Міністерство освіти і науки України
Дніпровський національний університет імені Олеся Гончара
Факультет прикладної математики
Кафедра математичного забезпечення ЕОМ

О. П. Луценко

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАНИХ

конспект лекцій
для здобувачів вищої освіти
другого (магістерського) рівня вищої освіти
зі спеціальності 121 – Інформаційні технології

Дніпро
ДНУ
2018

Розповсюдження і тиражування без офіційного дозволу Дніпровського національного університету імені Олеся Гончара забороняється

Конспект лекцій з дисципліни «Статистичний аналіз наддинамічних процесів» для студентів другого (магістерського) рівня спеціальності 121–Інженерія програмного забезпечення. Укладач: Луценко О.П. – Дніпро, РВВ ДНУ: 2018. – 49 с.

Укладач: канд.. техн. наук Луценко О. П.
Доцент кафедри математичного забезпечення ЕОМ Дніпровського національного університету імені Олеся Гончара

Рецензенти: д-р фіз.-мат. наук. Громов В.А.
Професор кафедри обчислювальної математики та математичної кібернетики Дніпровського національного університету імені Олеся Гончара

д-р техн.. наук, проф. Мороз Б. І.
декан технічного факультету
Університету митної справи та фінансів

Затверджено на засіданні кафедри МЗЕОМ 26.04.2018 р. (протокол № 9)
Протокол № 9 від 26.04.2018 р.

Затверджено Радою факультету прикладної математики Дніпровського національного університету імені Олеся Гончара
Протокол № 10 від 21.05.2018 р.

У конспекті лекцій викладено теоретичні відомості щодо методів аналізу часових рядів, утворених наддинамічними процесами моніторингу. Описані методи В-сплайн апроксимації, виявлення розладнань, відтворення щільності розподілів. Наведено приклади обчислювальних схем функцій ймовірності виникнення подій розладнання статистичних характеристик процесу.

ЗМІСТ

1 Згладжування часового ряду з використанням B-сплайнів

До того, як перейти до проблеми аналізу часового ряду, необхідно врахувати, що більшість процесів характеризуються великою кількістю дрібних стрибків курсу на тлі загальної тенденції, які утворюють шумову складову ряду і ускладнюють обробку даних. З ціллю виключити шумову складову перед застосуванням більшої частини існуючих методів аналізу даних потрібна додаткова первинна обробка часового ряду – згладжування.

Існує велика кількість способів згладити числовий ряд, як то: лінійна фільтрація, експоненціальне згладжування, згладжування методом найменших квадратів, розкладання в поліном Фур'є, апроксимація поліномом Лагранжа; метод «гусениці», що дозволяє провести спектральний аналіз і виділити перешкоди, а також безліч інших способів.

B-сплайни мають такі переваги [1]:

1) даний метод не дає лага за часом, на відміну, наприклад, від експоненціального згладжування і ковзних середніх;

2) крива на виході має прийнятну точність апроксимації;

3) метод дозволяє розрахувати значення проміжних точок у випадку, якщо для розрахунку використовуються не всі точки вибірки;

4) існують відносно швидкі способи обчислення точок B-сплайну, що критично у випадку великих обсягів вибірки;

5) ступінь згладжування ніяк не залежить від кількості точок у вибірці, а у разі зміни або поповнення вибірки достатньо перерахувати лише відрізок кривої, а не всю криву.

У загальному випадку математичний сплайн – це кусковий поліном ступеня K з неперервною похідної ступеня $K-1$ в точках з'єднання сегментів. Так, наприклад, кубічний сплайн має в точках з'єднання безперервність другого порядку. Кускові сплайни з багаточленів невисокого порядку дуже зручні для інтерполяції кривих, так як вони не вимагають великих обчислювальних витрат і не викликають чисельних відхилень, властивих многочленам високого порядку. За аналогією з фізичними сплайнами, зазвичай використовується серія кубічних сегментів, причому кожен сегмент проходить через дві точки.

B-сплайн – сплайн-функція, що має мінімальний носій для

заданого степеня, гладкості та області визначення.

Базис B-сплайна – неглобальний базис, що включає як частковий випадок поліном Бернштейна. B-сплайни неглобальні, так як з кожною вершиною пов'язана своя базисна функція. Тому вплив кожної вершини на криву проявляється тільки при тих значеннях параметра, де відповідна базисна функція не дорівнює нулю. Базис B-сплайна також дозволяє змінювати порядок базисних функцій і, отже, всієї кривої без зміни кількості вершин.

Крива, побудована на основі B-сплайн-базису, описується наступним чином [1]:

$$z(t_i) = \sum_{k=0}^{n} \overline{P}_k B_{kd}(t_i),$$

де $\overline{P}_k$ – вектор контрольних точок, побудований з вершин $y(t_i)$. Може включати як всі точки $y(t_i)$, так і кратні заданому кроку апроксимації;

n – кількість контрольних точок;

d – порядок кривої, $2 \le \mathrm{d} \le n+1$;

$B_{k,d}$ – вагова функция і-ої нормалізованої B-сплайн-базисної кривої порядку d (степені $d-1$), що задається співвідношенням (алгоритм Кокса – де Бура):

$$B_{k,1}(t_i) = \begin{cases} 1 & якщо\ u_k \le t_i \le u_{k-1} \\ 0 & якщо\ t_i \notin (u_k, u_{k-1}), \end{cases}$$
$$B_{k,d} = \frac{(u-u_k)B_{k,d-1}(u)}{u_{k+d-1}-u_k} + \frac{(u_{k+d}-u)B_{k+1,d-1}(u)}{u_{k+d}-u_{k+1}}, \qquad (1)$$

де u_k – елементи вузлового вектора.

Також $B_{k,d}$ називається стикувальною функцією.

Вузловий вектор є послідовністю дійсних цілих чисел u_i, таких що $u_i <= u_{i+1}$ для всіх u_i.

Єдина загальна вимога до вузлового вектору: $u_i <= u_{i+1}$, тобто це монотонно зростаюча послідовність дійсних чисел. Використовуються три типи вузлових векторів: рівномірні, відкриті рівномірні (або відкриті) і нерівномірні.

