ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ "МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ имени М.В.ЛОМОНОСОВА"
МЕХАНИКО-МАТЕМАТИЧЕСКИЙ ФАКУЛЬТЕТ
КАФЕДРА ОБЩИХ ПРОБЛЕМ УПРАВЛЕНИЯ

КУРСОВАЯ РАБОТА

"Исследование ценовых рядов на основании Математической Статистики и Теории Вероятностей"

Автор - Нарембеков Темирлан, студент 3-го курса кафедры Общих Проблем Управления, 312 группа

Научный руководитель - Заплетин Максим Петрович, доцент кафедры Общих Проблем Управления

Весна 2024

Содержание

1 СБОР И ПОДГОТОВКА ДАННЫХ

1.1 Сбор данных

Выберем временной период для анализа: май 2009г - апрель 2024г. со 180 ежемесячными показателями USD/KZT. Таким образом, интервал времени $\Delta = 1$ месяц является регулярным.

USD/KZT 446,435 +0,135 (+0,03%)

Дата	Цена	Откр.	Макс.	Мин.	Объём	Изм. %
01.04.2024	446,435	447,405	448,555	445,950		+0.12%
01.03.2024	445,920	451,105	455,310	445,680		-1.03%
01.02.2024	450,580	449,755	455,590	445,860		+0.30%
01.01.2024	449,230	456,810	458,905	444,290		-0.92%
01.12.2023	453,400	459,360	462,510	452,150		-0.81%
01.11.2023	457,100	468,860	471,200	455,910		-2.39%
01.10.2023	468,290	477,910	481,060	467,990		-1.91%
01.09.2023	477,390	458,360	486,210	454,595		+4.27%
01.08.2023	457,840	444,600	467,110	441,975		+3.10%
01.07.2023	444,080	449,100	450,305	439,175		-1.39%
01.06.2023	450,330	447,755	455,265	443,600		+0.94%
01.05.2023	446,130	452,155	453,560	440,740		-1.21%

01.03.2010	147,040	147,350	147,460	146,850	-0.19%
01.02.2010	147,325	148,025	148,235	147,115	-0.48%
01.01.2010	148,030	148,490	148,850	147,850	-0.33%
01.12.2009	148,520	148,750	150,000	148,280	-0.11%
01.11.2009	148,685	150,695	150,935	148,635	-1.38%
01.10.2009	150,770	151,010	151,050	150,060	-0.11%
01.09.2009	150,940	150,820	150,990	150,670	+0.07%
01.08.2009	150,830	150,690	150,930	150,650	+0.05%
01.07.2009	150,755	150,395	150,945	150,245	+0.22%
01.06.2009	150,430	150,490	150,590	149,340	-0.06%
01.05.2009	150,515	150,665	150,955	149,775	-0.14%

Максимум: 527,125 Минимум: 145,135 Разница: 381,990 Среднее: 294,247 Изм. %: 196,202

1.2 Подготовка Данных

Мы имеем наблюдения ежемесячной стоимости валюты за последние 15 лет, но сами по себе эти данные интересуют нас лишь косвенно по следующей причине.

В середине 20 века появились работы, в которых было доказано, что в поведении цен акций и товаров нет ни ритмов, ни трендов, ни циклов, а суммы логарифмов цен - являются случайным блужданием, которые описывают всю эволюцию цен.

Поэтому при исследовании ценовых рядов используются логарифмы цен $H_{tk} = \ln\left(\frac{S_{t_k}}{S_{t_{k-1}}}\right)$.

2 ИССЛЕДОВАНИЕ ДОХОДНОСТИ ТЕНГЕ НА СЛУЧАЙНОСТЬ

Проверим логарифмическую доходность тенге разными статистическими методами $H_{t_k} = \ln\left(\frac{S_{t_k}}{S_{t_{k-1}}}\right)$ с общим числом наблюдений равным 179.

2.1 Ранговый критерий Вальда-Вольфовица

Проверка статистических гипотез заключается в том, чтобы решить какой из исходов влечет наибольшие риски, а затем ставят задачу отклонить этот исход.

Данный критерий использует понятие Ранга R_k наблюдения $\ln\left(\frac{S_{t_k}}{S_{t_{k-1}}}\right)$, которое является номером наблюдения в соответсвующем вариационном ряду, для вычисления статистики $R^{\star} = \frac{R}{\sqrt{D[R]}}$ и улучшении ее до статистики $R^{\star\star} = R^{\star} + 1.1216 \cdot n^{-0.523}$

1)Построим по известному ряду логарифмических доходностей вариационный ряд.

