бПМИ ФКН ВШЭ, 3 курс, 3 модуль

Задание 3. Авторегрессионные и условно-гауссовские модели временных рядов

Вероятностные модели и статистика случайных процессов, весна 2018

Время выдачи задания: 5 марта (понедельник).

Срок сдачи: **19 марта (понедельник), 23:59.**

Среда для выполнения практического задания – PYTHON 2.x.

Правила сдачи

Выполнение работы в команде

1. Домашнее задание допускается выполнять в команде от 1 до 4 человек.

2. Командное решение достаточно загрузить в AnyTask только один раз. При этом в посылке следует указать состав команды.

3. Баллы, набранные командой, выставляются всем членам команды одинаковыми. Бонусные баллы выставляются всем членам команды одинаковыми. Это означает, что каждый член команды получает баллы, набранные его командой, независимо от его вклада в решение работы.

Инструкция по отправке:

1. Решения задач следует присылать единым файлом формата .pdf, набранным в LaTeX. Допускается отправка отдельных практических задач в виде отдельных файлов (ipython-тетрадок или исходных файлов с кодом на языке `python`).

Оценивание и штрафы:

1. Максимально допустимая оценка за работу – 10 баллов. Баллы, набранные сверх максимальной оценки, считаются бонусными и влияют на освобождение от задач на экзамене.

2. Дедлайн жесткий. Сдавать задание после указанного срока сдачи нельзя.

3. Задание выполняется командой независимо от других команд. «Похожие» решения считаются плагиатом и все студенты обеих команд (в том числе те, у кого списали) не могут получить за него больше 0 баллов (подробнее о плагиате см. на странице курса). Если вы нашли решение какого-то из заданий (или его часть) в открытом источнике, необходимо указать ссылку на этот источник в отдельном блоке в конце вашей работы (скорее всего вы будете не единственным, кто это нашел, поэтому чтобы исключить подозрение в плагиате, необходима ссылка на источник).

Необходимые теоретические сведения

1. Процесс скользящего среднего MA(q) – это процесс $X = (X_t)_{t \geqslant 0}$, задаваемый уравнением

$$X_t = b_1 \varepsilon_{t-1} + \ldots + b_q \varepsilon_{t-q} + \sigma^2 \varepsilon_t,$$

где параметры $b_i \in \mathbb{R}, i = 1, \ldots, q$, $\sigma^2 > 0$, а процесс $\varepsilon = (\varepsilon_t)_{t \geqslant 0}$ – последовательность независимых одинаково стандартно нормально распределенных случайных величин.

2. Авторегрессионный процесс AR(p) – это процесс $X = (X_t)_{t \geqslant 0}$, задаваемый уравнением

$$X_t = a_0 + a_1 X_{t-1} + \ldots + a_p X_{t-p} + \sigma^2 \varepsilon_t,$$

где параметры $a_i \in \mathbb{R}, i = 0, \ldots, p$, $\sigma^2 > 0$, а процесс $\varepsilon = (\varepsilon_t)_{t \geqslant 0}$ – последовательность независимых одинаково стандартно нормально распределенных случайных величин.

3. Пусть $X = (X_t)_{t \geqslant 0}$ – процесс авторегрессии. Уравнения Юла-Уолкера выражают коэффициенты автоковариации с заданным *лагом* k (т. е. величины $R(k) = \mathrm{E}[X_t X_{t+k}]$) процесса $X = (X_t)_{t \geqslant 0}$ через коэффициенты автоковариации с меньшими лагами:

$$R(k) = a_1 R(k-1) + \ldots + a_p R(k-p), \quad k \in \mathbb{N}.$$

Аналогичное соотношение справедливо и для коэффициентов *корреляции* $\rho_k = \rho(k) = R(k)/R(0)$, где $R(0)$ – дисперсия временного ряда X_t.

4. Модель ARMA(p, q) (autoregressive moving average) – это стохастическая модель временного ряда, в которой наблюдения $X = (X_t)_{t \geqslant 0}$ задаются соотношением

$$X_t = a_0 + \sum_{i=1}^{p} a_i X_{t-i} + \sum_{i=1}^{p} b_i \varepsilon_{t-i} + \sigma \varepsilon_t,$$

где имеются члены AR(p) и MA(q).