Зокрема, рівномірні вузлові вектори зазвичай починаються в нулі і збільшуються на 1 до деякого максимального значення або нормуються в діапазоні між 0 і 1 рівними десятковими значеннями, наприклад:

$$U=[0\ 0{,}25\ 0{,}5\ 0{,}75\ 1],$$
$$U=[0\ 1\ 2\ 3\ 4],$$
$$U=[-0{,}2\ -0{,}1\ 0\ 0{,}1\ 0{,}2].$$

Нерівномірні вузлові вектори відрізняються тим, що їх внутрішні вузлові величини розташовуються на різній відстані один від одного та/або поєднуються.

Вектори можуть бути відкритими ($U = [0\ 0\ 0\ 1\ 1\ 2\ 2\ 2]$) або періодичними ($U = [0\ 1\ 2\ 2\ 3\ 4]$).

У випадку вирішення задачі згладжування часового ряду маємо незамкнену криву, отже використовуємо неперіодичний вузловий вектор, що містить $n+2d-2$ елементів, де n – кількість точок контрольного полінома, d– порядок сплайну. Для кривої четвертого порядку відкритий рівномірний вузловий вектор має вигляд:

$$U=[0,0,0,0,1\ \ldots\ n-1,\ n,\ n,\ n,\ n].$$

Формула Кокса-де Бура (2.1) для розрахунку базисних функцій B-сплайна рекурсивна, тому функція порядку k залежить від базисних функцій нижчого порядку до 1. Нехай дана базисна функція $B_{k,d}$. Тоді цю залежність можна виразити у вигляді трикутника:

$B_{k,d}$,

$B_{k,d-1}$, $B_{k+1,d-1}$,

$B_{k,d-2}$, $B_{k+1,d-2}$, $B_{k+2,d-2}$,

...

$B_{k,1}$, $B_{k+1,1}$, $B_{k+2,1}$, $B_{k+3,1}$, ... $B_{k+d-1,1}$

Звідси, для одновимірного випадку, яким є задача згладжування часового ряду котирувань, значення функції $B_{k,d}$ можуть бути обчислені заздалегідь, щоб збільшити швидкість роботи алгоритму.

Для цього застосуємо наступну обчислювальну схему:

$$z(t) = P_0 B_{0,4}(t) + P_1 B_{1,4}(t) + P_2 B_{2,4}(t) + P_3 B_{3,4}(t) =$$

$$= \frac{1}{6}\begin{bmatrix} P_0\left(1 - 3t + 3t^2 - t^3\right) \\ + P_1\left(4 - 6t^2 + 3t^3\right) \\ + P_2\left(1 + 3t + 3t^2 - 3t^3\right) \\ + P_3\left(t^3\right) \end{bmatrix}.$$

Або в матричній формі:

$$z(t) = \frac{1}{6}\begin{bmatrix} P_0 & P_1 & P_2 & P_3 \end{bmatrix} 4 \begin{bmatrix} -1 & 3 & -3 & 1 \\ 3 & -6 & 3 & 0 \\ -3 & 0 & 3 & 0 \\ 1 & 4 & 1 & 0 \end{bmatrix} \begin{bmatrix} t^3 \\ t^2 \\ t \\ 1 \end{bmatrix}.$$

Значення стикувальної функції для решти інтервалів є здвигом отриманої функції вправо (рис. 1).

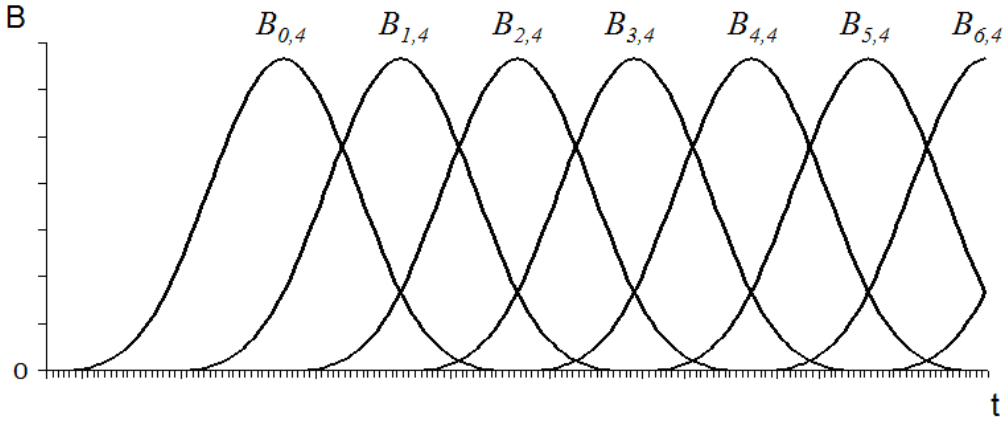

Рисунок 1 – Стикувальні функції на різних інтервалах

Керування гладкістю кривої досягається в наступні способи:

1) За рахунок введення похідних більш високого порядку. Порядок кривої визначає, наскільки вона гладка. По мірі того, як порядок зменшується, крива, що генерується, наближається до задаючого багатокутника. При $d = 2$ отримана крива є послідовністю прямих ліній, які ідентичні задаючому поліному. Крива четвертого порядку, яку будемо використовувати для рішення задачі, є кусково кубічним сплайном, неперервна за першою і другою похідними, а також по положенню точок вздовж всієї кривої.

2) За рахунок введення кратних вершин або за рахунок зміни числа і (або) положення кратних вершин в заданому багатокутнику.

В порівнянні з попереднім способом, цей спосіб не змінює порядок кривої, не впливає на складність розрахунків і не потребує реалізації окремих алгоритмів, тому йому була віддана перевага.

2 Сингулярно-спектральний аналіз та його використання в аналізі часових рядів

SSA (сингулярно-спектральний аналіз) являє собою новий метод аналізу часових рядів, що включає елементи класичного аналізу часових рядів, багатовимірного статистичного аналізу, багатовимірної геометрії, динамічних систем і обробки сигналів. Незважаючи на те, що багато імовірнісних і статистичних елементів, покладених в основу SSA-методів (вони пов'язані з техніками стаціонарності, ергодичності, головних компонент і початкового завантаження), SSA не статистичний метод в розумінні класичної статистики. Зокрема ми зазвичай не робимо жодних припущень стосовно ні сигналу, ні шуму при проведенні аналізу і дослідження властивостей алгоритмів [2].