2) Введем гипотезу H_0 и альтернативу H_1:

H_0 : Тренд отсутствует - т.е.случайность ряда,

H_1 : Тренд присутствует - неслучайность ряда.

3) В нашем случае ошибка I рода - это наиболее критическая.

Значит, ошибка I рода - наиболее критическая

Таким образом наша (статистическая) задача сделать ошибку I рода α как можно меньше, однако это увеличит ошибку II рода β(что не несет рисков кроме возможной потери прибыли).

4)Выберем уровень значимости α (Ошибка I рода). Для объема выборки $100 < n < 1000$ рекомендуется $\alpha = 0.01$.

Критерий двусторонний, гипотеза об отсутствии тренда принимается, если $R^{\star\star} \in [-2.57, 2.57]$, где 2.57 - критическое значение нормального распределения, соответствующее $\alpha = 0{,}01$.

5)Построим статистику R $= \sum_{i=1}^{n-1} \left(R_i - \frac{n+1}{2}\right)\left(R_{i+1} - \frac{n+1}{2}\right)$ для$n = 179$: R $= 68436$

6)Вычислим D[R] $= \dfrac{n^2 \cdot (n+1) \cdot (n-3) \cdot (5n+6)}{720}$.

7)$R^{\star} = \dfrac{R}{\sqrt{D[R]}} = 1.9201840279563078$

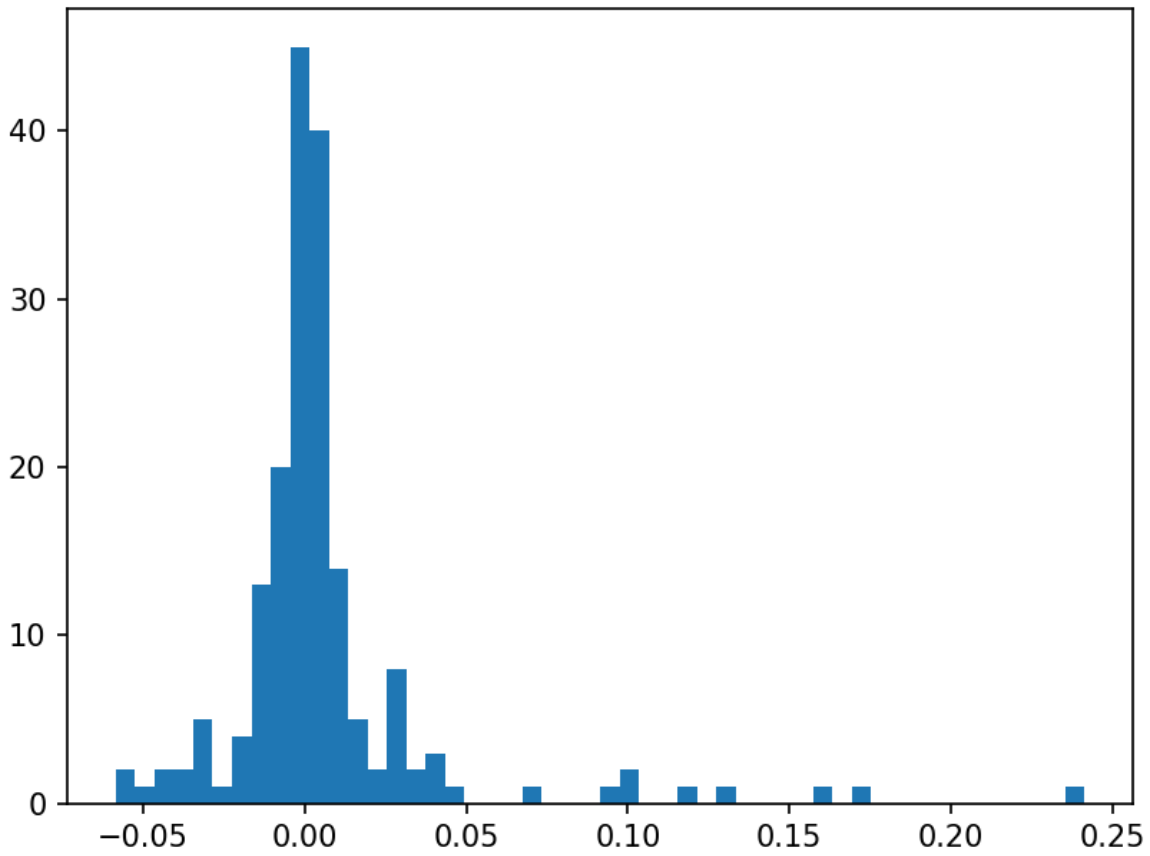
8) улучшим $R^{\star}$ до $R^{\star\star}$: $R^{\star\star} = 1.9945879562302098$

Таким образом принимаем гипотезу H_0 об отсутствии тренда и получаем, что ряд логарифмических доходностей - случаен.