5. Модель ARMAX(p, q, r) (autoregressive moving average with exogenous inputs) – это стохастическая модель временного ряда, в которой наблюдения $X = (X_t)_{t \geqslant 0}$ задаются соотношением

$$X_t = a_0 + \sum_{i=1}^{p} a_i X_{t-i} + \sum_{i=1}^{p} b_i \varepsilon_{t-i} + \sigma \varepsilon_t + \sum_{i=1}^{r} c_i u_{t-i},$$

где имеются члены AR(p), MA(q), и $u = (u_t)_{t \geqslant 0}$ – это некоторая заданная (возможно, случайная) последовательность.

6. Модель ARCH(p) (autoregressive conditional heteroscedasticity) – это стохастическая модель временного ряда, в которой наблюдения $X = (X_t)_{t \geqslant 0}$ задаются соотношением

$$X_t = \sigma_t \varepsilon_t, \qquad \sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 \ldots + \alpha_p X_{t-p}^2,$$

где параметры $\alpha_i \in \mathbb{R}, i = 0, \ldots, p$, а процесс $\varepsilon = (\varepsilon_t)_{t \geqslant 0}$ – последовательность независимых одинаково стандартно нормально распределенных случайных величин.

7. Модель GARCH(p, q) (generalized autoregressive conditional heteroscedasticity) – это стохастическая модель временного ряда, в которой наблюдения $X = (X_t)_{t \geqslant 0}$ задаются соотношением

$$X_t = \sigma_t \varepsilon_t, \qquad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{i=1}^{q} \beta_i \sigma_{t-i}^2,$$

где параметры $\alpha_i, \beta_j \in \mathbb{R}, i = 0, \ldots, p, j = 1, \ldots, q$, а процесс $\varepsilon = (\varepsilon_t)_{t \geqslant 0}$ – последовательность независимых одинаково стандартно нормально распределенных случайных величин.

8. Оператор сдвига индекса временного ряда L – это оператор, изменяющий индекс временного ряда на меньший согласно соотношению

$$LX_t = X_{t-1}, \quad t \in \mathbb{Z}.$$

Вариант 1

1. (3 балла) Для случайного процесса $h = (h_t)_{t \geqslant 0}$, заданного выражением:

   (a) $h_t = \varepsilon_t - 1.3\varepsilon_{t-1} + 0.4\varepsilon_{t-2}$,

   (b) $h_t - h_{t-1} = \varepsilon_t - 1.3\varepsilon_{t-1} + 0.3\varepsilon_{t-2}$,

   где $\varepsilon = (\varepsilon_t)_{t \geqslant 0}$ – последовательность независимых одинаково стандартно нормально распределенных случайных величин,

   (a) Записать эквивалентные представления с использованием оператора сдвига L;
   (b) Провести исследование на стационарность 2-го порядка (с доказательством);
   (c) Вычислить первые четыре автокорреляции $\rho_0, \rho_1, \rho_2, \rho_3$;
   (d) Вычислить дисперсию случайной величины h_t;
   (e) Определить тип процесса в терминах ARMA(p, q);
   (f) Записать уравнения Юла-Уолкера;
   (g) Решить эти уравнения с определением коэффициентов автокорреляции ρ_1, ρ_2.

2. (3 балла) Рассматривается процесс ARMAX, заданный уравнением

$$X_t - 1.5X_{t-1} + 0.7X_{t-2} = u_{t-1} + 0.5u_{t-2} + \varepsilon_t - \varepsilon_{t-1} + 0.2\varepsilon_{t-2},$$

   где $\varepsilon = (\varepsilon_t)_{t \geqslant 0}$ – процесс белого шума с дисперсией 0.25, последовательность $u = (u_t)_{t \geqslant 0}$ – случайная бинарная (± 1) последовательность.