Базова версія SSA складається з чотирьох кроків, які виконуються наступним чином. Нехай $F = (f_0, f_1,..., f_{n-1})$ часовий ряд довжиною N, та L — ціле число, яке будемо називати “довжиною вікна”.

Встановимо $K = N-L+1$ та визначимо L-зміщені вектори $X_j = (f_{j-1}..., f_{j+L-2})^T$, матрицю траєкторій

$$X = (f_{i+j-2})_{i,j=1}^{L,K} = [X_1 : ... : X_K], \quad j = 1,2..., K$$

Зауважимо, що матриця траєкторій X є матрицею Ханкеля, а це означає, що всі елементи по діагоналі $i+j=const$ рівні. Побудова матриці траєкторій являє собою перший крок алгоритму.

Другим кроком є сингулярне розкладання матриці X, яка може бути отримана за допомогою власних значень і векторів матриці $S = XX^T$ розміром $L \times L$. Це дає нам набір L сингулярних значень, які є квадратними коренями з власних значень матриці S, і відповідними лівим і правим сингулярними векторами. (Ліві сингулярні вектори X є ортонормованими векторами S, в літературі їх часто називають емпіричними ортогональними функціями або просто ЕОФ. Праві сингулярні вектори можна розглядати як власні вектори матриці X^TX). Таким чином, ми отримуємо представлення

X у вигляді суми біортогональних матриць першого порядку X_i *(i = 1, ..., d),* де *D (D ≤ L)* число ненульових сингулярних значень *X*.

На третьому кроці розіб'ємо множину індексів *I = {1,. . . , d}* на кілька груп $I_1. . . , I_m$ і підсумуємо матриці X_i в кожній групі. В результаті отримаємо представлення:

$$X = \sum_{k=1}^{m} X_{I_k}, X_{I_k} = \sum_{i \in I_k} X_i$$

На четвертому етапі, виконується усереднення по діагоналі $I + J = const$ X_{Ik} . Це дає нам SSA розкладання, тобто розкладання вихідної серії F на суму рядів

$$f_n = \sum_{k=1}^{m} f_n^{(k)}, \quad n = 0,..., N-1 \tag{2}$$

Де для кожного k послідовність $f_n^{(k)}$ – результат діагонального усереднення матриці X_{I_k} .

Принципова схема SSA для аналізу часових рядів і деякі модифікації цієї схеми описані в присвяченій SSA літературі, що наведена вище. Зверніть увагу, що SSA, як правило, розглядається як метод виявлення та вилучення коливальних компонент з оригінального ряду; див., наприклад, Yiou співавт. (1996), Ghil і Taricco (1997), Fowler і Kember (1998). Однак, стандартна література, присвячена SSA, не приділяє достатньої уваги теоретичним аспектам, які дуже важливі для розуміння того, як обирати параметри SSA і, в першу чергу, довжину вікна *L* для різних класів часових рядів. Поняття розділимості та пов'язані з ним методологічні аспекти та теоретичні результати дають нам змогу це зрозуміти. Вивчення розділимості складає велику різницю між нашим дослідженням SSA аналізу і стандартним підходом.

Вибір параметрів в процесі SSA-розкладання (довжини вікна *L* і способу групування матриці X_i) повинен залежати від властивостей ряду і мети аналізу.

Загальна мета SSA-аналізу — розкладання (2) на складові компоненти $f_n^{(k)}$, що є часовими рядами — «незалежними» і «такими, що ідентифікуються». Говорячи про аналіз структури часових рядів за допомогою SSA, ми маємо на увазі саме це. Іноді інтерес представляють конкретні завдання, такі як отримання сигналу з шуму, отримання коливальних компонент і згладжування.

За умови правильного SSA розкладання компоненту $f_n^{(k)}$ в (2.1) можна вважати трендом оригінального ряду, коливання ряду (наприклад, сезонного) або шуму. Коливальні ряди — це періодичні або квазіперіодичні ряди, що можуть бути чистими або амплітудно модульованими. Шум — будь-який аперіодичний ряд. Тренд ряду, грубо кажучи, є повільно змінюваною компонентою ряду з видаленням усіх коливань.

Базовий SSA виконується в чотири кроки. На першому етапі (що називається кроком згортання), одномірний ряд представляється у вигляді багатовимірного ряду, розмірність якого називається довжиною вікна. Багатовимірні часові ряди (які є послідовністю векторів) утворюють матрицю траєкторій. Єдиним (і дуже важливим) параметр цього кроку є довжина вікна.

Другий крок – сингулярне розкладання матриці траєкторій на суму біортогональних матриць першого порядку. Перші два етапи разом розглядаються як етап розкладання базового SSA.

Наступні два кроки – етап відновлення. На кроці групування відбувається поділ матриці, розрахованої на кроці сингулярного розкладання, на кілька груп і сумовування матриць в кожній групі. Результат кроку – представлення матриці траєкторій у вигляді суми кількох результуючих матриць.

На останньому кроці кожна результуюча матриця перетворюється на часовий ряд, що є аддитивною компонентою оригінального ряду. Ця операція називається діагональним усередненням. Це лінійна операція, що відображає матрицю траєкторій оригінального ряду на сам оригінальний ряд. Таким чином, ми отримаємо розкладання вихідного ряду на кілька адитивних компонент.

Опишемо ці кроки формально і обговоримо їх значення та функції.

2.1 Базовий SSA: описання

Нехай N > 2. Розглянемо часовий ряд дійсних значень $F = (f_0,...,f_{N-1})$ довжини N. Припустимо, що F – ненульовий ряд, тобто існує принаймні одне значення i, таке, що $f_i = 0$. Хоча звичайно можна припустити, що $f_i = f(i\Delta)$ для деякої функції від

часу *f(t)* і певного часового інтервалу *Δ*, це не грає особливої ролі в наших міркуваннях.