3 ПРОВЕРКА НА НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

3.1 Гистограмма

Рассмотрим гистограмму: На практике существенным отклонением



от нормальности считается:

а)Наличие выбросов;

б)Отсутствие колоколообразной формы;

в)Отсутствие симметричности(при маленьких выборках к этому относятся снисходительно).

На гистограмме наблюдаются выбросы справа - это сигнализирует о существенном отклонении от нормальности. Если от них избавится, то на практике можно считать что распределение условно нормально, так как гистограмма имеет колоколообразную форму и небольшую ассиметрию. Однако нашем случае распределение нормальным не будет, далее мы убедимся в этом.

Тест Шапиро-Уилка

$$H_0 : \text{Распределение нормальное},$$
$$H_1 : \text{Распределение не нормальное}.$$

Статистика Критерия

$$W = \frac{1}{s^2}\left[\sum_{i=1}^{n} a_{n-i+1}(x_{n-i+1} - x_i)\right]^2,$$

$$\text{где} \quad s^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2, \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

с ростом n приближается нормальным распределением.

```
50  df = pd.DataFrame(logarithmic_returns_array, columns=['LogReturns'])
51  res = stats.shapiro(df)
52  print('p-value: ', res[1])
```

PROBLEMS 10 OUTPUT DEBUG CONSOLE TERMINAL PORTS

```
p-value:  4.006752775774608e-19
```

p-значение намного меньше, значит гипотеза отклоняется.

4 РАСПРЕДЕЛЕНИЕ ЛОГАРИФМИЧЕСКИХ ДОХОДНОСТЕЙ H_K

Было выяснено, что ряд логарифмических доходностей является случайным(с распределением отличным от нормального)

Это означает, что H $= (H_{t_k})$ - выборка случайных величин. Для дальнейшего анализа, будем проверять являются ли H_{t_k} независимыми одинаково распределенными случайными величинами(Н.О.Р.) Для этого сперва посчитаем **Автокорреляционную функцию.**

