Аппроксимация временного ряда цен на золото

Романовский Илья, Серов Кирилл
Научный соруководитель: Коваленко В. В., аспирант кафедры ОПУ
Научный руководитель: к.ф. - м.н. Заплетин М.П., доцент кафедры ОПУ

Введение

Цель работы

Разработка и апробация модели на основе ряда Фурье для анализа цен на золото

Задачи исследования

- Предобработка и нормализация исторических данных
- Построение модели ряда Фурье с оптимизацией параметров
- Проверка точности аппроксимации и статистический анализ
- Визуализация результатов и сравнение с исходными данными

Ряды Фурье

Теорема Дирихле (ключевая теорема)

Для кусочно-непрерывных и монотонных функций ряд Фурье сходится:

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left(a_n \cos \frac{2\pi nt}{T} + b_n \sin \frac{2\pi nt}{T} \right)$$

где коэффициенты Фурье вычисляются по формулам:

$$a_n = \frac{2}{T} \int_0^T f(t) \cos \frac{2\pi nt}{T} dt$$

$$b_n = \frac{2}{T} \int_0^T f(t) \sin \frac{2\pi nt}{T} dt$$

Представление ряда Фурье только через косинусы

$$f(t) = A_0 + \sum_{n=1}^{N} A_n \cos\left(\frac{2\pi n t}{T} + \phi_n\right)$$

- $$A_0 = \frac{a_0}{2}$$
  $$A_n = \sqrt{a_n^2 + b_n^2} \quad \text{(амплитуда)}$$
- $$\phi_n = -\arctan\left(\frac{b_n}{a_n}\right) \quad \text{(фаза)}$$

Преимущества

- Меньше параметров
- Снижение переобучения

Метод ADAM: описание алгоритма

Инициализация параметров

- $\theta_0 = \left\{ \mathrm{A}_0^{(0)}, \mathrm{A}_1^{(0)}, \ldots, \mathrm{A}_\mathrm{N}^{(0)}, \phi_1^{(0)}, \ldots, \phi_\mathrm{N}^{(0)} \right\}$
- $\mathrm{m}_0 = \mathbb{E}[\mathrm{g}_0] = 0,$
- $\mathrm{v}_0 = \mathbb{E}[\mathrm{g}_0^2] = 0,$
- $\beta_1 = 0.9$, $\beta_2 = 0.999$
- Learning rate: $\alpha = 0.001$
- $\epsilon = 10^{-8}$
- $\mathcal{L}(\theta)$ - целевая функция, дифференцируемая по параметрам θ

Метод ADAM: описание алгоритма

Итеративное обновление

На каждой итерации t:

1. Вычисление градиента:

$$g_t = \nabla_\theta \mathcal{L}(\theta_{t-1}) \tag{1}$$

2. Обновление смещённой оценки первого момента (среднего градиентов):

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t \tag{2}$$

3. Обновление смещённой оценки второго момента (дисперсии градиентов):

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \tag{3}$$

Метод ADAM: описание алгоритма

Итеративное обновление

4. Коррекция смещения:

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1 - \beta_2^t} \tag{4}$$

5. Обновление параметров:

$$\theta_t = \theta_{t-1} - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} \tag{5}$$

Метод ADAM: описание алгоритма

Контроль сходимости

1. Алгоритм прекращает работу, когда изменение параметров становится меньше заданного порога ϵ_{params}:

$$\|\theta_{\text{t}} - \theta_{\text{t}-1}\|_2 < \epsilon_{\text{params}}$$

2. Дополнительный критерий основан на изменении значения функции потерь:

$$|\mathcal{L}_{\text{t}} - \mathcal{L}_{\text{t}-1}| < \epsilon_{\text{loss}}$$

Метод ADAM: описание алгоритма

Контроль сходимости

3. Для обеспечения устойчивости в алгоритме применяется регуляризация параметров:
   - Ограничение амплитуд: $A_n \in [-10, 10]$
   - Ограничение фаз: $\phi_n \in [-\pi, \pi]$
4. В качестве страховочного механизма установлено максимальное число итераций $T_{max} = 5000$, что гарантирует завершение алгоритма даже при медленной сходимости.