Крім того, числа *0. . . , N-1* можна інтерпретувати не тільки як дискретні моменти часу, але і в якості міток будь-яких інших лінійно впорядкованих структур. Нумерація значень часових рядів починається з $i = 0$, на відміну від більш стандартної $i = 1$; це тільки для зручності запису.

Як уже згадувалося, базовий SSA складається з двох доповнюючих один одного етапів: розкладання (декомпозиції) та відновлення (реконструкції).

2.1.1 Етап розкладання

1-й крок: згортання.

Процедура згортання представляє оригінальний ряд у вигляді послідовності багатовимірних векторів з відставанням.

Нехай L – ціле число (довжина вікна), $1 < L < N$. Процедура згортання утворює $K = N - L + 1$ векторів з відставанням.

$$X_i = (f_{i-1},..., f_{i+L-2})^T, \quad 1 < i < K,$$

що мають розмірність L. Якщо потрібно підкреслити розмірність векторів, будемо називати їх *векторами з L-відставанням*.

Строки матриці *L*-траєкторій (або просто матриці траєкторій) ряду *F*:

$$X = [X_1 : ... : X_K]$$

являють собою вектори з затримкою. Іншими словами, матриця траєкторій:

$$X = (x_{ij})_{i,j=1}^{L,K} = \begin{pmatrix} f_0 & f_1 & f_2 & \cdots & f_{K-1} \\ f_1 & f_2 & f_3 & \cdots & f_K \\ f_2 & f_3 & f_4 & \cdots & f_{K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f_{L-1} & f_L & f_{L+1} & \cdots & f_{N-1} \end{pmatrix} \qquad (3)$$

Очевидно, що $x_{ij} = f_{i+j-2}$, і матриця X має однакові елементи на діагоналях $i + j = const$. (Таким чином, матриця траєкторій є матрицею Ханкеля.) Звичайно, якщо N і L фіксовані, між матрицею траєкторій і часовим рядом існує відповідність один-до-одного.

2-й крок: сингулярне розкладання.

Результат кроку – сингулярне розкладання матриці траєкторій. Нехай $S=XX^T$. Позначимо як $\lambda_1,...,\lambda_L$ власні значення S, розташовані в порядку зменшення $(\lambda_1 \geq ... \geq \lambda_L \geq 0)$ і як $U_1,...,U_L$ ортонормальну систему власних векторів матриці S, що відповідають цим власним значенням. Нехай $d = \max\{i, таких, що\ \lambda_i > 0\}$.

Якщо позначимо $V_i = X^T U_i / \sqrt{\lambda_i}$ $(i = 1,...,d)$, то сингулярне розкладання матриці траєкторій X може бути записане як:

$$X = X_1 + ... + X_d \qquad (4)$$

де $X_i = \sqrt{\lambda_i} U_i V_i^T$. Матриці X_i мають порядок 1; отже, це елементарні матриці. Набір $(\sqrt{\lambda_i}, U_i, V_i^T)$ буде називатися i-ю власною трійкою сингулярного розкладання.

Після отримання розкладання (2) процедура групування розділяє множину індексів {1,…,d} на m підмножин $I_1 ... I_m$, що не перетинаються.

Нехай $I=\{i_1,...,i_p\}$. Тоді результуюча матриця X_I , що відповідає групі I , визначається як $X_I = X_{i1} + ... + X_{ip}$. Ці матриці обчислюються для $I=I_1,...,I_m$, і розкладання (1.2) приводить до розкладання:

$$X_I = X_{I_1} + ... + X_{I_m}$$

Процедура вибору множин $I_1,...,I_m$ називається групуванням по власним трійкам.

4-й крок: діагональне усереднення

На останньому кроці базового SSA кожна матриця згрупованого розкладання (1.3) перетворюється у новий ряд довжиною N.

Нехай Y – матриця $L \times K$ з елементами $y_{ij}, 1 \leq i \leq L, 1 \leq j \leq K$. Задамо $L^* = \min(L, K), K^* = \max(L, K)$ і $N = L + K - 1$. Нехай $y^*_{ij} = y_{ij}$, якщо $L < K$, і $y^*_{ij} = y_{ji}$ в іншому випадку.

Діагональне усереднення перетворює матрицю Y в ряд за формулою:

$$g_k = \begin{cases} \frac{1}{k+1}\sum_{m=1}^{k+1} y^*_{m,k-m+2} & для\ 0 \le k < L^* - 1 \\ \frac{1}{L}\sum_{m=1}^{L^*} y^*_{m,k-m+2} & для\ L^* - 1 \le k < K^* \\ \frac{1}{N-k}\sum_{m=k-K^*+2}^{N-K^*+1} y^*_{m,k-m+2} & для\ K^* - 1 \le k < N \end{cases} \quad (5)$$

Вираз (5) відповідає усередненню елементів матриці по діагоналям $i + j = k + 2$: вибір k = 0 дає $g_0 = y_{11}$, для k = 1 отримаємо $g_1 = (y_{12} + y_{21})/2$, і так далі. Відзначимо, що якщо матриця Y є матрицею траєкторій деякої послідовності $(h_0,...,h_{N-1})$ (іншими словами, Y є матрицею Ханкеля), то $g_i = h_i$ для будь-якого *i*.

Діагональне усереднення (2.3), застосоване до результуючої матриці X_{I_k}, породжує ряд $\tilde{F}^{(k)} = (\tilde{f}_0^{(k)},..., \tilde{f}_{N-1}^{(k)})$, і отже початковий ряд $f_0,..., f_{N-1}$ розкладається на суму *m* рядів:

$$f_n = \sum_{k=1}^{m} \tilde{f}_n^{(k)} \quad (6)$$

2.2 Кроки базового SSA: коментарі

Формальний опис кроків базового SSA вимагає деякого роз'яснення. У цьому розділі коротко обговоримо зміст задіяних процедур.

1.2.1 Згортання

Згортання може розглядатися як перехід від одновимірного часового ряду $F = (f_0,..., f_{N-1})$ до багатовимірного $X_i,...,X_k$ з векторами $X_i = (f_{i-1},..., f_{i+L-2})^T \in R^L$, де $K = N - L + 1$. Вектори X_i називаються векторами з L-відставанням.