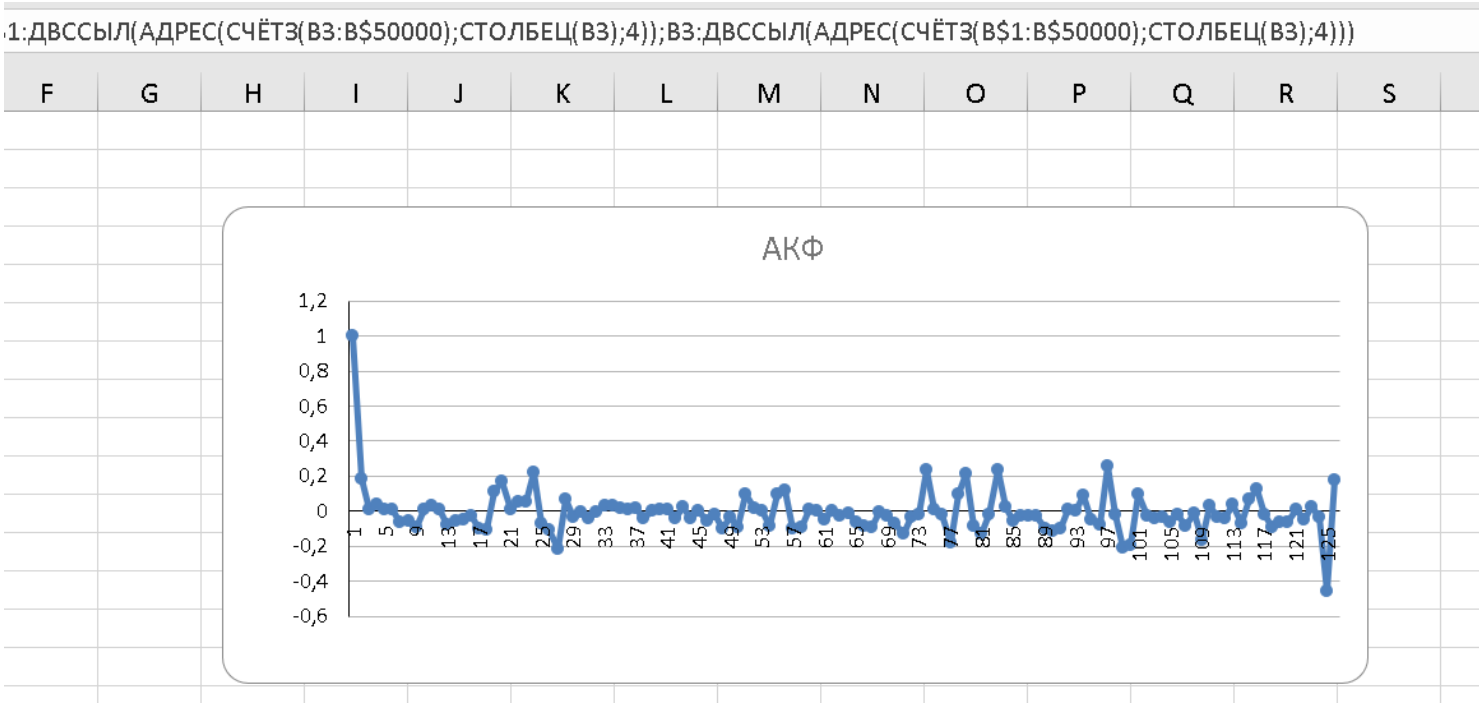
4.1 АВТОКОРРЕЛЯЦИОННАЯ ФУНКЦИЯ

$$R(K) = \frac{1}{(n-k)\sigma^2} \sum_{t=1}^{n-k} (H_t - \bar{H})(H_{t+k} - \bar{H})$$

- автоковариационная функция

$$\hat{\rho}(k) = \frac{\hat{R}(k)}{\hat{R}(0)}, \quad -n < k < n.$$

- автокорреляционная функция



4.2 ОДИНАКОВАЯ РАСПРЕДЕЛЕННОСТЬ

Рассмотрим график ряда на временной шкале:

Видно явное различие между первыми 80 и оставшимися величинами. Проверим эти 2 подвыборки на однородность.

Тест Колмогорова-Смирнова

Пусть эмпирическая функция распределения (ЭФР) F_n построенная по выборке $X = (X_1, \ldots, X_n)$, имеет вид: $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leqslant x}$, где $I_{X_i \leqslant x}$ указывает, попало ли наблюдение X_i в область $(-\infty, x]$:

$$I_{X_i \leqslant x} = \begin{cases} 1, & X_i \leqslant x; \\ 0, & X_i > x. \end{cases}$$

Таким образом имеем 2 эмпирические функции распределения F_{80} и F_{99}

H_0 : Распределения подвыборок совпадают и данные одинаково распределены.

H_1 : Распределения отличаются и данные не являются одинаково распределёнными.

Теорема Смирнова

Пусть $F_{1,n}(x), F_{2,m}(x)$ — эмпирические функции распределения, построенные по независимым выборкам объёмом n и m случайной величины ξ. Тогда, если $F(x) \in C^1(\mathbb{X})$, то

$$\forall t > 0 : \lim_{n,m\to\infty} P\left(\sqrt{\frac{nm}{n+m}} D_{n,m} \leqslant t\right) = K(t) = \sum_{j=-\infty}^{+\infty} (-1)^j e^{-2j^2t^2},$$

где $D_{n,m} = \sup_x |F_{1,n} - F_{2,m}|$.

Построим критическое множество $\mathrm{R} = \mathrm{D} \mid \mathrm{D} \geq K_{0.01} = 0.121$
Вычислим статистику: $D = \sqrt{\frac{80 \cdot 99}{80+99}} \cdot D_{80,99}$ Результат: $D = 0.36$
Таким образом выборки не одинаково распределены.

