T}}$ 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.

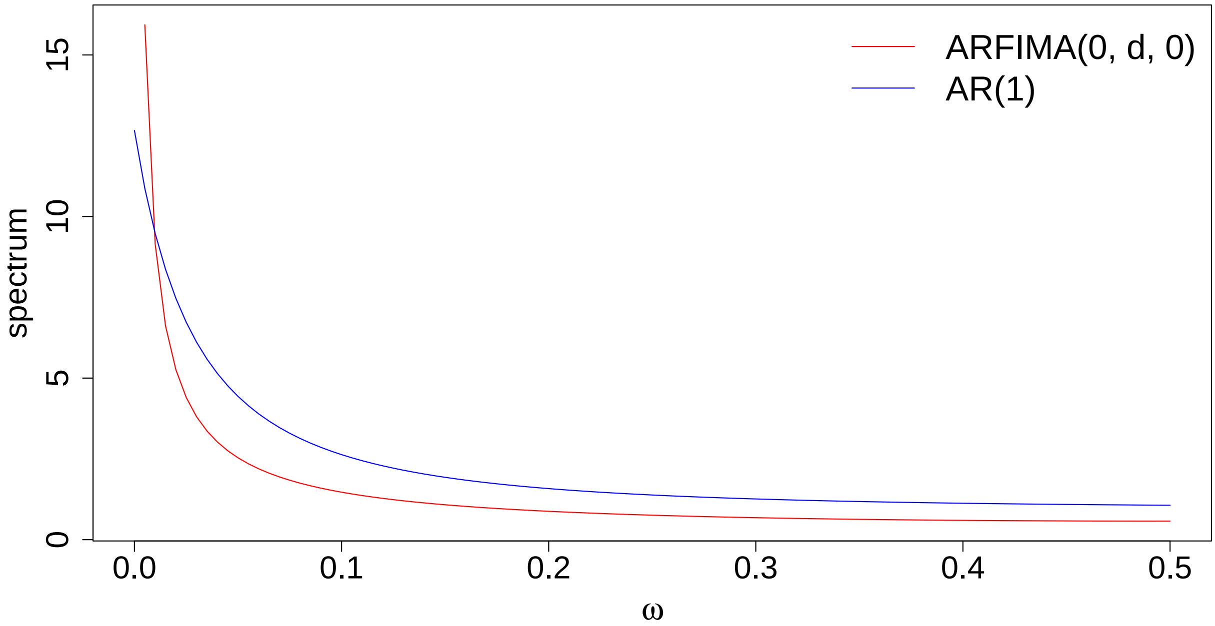


Рис. 2.1. Спектральная плотность процессов с одинаковой дисперсией

2.2.5. Сравнение MC-SSA при разных моделях шума

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

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

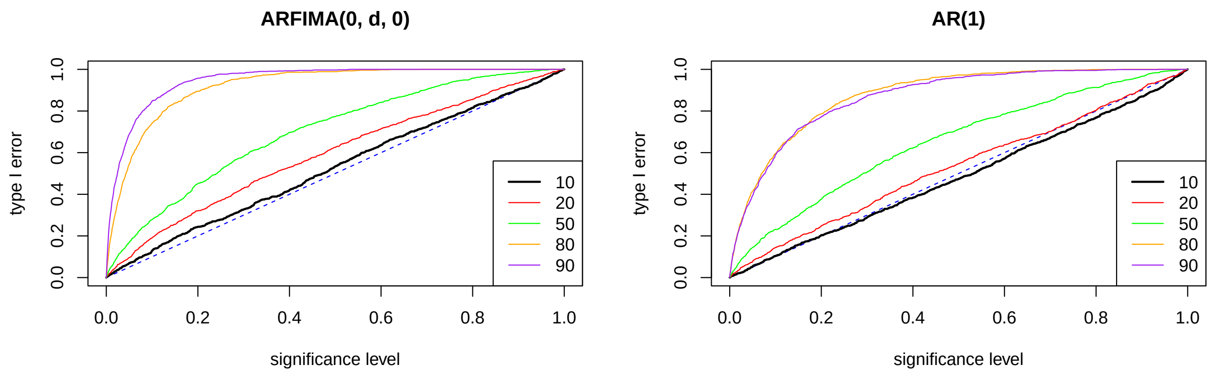
Тогда дисперсии процессов равны тогда и только тогда, когда