Метод градиентного спуска: описание алгоритма

Инициализация параметров

- Начальный вектор оптимизируемых параметров: $\theta_0 = \left\{ \mathrm{A}_0^{(0)}, \mathrm{A}_1^{(0)}, \ldots, \mathrm{A}_\mathrm{N}^{(0)}, \varphi_1^{(0)}, \ldots, \varphi_\mathrm{N}^{(0)} \right\}$
- Скорость обучения (learning rate): $\alpha = 0.001$
- Максимальное число итераций: $\mathrm{T}_{\mathrm{max}} = 5000$
- Порог остановки: ϵ — минимальное изменение функции потерь или параметров для прекращения итераций

Метод градиентного спуска: описание алгоритма

Итеративное обновление

На каждой итерации t:

1. Вычисление градиента:

$$\mathrm{g_t} = \nabla_\theta \mathcal{L}(\theta_{\mathrm{t}-1})$$

2. Обновление параметров:

$$\theta_{\mathrm{t}} = \theta_{\mathrm{t}-1} - \alpha \cdot \mathrm{g_t}$$

$$\mathrm{A}_{\mathrm{n}}^{(\mathrm{t})} = \mathrm{A}_{\mathrm{n}}^{(\mathrm{t}-1)} - \alpha \cdot \frac{\partial \mathcal{L}}{\partial \mathrm{A_n}}, \quad \varphi_{\mathrm{n}}^{(\mathrm{t})} = \varphi_{\mathrm{n}}^{(\mathrm{t}-1)} - \alpha \cdot \frac{\partial \mathcal{L}}{\partial \varphi_{\mathrm{n}}}$$

Метод градиентного спуска: описание алгоритма

Итеративное обновление

3. Проверка критериев остановки: Алгоритм прекращает работу, если выполнено хотя бы одно из условий:
   - Изменение параметров становится меньше заданного порога:

     $$\|\theta_{\mathrm{t}} - \theta_{\mathrm{t}-1}\|_2 < \epsilon_{\mathrm{params}}$$

   - Изменение функции потерь становится меньше порога:

     $$|\mathcal{L}(\theta_{\mathrm{t}}) - \mathcal{L}(\theta_{\mathrm{t}-1})| < \epsilon_{\mathrm{loss}}$$

   - Достигнуто максимальное число итераций $\mathrm{T}_{\mathrm{max}}$

Метод градиентного спуска: описание алгоритма

Регуляризация параметров

- Амплитуды: $\mathrm{A_n} \in [-15, 15]$
- Фазы: $\varphi_{\mathrm{n}} \in [-\pi, \pi]$

Преимущества метода

- Простота реализации
- Низкие вычислительные затраты на каждой итерации
- Быстрая сходимость при применении к простым функциям потерь

Метод винзоризации

Для данных $X = \{x_1, \ldots, x_n\}$ и параметра $\alpha \in [0, 0.5]$:

$$x_i^* = \begin{cases} Q_\alpha, & \text{если } x_i < Q_\alpha \\ x_i, & \text{если } Q_\alpha \le x_i \le Q_{1-\alpha} \\ Q_{1-\alpha}, & \text{если } x_i > Q_{1-\alpha} \end{cases}$$

Где:

- Q_α — нижний α-процентиль
- $Q_{1-\alpha}$ — верхний $(1-\alpha)$-процентиль

Применение

В работе использовано $\alpha = 0.1$ (10% границы)

Метод медианной замены

1. Вычисляем границы выбросов:

$$\begin{cases} Q_{15} = \inf\{x | F_n(x) \geq 0.15\} \\ Q_{85} = \inf\{x | F_n(x) \geq 0.85\} \end{cases}$$

2. Находим медиану выборки:

$$m = \mathrm{med}(\ell_i) = F_n^{-1}(0.5)$$

3. Заменяем аномальные значения:

$$\ell_i^* = \begin{cases} m, & \ell_i \notin [Q_{15}, Q_{85}] \\ \ell_i, & \text{иначе} \end{cases}$$

Статистический анализ остатков аппроксимации

Показатель Херста

Измеряет степень долгосрочной памяти временного ряда:

- $H \approx 0.5$ - белый шум (случайные колебания)
- $H > 0.5$ - персистентность (сохранение тенденций)
- $H < 0.5$ - антиперсистентность (частая смена направлений)

оценивается из следующего соотношения:

$$\frac{R(n)}{S(n)} \sim C \cdot n^{H}$$

Показатель Херста: алгоритм оценки

1. Разбиваем ряд на блоки размеров $n_1, n_2, \ldots, n_k$
2. Для каждого блока вычисляем:
   - Кумулятивную сумму $W_k = \sum_{i=1}^{k} X_i$
   - Размах $R = \max W_k - \min W_k$
   - Нормированный размах R/S , где S - стандартное отклонение
3. Усредняем по всем блокам:

$$\langle R/S \rangle_n = \frac{1}{m} \sum_{j=1}^{m} (R/S)_j$$

4. Строим регрессию:

$$\log(R/S)_n = \log C + H \cdot \log n + \epsilon$$

Показатель Херста: алгоритм оценки

5. Оцениваем показатель Херста (МНК)

$$\hat{H} = \frac{\sum_{i=1}^{k}(\log n_i - \overline{\log n})(\log\langle R/S\rangle_{n_i} - \overline{\log\langle R/S\rangle})}{\sum_{i=1}^{k}(\log n_i - \overline{\log n})^2}$$

где $\overline{\log n}$, $\overline{\log\langle R/S\rangle}$ – средние значения.

Коррекция для малых выборок

Для рядов с $N < 1000$ (в нашем случае $N = 558$) применяем поправку Аньюса-Ллойда:

$$\hat{H}_{корр} = \hat{H} - \frac{\log \mathbb{E}[(R/S)_n]}{\log n}$$

Коэффициент детерминации R^2

Определение

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

- $SS_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ (необъяснённая вариация)
- $SS_{tot} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ (общая вариация)

Интерпретация

- $R^2 = 1$: идеальная аппроксимация
- $R^2 \approx 0$: модель не лучше среднего
- Близость к 1: лучше объяснение данных

Критерий Шапиро-Уилка

Основные положения

- Нулевая гипотеза H_0: ошибки $\sim \mathcal{N}(\mu, \sigma^2)$
- Альтернатива H_1: распределение не нормальное
- Особенно эффективен для $n < 50$, но применим для любых n

Статистика теста

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

- $x_{(i)}$ - порядковые статистики
- a_i - табличные коэффициенты

Критерий Шапиро-Уилка

Если $m = (m_1, m_2, \ldots, m_n)^T$ — вектор математических ожиданий порядковых статистик стандартного нормального распределения, а V — их ковариационная матрица, то:

$$a = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}}$$

Правила принятия решений

- $W \approx 1$ - нет оснований отвергать H_0
- $W \ll 1$ - отклоняем H_0 (распределение не нормальное)
- p-value < 0.05 - статистически значимые отклонения

Критерий Колмогорова-Смирнова

Основные положения

- Нулевая гипотеза H_0: ошибки $\sim \mathcal{N}(\mu, \sigma^2)$
- Альтернатива H_1: распределение отлично от нормального
- Непараметрический критерий (не зависит от параметров распределения)

Статистика критерия

$$D_n = \sup_x |F_n(x) - F(x)|$$

- $F_n(x)$ - эмпирическая функция распределения
- $F(x)$ - теоретическая нормальная $\mathcal{N}(\bar{x}, s_x^2)$

Критерий Колмогорова-Смирнова

Интерпретация результатов

- $D_n \to 0$ - нет оснований отвергать H_0
- $D_n > D_{crit}$ - отклоняем H_0
- p-value < 0.05 - значимые отклонения от нормальности

Особенности критерия

- Чувствителен к различиям в центре и хвостах распределения
- Требует большого объема выборки для хорошей мощности

Итоги исследования

Что сделано

- Разработана модель на основе ряда Фурье с фазовым сдвигом
- Реализована оптимизация параметров (градиентный спуск и Adam)
- Рассмотрены методы борьбы с выбросами
- Проведена комплексная статистическая проверка модели
- Выполнена программная реализация на Python