Хоть распределения и разные, может оказаться так, что они имеют одинаковый закон, но разные параметры, либо же вообще разные законы.
Стандартизируем, т.е. перейдем к

$$\frac{X - \bar{X}}{\sigma^{1/2}}$$

в каждой подвыборке, и выясним какой случай у нас, снова применив Тест:

```
80   standartized_subsample1 = [(x - stat.mean(sample_sub1))/stat.stdev(sample_sub1) for x in sample_sub1]
81   standartized_subsample2 = [(x - stat.mean(sample_sub2))/stat.stdev(sample_sub2) for x in sample_sub2]
82
83   ks_stat, ks_p_value = stats.ks_2samp(standartized_subsample1, standartized_subsample2)
84   print("Тест Колмогорова-Смирнова для двух выборок:")
85   print(f"Статистика K-S: {ks_stat}")
86   print(f"P-значение: {ks_p_value}")
87
PROBLEMS 1   OUTPUT   DEBUG CONSOLE   TERMINAL   PORTS                                   powershell +

Тест Колмогорова-Смирнова для двух выборок:
Статистика K-S: 0.36464646464646466
P-значение: 9.145329456492403e-06
```

Выходит, что законы распределений различны.

4.3 ОЧИСТКА ДАННЫХ

Исключим выбросы основанные на межквартильном расстоянии:

	LogReturns
56	0.170719
74	0.241325
75	0.127618
129	0.159723
152	0.116695

Далее, заполним пропуски средним арифметическим соседних значений либо линейной интерполяцией.

$$x_i = \frac{x_{i-1} + x_{i+1}}{2}$$

$$x_i = x_a + \frac{b-a}{i-a} \cdot (x_b - x_a)$$

Где:

- x_a — последнее допустимое значение перед выбросами,
- x_b — первое допустимое значение после выбросов,
- i — индекс выброса (между a и b).

Тест Дики-Фуллера

Теперь протестируем очищенный ряд на стационарность тестом Дики-Фуллера

H_0 : существует единичный корень, ряд нестационарный.

$$y_t = a \cdot y_{t-1} + \varepsilon_t$$

- авторегрессионное уравнение 1 порядка AR(1)

$$\triangle y_t = b \cdot y_{t-1} + \varepsilon_t$$

Статистика критерия $DF = \frac{b}{\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{b})^2}{n^2}}}$ имеет распределение Дики-Фуллера.

```
df_cleaned = df[(df['LogReturns'] >= mean - 3 * std_dev) & (df['LogReturns'] <=
# Выполняем тест Дики-Фуллера на стационарность
result = adfuller(df['LogReturns'])
```

```
ADF Statistic: -10.986895194268861
p-value: 7.2213861252741 87e-20
Critical Values: {'1%': -3.467631519151906,
Ряд стационарен (отклоняем нулевую гипотезу)
```

Таким образом ряд доходностей стационарен, не имеет пропусков и выбросов, и не распределен нормально.

4.4 СОВМЕСТНАЯ ПЛОТНОСТЬ РАСПРЕДЕЛЕНИЯ H_{t_k}

Было выяснено, что ряд доходностей H1,...,H179 распределен не одинаково - это означает, что ряд состоит из k<179 случайных величин(а именно их реализаций).
Значит, речь пойдет о совместном распределении.

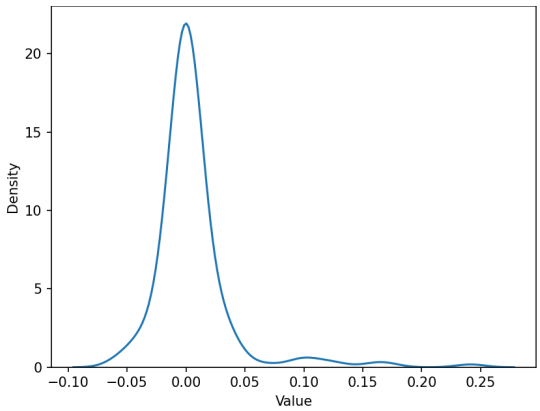


Рис. 1: KDE до удаления выбросов

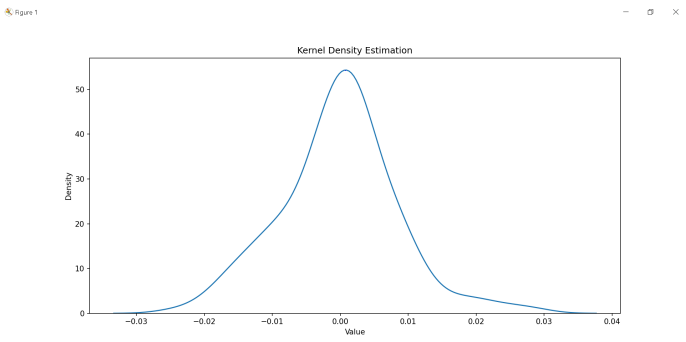


Рис. 2: KDE после удаления выбросов

Рис. 3: Ядерные оценки KDE совместной плотности H_{t_k} доипослепреобразованияданных.