Єдиний параметр згортання – довжина вікна L, ціле число, таке що $2 \le L \le N - 1$

Згортання є стандартною процедурою в аналізі часових рядів. Після виконання згортання, подальші розрахунки залежать від мети дослідження.

Серед фахівців в області динамічних систем поширеним методом є отримання емпіричного розподілу всіх попарних відстаней між векторами з відставаням X_i і X_j, а потім обчислення так званої кореляційної розмірності цього ряду. Ця розмірність пов'язана з фрактальною розмірністю атрактора динамічної системи, яка генерує часовий ряд (див відповідний алгоритм у Takens, 1981;. Sauer, Yorke і Casdagli 1991 року, для теорії і Nicolis і Prigogine, 1989, Додаток IV). Зауважимо, що при такому підході величина L повинна бути відносно невеликою, а K повинна бути дуже великою (формально, $K \to \infty$).

Якщо величина L достатньо велика, то можна розглянути кожний вектор з L-відставанням X_i як окремий ряд та дослідити певні динамічні характеристики для такого набору рядів. Найпростішим прикладом такого підходу є відомий метод «ковзного середнього», де обчислюються середні вектори з відставанням. Є й багато більш складних алгоритмів.

Наприклад, якщо початковий ряд можна розглядати як локально стаціонарний процес, то можемо розкласти кожний вектор з відставанням X_i у будь-якому фіксованому базисі (наприклад, базис Фур'є або певний вейвлетний базис) і вивчити динаміку такого розкладення. Ці ідеї відповідають динамічному аналізу Фур'є. Очевидно, що можуть бути застосовані і інші базиси.

Апроксимація стаціонарних рядів за допомогою моделі авторегресії може бути виражена в термінах згортання: якщо ми маємо справу з моделлю

$$f_{i+L-1} = a_{L-1} f_{i+L-2} + a_1 f_i + \varepsilon_{i+L-1}, \quad i \geq 0 \qquad (7)$$

Тоді знайдемо вектор $A = (a_1, ..., a_{L-1}, -1)^T$, такий що внутрішні добутки (X_i, A) описувалися як деякі серії шуму.

Зауважимо, що цей і багато інших методів, які використовують згортання, можна розділити на дві великі частини, які можна назвати «глобальні» і «динамічні». Глобальні методи розглядають X_i як L-мірні вектори, і не використовують їх упорядкування.

Наприклад, якщо обчислити емпіричний розподіл попарних відстаней між векторами з відставанням, то результат не залежить від порядку, в якому ці вектори з'являються. Аналогічна ситуація має місце в моделі авторегресії (7), якщо коефіцієнти a_i

розраховуються по всій колекції векторів з відставанням (наприклад, методом найменших квадратів).

Ця інваріантність щодо перестановки векторів з відставанням не є дивною, оскільки обидві моделі мають справу з рядами стаціонарного типу і призначені для знаходження глобальних характеристик всього ряду. Кількість векторів з відставанням в даних міркуваннях грає роль «величини вибірки», і, отже, вона повинна бути досить великою. Теоретично, в цих підходах величина L повинна бути фіксована, і $K \to \infty$.

Ситуація інша, коли маємо справу з динамічним аналізом Фур'є і подібними методами, і навіть з ковзними середніми. При цьому порядок векторів з відставанням є важливим і описує динаміку, що нас інтересує. Таким чином, нестаціонарний сценарій – основна область застосування цих підходів. Що стосується L і K, їхнє співвідношення взагалі можу бути довільним і має залежати від конкретних даних і конкретної задачі.

У всякому разі, довжина вікна L повинна бути достатньо великою. Значення L повинно бути достатньо великим, щоб кожен вектор з L-відставанням включав важливу частину поведінки початкового ряду серії $F = (f_0,..., f_{N-1})$.

Відповідно до формального опису кроку згортання, результатом цього кроку є матриця траєкторій

$$X = [X_1 : ... : X_K],$$

а не просто набір векторів з відставанням X_i . Це означає, що в цілому ми зацікавлені в динамічних ефектах (хоча деякі характеристики, які є інваріантними щодо перестановок векторів з відставанням, також є важливими).

Матриця траєкторій (2.1) має очевидну властивість симетрії: транспонована матриця X^T є матрицею траєкторій того ж ряду $f_0,...,f_{N-1}$ з довжиною вікна K, а не L.

Сингулярне розкладання матриці траєкторій – другий крок у базовому SSA. Сингулярне розкладання може бути описане в різних умовах і використовуватися для різних цілей. (див. математичні результати у главі). Більшість характеристик сингулярного розкладання дійсні для загальних L × K матриць, але Ханкелева структура матриці траєкторій додає ряд специфічних особливостей. Почнемо з загальних властивостей сингулярного розкладання, важливих для SSA.

Як уже згадувалося, сингулярне розкладання довільної ненульової матриці L × K $X = [X_1 : ... : X_K]$ є розкладанням X у вигляді

$$X = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^T \qquad (8)$$

де $\lambda_i, i = (1,..., L)$ – власні числа мастриці $S=XX^T$, впорядковані за зменшенням.

$$d = \max\{i, таких, що\ \lambda_i > 0\} = rank\ X$$

$\{U_1,...,U_d\}$ – відповідна система ортонормованих власних векторів матриці S, $V_i = X^T U_i / \sqrt{\lambda_i}$.

Стандартна термінологія сингулярного розкладання називає $\sqrt{\lambda_i}$ сингулярними значеннями, U_i і V_i – ліві і праві сингулярні вектори матриці X, відповідно. Сукупність $(\sqrt{\lambda_i}, U_i, V)$ називається і-ю власною трійкою матриці X. Якщо покласти $X_i = \sqrt{\lambda_i} U_i V_i^T$, то представлення (8) можна переписати у вигляді (1.2), тобто у вигляді представлення X як суми елементарних матриць X_i .

Якщо всі власні значення мають кратність один, то розкладання (3) визначається однозначно. В іншому випадку, якщо існує хоча б одне власне значення з кратністю більше 1, то у виборі відповідного власного вектора є елемент свободи. Ми будемо вважати, що власні вектори певним чином вибрані, і вибір фіксований.