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

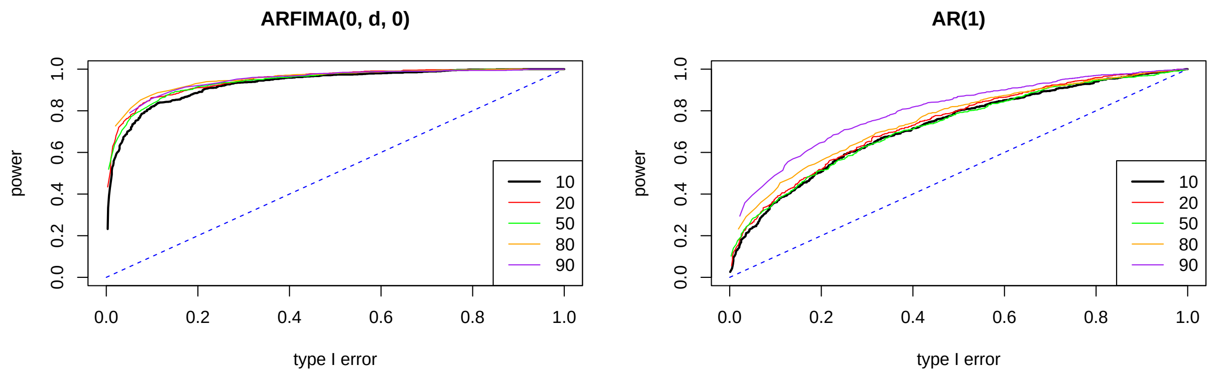
Пусть $d = 0.4$. Тогда при $\phi \approx 0.719$ процессы $\boldsymbol{\xi}$ и $\boldsymbol{\eta}$ имеют одинаковую дисперсию. На рис. 2.1 изображены спектральные плотности процессов. На нем видно, что процесс $\boldsymbol{\eta}$ имеет меньшее значение плотности для всех значений ω, за исключением близких к нулю.

Предполагается, что если рассмотреть сигнал с частотой $\omega : f_{\boldsymbol{\eta}}(\omega) < f_{\boldsymbol{\xi}}(\omega)$, то критерий MC-SSA с суррогатными реализациями процесса $\boldsymbol{\eta}$ даст мощность больше, чем тот же критерий с суррогатными реализациями процесса $\boldsymbol{\xi}$. Убедимся в этом. Пусть длина ряда $N = 100$ и

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



(*а*) Ошибка первого рода



(*б*) ROC-кривая

Рис. 2.2. Сравнение мощностей MC-SSA

На рис. 2.2, *а* изображены график ошибок первого рода критериев MC-SSA для разных длин окна L. По нему видно, что рассматриваемые критерии являются радикальными, поэтому сравнивать их по мощности будем с помощью ROC-кривых, которые являются графиком мощности критериев, к которым была применена поправка из раздела 2.1.2. По рис. 2.2, *б* видно, что, действительно, замена модели AR(1) на ARFIMA$(0, d, 0)$ увеличила мощность критерия против данной альтернативы.

Заключение

TODO

Список литературы

1. Palma Wilfredo. Long-Memory Time Series: Theory and Methods. — Wiley, 2006.
2. Time Series Analysis: Forecasting and Control / Box G., Jenkins G., Reinsel G., and Ljung G. — Fifth ed. — 2016.
3. Hassler Uwe. Time Series Analysis with Long Memory in View. — Wiley, 2018.
4. Hipel Keith W., McLeod Ian. Time series modelling of water resources and environmental systems. — Elsevier, 1994.
5. Haslett John, Raftery Adrian 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 Maria C., Bhuiyan Md Al Masum, Tweneboah Osei K. and Gonzalez-Huizar Hector // 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. Guglielmo Maria Caporale Luis Gil-Alana, Plastun Alex. Long memory and data frequency in financial markets // Journal of Statistical Computation and Simulation. — 2019. — Vol. 89, no. 10. — P. 1763–1779.
9. Whittle P. The Analysis of Multiple Stationary Time Series // Journal of the Royal Statistical Society. Series B (Methodological). — 1953. — P. 125–139.
10. R Core Team. — R: A Language and Environment for Statistical Computing. — R Foundation for Statistical Computing, Vienna, Austria, 2024. — Access mode: `https://www.R-project.org/`.
11. Veenstra J.Q. — arfima: Fractional ARIMA (and Other Long Memory) Time Series Modeling : 2012.
12. Maechler Martin. — fracdiff: Fractionally Differenced ARIMA aka ARFIMA(P,d,q) Models : 1999.
13. Ларин Е. С. Метод SSA для проверки гипотезы о существовании сигнала во временном ряде : квалификационная работа магистра ; СПбГУ. — 2022.
14. 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.