Займемся приближением и нахождением аналитического вида совместной функции плотности ряда доходностей H_{t_k}.

GAUSSIAN MIXTURE MODEL(GMM)

Приблизим искомую функцию смесью нормальных распределений. Т.е. функция примет вид:

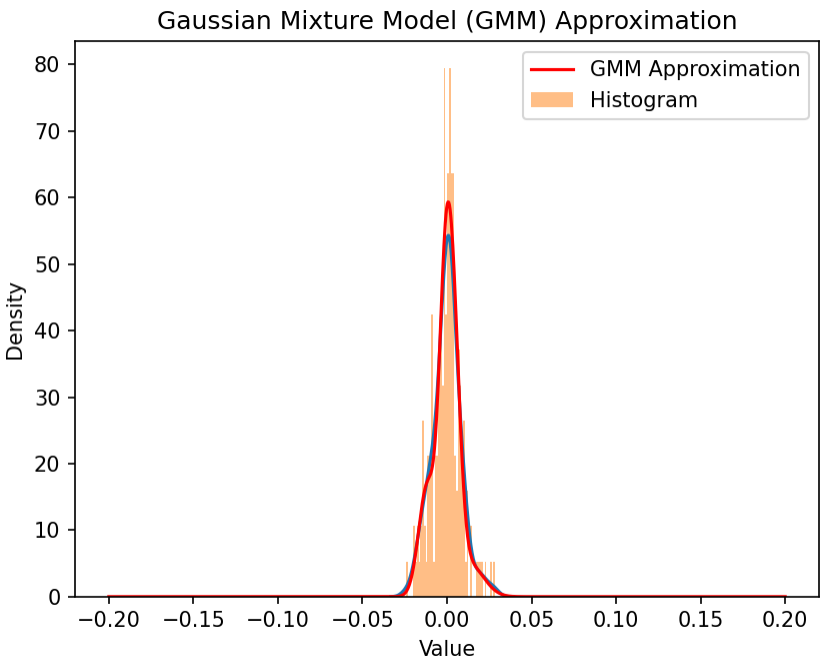
$$f(x) = \sum_{i=1}^{k} w_i \cdot N(x \mid \mu_i, \sigma_i^2),$$

где

$$N(x \mid \mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right)$$

— плотность нормального распределения с параметрами μ_i, σ_i^2, а w_i — вес i-й компоненты (все веса суммируются в 1). Остается найти число компонент k, при котором приближение наилучшее, и значения параметров.

```python
gmm = GaussianMixture(n_components=num_components, covariance_type='full', random_state=42)
logarithmic_returns_gmm = np.array(logarithmic_returns_list).reshape(-1, 1)
gmm.fit(logarithmic_returns_gmm)
# Получаем параметры смеси
means = gmm.means_.flatten()  # Средние (μ)
variances = gmm.covariances_.flatten()  # Дисперсии (σ^2)
weights = gmm.weights_.flatten()  # Веса
# Выведем найденные параметры
for i in range(num_components):
    print(f"Компонента {i+1}: μ = {means[i]:.4f}, σ^2 = {variances[i]:.4f}, w = {weights[i]:.4f}")
# Строим аппроксимацию плотности GMM
x_values = np.linspace(-0.2, 0.2, 1000)
gmm_pdf = np.exp(gmm.score_samples(x_values.reshape(-1, 1)))  # Оценка плотности
# Визуализируем
plt.plot(x_values, gmm_pdf, label="GMM Approximation", color="red")
plt.hist(logarithmic_returns_gmm, bins=50, density=True, alpha=0.5, label="Histogram")
plt.title("Gaussian Mixture Model (GMM) Approximation")
plt.legend()
plt.show()
```



```
Компонента 1: μ = 0.0136, σ^2 = 0.00006843, w = 0.0964
Компонента 2: μ = -0.0117, σ^2 = 0.00002495, w = 0.2057
Компонента 3: μ = -0.0016, σ^2 = 0.00001387, w = 0.2894
Компонента 4: μ = 0.0033, σ^2 = 0.00001849, w = 0.4084
```

ЛИТЕРАТУРА

1. "Основы Стохастической Финансовой Математики Том 1, А.Н. Ширяев.
2. "Критерии проверки гипотез о случайности и отсутствии тренда Б.Ю. Лемешко, И.В. Веретельникова.
3. Medium.com/Gaussian Mixture Modeling (GMM)