Так як сингулярне розкладання має справу з усією матрицею X, воно не є інваріантним при перестановці стовпців $X_1,...,X_K$. Крім того, сингулярне розкладання має наступну властивість симетрії: $V_1,..., V_d$ утворюють ортонормовану систему власних векторів для матриці X^TX, відповідно до тих же власних значень $\sqrt{\lambda_i}$. Зазначимо, що рядки і стовпці матриці траєкторій є підрядами оригінального часового ряду. Таким чином, ліві і праві сингулярні вектори мають часову структуру і, отже, їх можна розглядати як часові ряди.

Сингулярне розкладання має кілька оптимальних властивостей. Одна з таких властивостей полягає в наступному: з усіх матриць $X^{(r)}$ рангу $r < d$, матриця $\sum_{i=1}^{r} X_i$ забезпечує краще

наближення до матриці траєкторій X, так що значення $\| X - X^{(r)} \| \mu$ є мінімальним.

Тут і нижче норма (Фробеніуса) матриці Y – $\sqrt{\langle Y, Y \rangle_\mu}$, де внутрішній добуток двох матриць $Y = (y_{ij})_{i,j=1}^{q,s}$ і $Z = (z_{ij})_{i,j=1}^{q,s}$ визначається як:

$$\langle Y, Z \rangle_\mu = \sum_{i,j=1}^{q,s} y_{ij} z_{ij}$$

Відзначимо, що $\| X \|_\mu^2 = \sum_{i=1}^{d} \lambda_i$ і $\lambda_i = \| X \|_\mu^2$ для $i = 1,...,d$. Отже, будемо розглядати відношення $\lambda_i / \| X \|_\mu^2$ як характеристику вкладу матриці Xi в розкладенні повної матриці траєкторій X. Звідси, $\sum_{i=1}^{r} \lambda_i / \| X \|_\mu^2$, сума перших r співвідношень, є характеристикою оптимальної апроксимації матриці траєкторій матрицями розмірності r.

Розглянемо тепер матрицю траєкторій X як послідовність векторів з L-відставанням. Позначимо $\zeta^{(L)} \subset R^L$ лінійний простір, утворений векторами $X_1,...,X_k$. Будемо називати цей простір *простором L-траєкторій* (або просто *простором траєкторій*) ряду F. Щоб підкреслити роль ряду F, будемо використовувати запис $\zeta^{(L)}F$ замість $\zeta^{(L)}$. Рівність (1.7) показує, що $U = (U_1,...,U_d)$ – ортонормований базис в d-мірному просторі траєкторій.

Встановлюючи $Z_i = \sqrt{\lambda_i} V_i$, $i = 1,...,d$, можемо переписати розкладення (1.7) в формі:

$$X = \sum_{i=1}^{d} U_i Z_i^T , \qquad (9)$$

і для векторів з відставанням маємо:

$$X_j = \sum_{i=1}^{d} z_{ij} U_i \qquad (10)$$

де z_{ij} – компоненти вектору Z_i.

z_{ij} - і-й компонент вектору Xj, представлений в базисі U. Іншими словами, вектор Z_i складається з і-х компонент векторів з затримкою, представлених в базисі U.

Розглянемо тепер транспоновану матрицю X^T. Вводячи $Y_i = \sqrt{\lambda_i} U_i$, отримуємо розкладення:

$$X^T = \sum_{i=1}^{d} V_i Y_i^T ,$$

що відповідає представленню послідовності векторів з *K*-затримкою в ортонормованому базисі $V_1,...,V_d$. Таким чином,сингулярне розкладання дає підставу для двох геометричних описів матриці траєкторій *X*.

Оптимальна особливість сингулярного розкладання, розглянута вище можна переформулювати на мові багатовимірної геометрії для векторів з *L*-відставанням наступним чином. Нехай *r*<*d*. Тоді серед усіх *r*-мірних підпросторів ζ_r *з* R^L, підпростір $\zeta_r^{(0)} \overset{def}{=} \zeta(U_1,...,U_r)$, утворений $(U_1,...,U_r)$, апроксимує ці вектори найкращим чином; тобто мінімум досягається при $\zeta_r^{(0)}$. Відношення $\sum_{i=1}^{r} \lambda_i / \sum_{i=1}^{d} \lambda_i$ є характеристикою найкращої r-мірної апроксимації векторів з відставанням.

Ще одна оптимальна особливість відноситься до властивостей напрямів, що визначаються власними векторами $U_1,...,U_d$. Зокрема, перший власний вектор U_1 визначає напрям, на якому варіація проекції вектора з відставанням є максимальною.

Кожен наступний власний вектор визначає напрям, який є ортогональним до всіх попередніх напрямів, і на якому варіація проекції вектора з також є максимальною. Тому, природно назвати напрямок і-го власного вектора Ui і-м основним напрямом. Зазначимо, що елементарні матриці $X_i = U_i Z^T$ будуються з проекцій векторів з відставанням на і-й напрям.

Ця точка зору на сингулярне розкладання матриці траєкторій, що складається з векторів з L-відставанням, і звернення до асоціації з аналізом головних компонент веде до наступної термінології. Будемо називати вектор Ui і-м (основним) власним вектором,

вектор Vi називатимемо вектор і-го фактора, а вектор Zi – вектором і-х головних компонент.

3 Виявлення розладнань статистичних характеристик часових рядів

Розбиття ряду на ділянки з подібними статистичними характеристиками дозволяє отримати компактну інформацію про динаміку коливання цін з великих вибірок даних, значно зменшуючи розмірність подальших розрахунків. Скорочення об'ємів вхідних даних у такий спосіб дозволяє працювати в процесі прогнозування з великими часовими проміжками історії цін, використовуючи при цьому вхідні набори даних помірної величини.

Розладнанням випадкового процесу називається стрибкоподібна зміна його властивостей, що відбувається в невідомий момент часу τ, або не відбувається взагалі. Завданням виявлення розладнання є встановлення факту розладнання, і якщо таке сталося, оцінювання моменту часу τ [3].