   (a) Сгенерировать траекторию длины $N = 500$ указанного процесса;

(b) Выписать функционал правдоподобия указанного процесса и выражения для оценок максимального правдоподобия его параметров в предположении, что известны порядки частей AR, MA и X процесса;

(c) По сгенерированной выборке оценить параметры процесса ARMAX в предположении, что полностью известна модель (известны порядки частей AR, MA и X процесса);

(d) Построить графики зависимости оценок параметров процесса ARMAX (предполагается, что полностью известна модель процесса) от объема использованной выборки. Сходятся ли эти оценки к настоящим значениям?

3. (3 балла) Доходность акций часто оценивают в процентах согласно следующей формуле. Если S_n – цена закрытия акции в день $n, n \geqslant 1$, то ее доходность равна величине $X_n = \frac{S_n - S_{n-1}}{S_n}$ (отношение прироста цены закрытия за текущий день к цене предыдущего дня, daily returns in percent).

Для моделирования временного ряда, соответствующего процентной доходности некоторого фондового индекса, использовалось 1000 наблюдений, к которым была подогнана GARCH-модель. Результатом подгонки являются уравнения:

$$X_n = 0.102 + h_n, \quad h_n = \sigma_n \varepsilon_n,$$
$$\sigma_n^2 = 0.146 + 0.1068 h_{n-1}^2 + 0.8212 \sigma_{n-1}^2.$$

Здесь X_n – доходность в день n, $\varepsilon = (\varepsilon_t)_{t=1,2,\dots}$ – процесс гауссовского белого шума.

Доходность в день $n = 1000$ (2013-05-31) составляла $X_{1000} = -1.354$, причем $\sigma_{1000}^2 = 2.27$. В день $n+1 = 1001$ (2013-06-03) наблюдаемая доходность составила $X_{1001} = -10.47$, в то время как прогноз дисперсии в этот день составил 2.24.

(a) Записать формулу для подсчета σ_{1001}^2, подставить числовое значение, подсчитать результат;

(b) Подсчитать распределение доходности X_{1001};

(c) Обосновать ожидаемость или невозможность наблюдений фактической доходности $X_{1000} = -1.354$;

(d) Используя приведенный пример, обосновать утверждение о том, что GARCH является моделью с условной дисперсией;

(e) Используя уравнения модели, охарактеризовать *безусловную* дисперсию доходности.

4. (3 балла) Вам выдан файл `aapl.txt`, содержащий значения наблюдаемой доходности $X_n = \frac{S_n - S_{n-1}}{S_n}$ акций компании Apple в период с 01.01.2007 по 01.01.2017.

(a) Записать формулу для правдоподобия условно-гауссовской модели ARCH(p);

(b) Используя несколько различных значений p, оценить параметры модели ARCH(p), прокомментировать качество оценки для различных p;

(c) Определить, характеризуются ли данные кластерами волатильности (volatility clustering), т.е. дает ли модель ARCH(p) преимущества по сравнению с однородной моделью, для которой $\sigma_n^2 = \mathrm{const}$;

(d) Для выбранного вами значения p нарисовать график волатильности σ_n^2;

(e) Описать процедуру выбора оптимального значения p с помощью какого-либо информационного критерия.

Вариант 2

1. (3 балла) Для случайного процесса $h = (h_t)_{t=1,2,\dots}$, заданного выражением:

   (a) $h_t - 0.5h_{t-1} = \varepsilon_t - 1.3\varepsilon_{t-1} + 0.4\varepsilon_{t-2}$,

   (b) $h_t - 1.5h_{t-1} + 0.6h_{t-2} = \varepsilon_t$,

   где $\varepsilon = (\varepsilon_t)_{t\geqslant 0}$ – последовательность независимых одинаково стандартно нормально распределенных случайных величин,

   (a) Записать эквивалентные представления с использованием оператора сдвига L;
   (b) Провести исследование на стационарность 2-го порядка (с доказательством);
   (c) Вычислить первые четыре автокорреляции $\rho_0, \rho_1, \rho_2, \rho_3$;
   (d) Вычислить дисперсию случайной величины h_t;
   (e) Определить тип процесса в терминах ARMA(p, q);
   (f) Записать уравнения Юла-Уолкера;
   (g) Решить эти уравнения с определением коэффициентов автокорреляции ρ_1, ρ_2.