Найчастіше розглядається випадок з дискретним часом, $t = 1,2...N$. Випадковий процес в цьому випадку є часовим рядом.

З математичної точки зору, загальна постановка задачі виявлення розладнання часового ряду $y(t), t = \overline{0,t_n}$ полягає в перевірці гіпотези H_0 про те, що випадкові величини y мають один і той же безумовний розподіл F_0 з деякої множини розподілів. Альтернативною є гіпотеза H_1 про кускову стаціонарність, тобто про існування такого моменту часу $\tau \geq 1$, що при $t < \tau$ розподілом випадкових величин y_t є F_0, а при $t \geq \tau$ відмінний від F_0 деякий розподіл F_1.

3.1 Класифікація задач про розладнання

Задачі пошуку розладнання зазвичай класифікуються за такими критеріями:

1. Метод отримання даних.

Відповідно до цього критерію виділяється два класи задач.

У першому випадку дані надходять поступово, і необхідно виявити розладнання якомога швидше після його появи, але не

потрібно точно вказувати момент часу, коли сталося розладнання. Ця задача, відома як задача якнайшвидшого виявлення розладнання, часто виникає при поточному контролі якості продукції, в радіолокації, гідроакустиці і скрізь, де функція втрат залежить від часу між моментом появи розладнання і моментом його виявлення. Методи якнайшвидшого виявлення розладнання також називають послідовними методами.

Другий основний клас задач зводиться до оцінювання моменту появи розладнання за наявності повної вибірки експериментальних даних, яка збирається до початку рішення задачі. Задача полягає в тому, щоб оцінити момент появи розладнання якомога точніше. В деяких випадках сам факт наявності розладнання в межах аналізованої вибірки заздалегідь невідомий, і перевірка його наявності також є предметом рішення. Методи вирішення даного класу задач, відомі як апостеріорні, використовуються при обробці геофізичних даних з сейсмоприймачів, а також в інших галузях, де критична точність визначення розладнання в часі.

Можливий комбінований підхід – алгоритми, в яких за допомогою послідовних методів виявляється факт наявності розладнання, після чого він уточнюється за допомогою апостеріорних алгоритмів.

Виявлення зміни ринкової тенденції належить до задач послідовного виявлення розладнань, для якої можливість роботи з найновішими даними є важливішою за максимізацію точності виявлення моменту розладнання у часі, тому розділ присвячений переважно послідовним алгоритмам. Апостеріорні методи розглядаються переважно у їхньому зв’язку з послідовними аналогами.

2. Повнота апріорно відомої статистичної інформації (ознака параметризації). Згідно з цим критерієм, виділяють параметричні, семіпараметричні і непараметричні методи виявлення розладнання.

3. Розмірність розглянутих даних. На підставі цього критерію, задачі про розладнання поділяються на задачі про розладнання випадкового процесу і задачі про розладнання випадкового поля.

4. Характер розладнання. За цією ознакою виділяють процеси з одиничним розладнанням і багаторазовими розладнанням. Для виявлення багаторазового розладнання як правило можуть бути

використані ті ж методи, що і для виявлення одиничного, застосовані декілька разів. Також іноді виділяють особливий тип розладнання, при якому зміни відбуваються протягом деякого інтервалу часу, а не стрибкоподібно.

3.2 Опис існуючих методів виявлення розладнання

Розглядаючи послідовні методи виявлення розладнання, насамперед можна виділити кілька великих групах методів, заснованих на загальних підходах: алгоритм Гіршика-Рубіна-Ширяєва (ГРШ, GRSh) і його похідні; алгоритми, засновані на накопиченні кумулятивних сум (також відомі як АКС або CUSUM); алгоритми, засновані на лемі Неймана-Пірсона; алгоритми, що використовують експоненціальне згладжування. У першу чергу розглянемо основні методи, що належать до цих груп.

Алгоритм Гіршика-Рубіна-Ширяєва. Розглядається випадок поточного контролю виробничого процесу, який може знаходитися в двох станах – налагодженому і розладнаному і має відомі ймовірності переходу з одного стану в інший. Авторами було запропоновано правило виявлення розладнання, згідно з яким на кожному кроці виявлення розраховувалася ймовірність того, що процес перебуває в розладнаному стані. Правило подачі сигналу про розладнання в цьому випадку виглядає наступним чином [4]:

$$\tau = \inf(t : \pi_t > \lambda),$$

де τ – момент розладнання;

λ – поріг визначення.

Нехай в першому стані щільність ймовірності значень ряду дорівнює $f(x_t, \theta_0)$, а в другому (розладнаному) – $f(x_t, \theta_0)$. Тоді:

$$w = \frac{f(x_t, \theta_0)}{f(x_t, \theta_1)}.$$

На кожному кроці накопичується добуток:

$$W_t = w_t (1 + W_{t-1}),$$

$$W_0 = 0.$$

Правило подачі сигналу про розладнання має вигляд:

$$\tau = \inf(t : W_t \geq b),$$

де b – чутливість виявлення.

Існує інше тлумачення методу.

Нехай момент τ розподілений по геометричному закону, і

ймовірність переходу в розладнаний стан становить

$$P\{t_0 = k\} = a\,(1-a)^{k-1},$$

тоді процес можна промоделювати за допомогою марківського ланцюга з матрицею станів:

$$P = \begin{vmatrix} 1-a & 0 \\ a & 1 \end{vmatrix}.$$

Нехай ймовірність початкових станів $p_0 = 1-\pi, \quad p_1 = \pi$. Тоді на кожному кроці ймовірність переходу в розладнаний стан можна розраховувати по формулі:

$$z_t = \pi_t /(1-\pi_t).$$

Рекурентна формула для значень ймовірності:

$$z_t = (1/(1-a))(z_{t-1} + a)(\omega_2(x_t)/\omega_1(x_t)).$$

Прологарифмувавши обидві частини і позначивши $\ln z_t = g_t$, отримуємо:

$$g_t = \ln(a + e^{g_{t-1}}) - \ln(1-a) + \ln(\omega_2(x_t)/\omega_1(x_t)).$$

Для випадку зміни середнього значення нормального розподілу:

$$g_t = \ln(a + e^{g_{t-1}}) - \ln(1-a) + ((\theta_2 - \theta_1)/\sigma^2)(x_t - (\theta_2 + \theta_1)/2).$$