2. (3 балла) Сгенерировать траекторию длины $N = 80$ авторегрессионного процесса

$$(1 + 1.5L^{-1} + 0.5625L^{-2})X_t = 0.1\varepsilon_t,$$

   где $\varepsilon = (\varepsilon_t)_{t\geqslant 0}$ – последовательность независимых одинаково стандартно нормально распределенных случайных величин.

   (a) Подогнать к сгенерированной траектории процессы скользящего среднего MA(q) порядков $q \in \{1, 2, 3, 4\}$. Какая из этих моделей, согласно критерию AIC, является наиболее подходящей для моделирования сгенерированного временного ряда?

(b) Подогнать к сгенерированной траектории процессы авторегрессии AR(p) порядков $p \in \{1, 2, 3, 4\}$. Использовать для оценки параметров процессов авторегрессии метод максимального правдоподобия. Сравнить значения полученных оценок коэффициентов процесса авторегрессии со значениями оценок заданного процесса авторегрессии, который использовался для генерации данных.

(c) Подогнать к сгенерированной траектории процесс авторегрессии порядка $p = 2$ (то есть процесс авторегрессии правильного порядка), используя уравнения Юла-Уолкера (т. е. решая их относительно коэффициентов $a_0, a_1, \ldots, a_p$. Исследовать зависимость ошибки оценки параметров процесса авторегрессии от объема выборки.

3. (3 балла) Доходность акций часто оценивают в процентах согласно следующей формуле. Если S_n – цена закрытия акции в день $n, n \geqslant 1$, то ее доходность равна величине $X_n = \frac{S_n - S_{n-1}}{S_n}$ (отношение прироста цены закрытия за текущий день к цене предыдущего дня, daily returns in percent).

Для моделирования временного ряда, соответствующего процентной доходности некоторого фондового индекса, использовалось 1000 наблюдений, к которым была подогнана GARCH-модель. Результатом подгонки являются уравнения:

$$X_n = 0.102 + h_n, \quad h_n = \sigma_n \varepsilon_n,$$
$$\sigma_n^2 = 0.146 + 0.1068 h_{n-1}^2 + 0.8212 \sigma_{n-1}^2.$$

Здесь X_n – доходность в день n, $\varepsilon = (\varepsilon_t)_{t=1,2,\ldots}$ – процесс гауссовского белого шума.

Доходность в день $n = 1000$ (2013-05-31) составляла $X_{1000} = -1.354$, причем $\sigma^2_{1000} = 2.27$. В день $n + 1 = 1001$ (2013-06-03) наблюдаемая доходность составила $X_{1001} = -10.47$, в то время как прогноз дисперсии в этот день составил 2.24.

(a) Записать формулу для подсчета σ^2_{1001}, подставить числовое значение, подсчитать результат;

(b) Подсчитать распределение доходности X_{1001};

(c) Обосновать ожидаемость или невозможность наблюдений фактической доходности $X_{1000} = -1.354$;

(d) Используя приведенный пример, обосновать утверждение о том, что GARCH является моделью с условной дисперсией;

(e) Используя уравнения модели, охарактеризовать *безусловную* дисперсию доходности.

4. (3 балла) Вам выдан файл `goog.txt`, содержащий значения наблюдаемой доходности $X_n = \frac{S_n - S_{n-1}}{S_n}$ акций компании Google в период с 01.01.2007 по 01.01.2017.

(a) Записать формулу для правдоподобия условно-гауссовской модели GARCH(p, q);

(b) Используя несколько различных значений p и q, оценить параметры модели GARCH(p, q), прокомментировать качество оценки для различных p и q;

(c) Определить, характеризуются ли данные кластерами волатильности (volatility clustering), т.е. дает ли модель GARCH(p, q) преимущества по сравнению с однородной моделью, для которой $\sigma_n^2 = \text{const}$;

(d) Для выбранных вами значений p, q нарисовать график волатильности σ_n^2;

(e) Описать процедуру выбора оптимального значения параметров p и q с помощью какого-либо информационного критерия.