Правило зупинки в цьому рішенні застосовується з новим порогом *h*:

$$\tau = \inf\{t : g_t \geq h\}.$$

У випадку, якщо можливий перехід в налагоджений стан з розладнаного з ймовірністю *b*, матриця станів набуває вигляду:

$$P = \begin{vmatrix} 1-a & b \\ a & 1-b \end{vmatrix}.$$

Алгоритми, засновані на накопиченні кумулятивних сум. Алгоритм кумулятивних сум (АКС, CUSUM) був розроблений Е. С. Пейджем [5]. Він являє собою послідовний критерій відношення ймовірності (ПКВЙ) для двох простих гіпотез H_1 (немає розладнання): $\theta = \theta_1$ і H_2 (є розладнання): $\theta = \theta_2$, де θ – деякий скалярний параметр щільності розподілу $f(x_t/\theta)$.

Ідея Е. С. Пейджа полягає в аналізі поведінки кумулятивної суми

$$S_t = S_{t-1} + \ln(f(y_t/\theta_2)f(y_t/\theta_1)).$$

Член $\ln(f(y_t/\theta_2)f(y_t/\theta_1)$ є логарифмом відношення

правдоподібності і забезпечує облік інформації про нові дані процесу, що спостерігається.

У ПКВЙ кумулятивна сума S_t порівнюється на кожному кроці з двома порогами: ε і h (ε і $h>0$). Якщо на кроці t $g_t>h$, приймається гіпотеза H_2, якщо $g_t<-\varepsilon$ – гіпотеза H_1. У такому варіанті ПКВЙ порушено припущення про приналежність всієї вибірки до гіпотези H_1 або H_2, тому застосувати його до задачі про розладнання не можна.

Пейдж запропонував на кроці t відновлювати накопичення суми з нуля після того, як на кроці $t-1$ була прийнята гіпотеза H_1 (нерозладнаність процесу). Після того, як гіпотеза H_2 змінить H_1, тобто після розладнання, математичне очікування логарифма правдоподібності буде позитивним, і сума S почне рости. Поріг ε Пейдж запропонував встановити рівним нулю (оптимальність такого вибору згодом встановили А. Ширяєв і Т. Лорден). Таким чином:

$$S_t = \max(0,\, S_{t-1} + g_t),$$

$$g_t = \ln(f(y_t / \theta_2) / f(y_t / \theta_2)).$$

Сигнал про розладнання подається у момент часу:

$$\tau = \inf(t \geq 1 : S_t > h).$$

Існує інше тлумачення АКС. На кожному кроці із заданим порогом порівнюється різниця

$$g_t = S_t - \min_{k<t} S_k. \qquad (11)$$

Вказані формули справедливі для випадку, коли середня величина θ збільшується. У випадку, якщо необхідно виявляти зміни θ у бік зменшення, різниця має вигляд:

$$g_t = \max_{k<t} S_k - S_t. \qquad (12)$$

Сигнал про розладнання подається у момент часу $\tau = \inf(t \geq 1 : g_t > h)$. Порівняння сум (11) і (12) з контрольною межею h може проводитися одночасно, щоб виявляти відхилення у будь-який бік.

Знаючи тип розподілу та визначаючи одну з імовірнісних характеристик розподілу як параметр θ, можна отримати рекурентні формули накопичення кумулятивної суми. Так, у разі нормального розподілу, щільність якого дорівнює

$$f = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(y-m)^2}{2\sigma^2} \right).$$

Шляхом простих підстановок можемо отримати формули для приросту кумулятивної суми:

$$g_t = \frac{m_2 - m_1}{\sigma^2} \left(y - \frac{m_1 + m_2}{2} \right),$$

де m_1 – математичне очікування величини до розладнання;

m_2 – передбачуване математичне очікування величини після розладнання;

σ – середнє квадратичне відхилення;

y – поточне значення спостереження.

Наряду з послідовним методом, Пейджем була запропонована обчислювальна схема АКС для апостеріорного застосування:

$$t_0 = \inf\{k : S_{k-1} \geq S_j, j = 2, N\},$$

$$S_k = \sum_{i=1}^{k} (\ln f(y_i / \theta_1) - \ln f(y_i / \theta_2)).$$

На підставі послідовного аналізу Вальда Г. Лорденом у роботах [6] була розроблена процедура максимальної правдоподібності. Як і АКС Пейджа, процедура заснована на однобічній процедурі послідовного критерію відношення ймовірності:

$$\tau = \inf\left\{ t \geq 1 : \max_{k \leq t} \sum_{i=1}^{t} \ln f\left(y_i / \theta_2\right) - \ln f\left(y_i / \theta_1\right) \geq h \right\}.$$

Нехай розподіл y відноситься до експоненціального сімейства розподілів зі щільністю:

$$f(y_t \mid \theta) = \exp(\theta T(y) - b(\theta)),$$

де $b(\theta)$ – строго вигнута вгору функція, що диференціюється на всієї області визначення.

Для прийнятого сімейства розподілів можна припустити, що при $\theta = 0$ $b(\theta) = 0$, за необхідності центруючи вибірку відносно середнього значення. Логарифм відношення правдоподібності гіпотези про присутність розладнання проти гіпотези про стаціонарність ряду рівний $\theta S_n - nb(\theta)$. Правило максимальної правдоподібності переписується у вигляді:

$$\tau = \inf\left\{ t \geq 1 : \sup_{k \leq t} \sum_{i=k}^{t} T\left(y_i\right) - (t-k+1)b\left(\theta\right) \geq h \right\}.$$

Правило двобічного виявлення можна представити у формі U-маски: на кожному кроці накопичуючи суму $S_k^t = \sum_{i=k}^{t} T(y_i)$, можна порівнювати її з криволінійними порогами:

$$\bar{c}_l = \inf_{\theta > \widetilde{\theta}} \{h / \theta + lb(\theta) / \theta\},$$

$$\underline{c}_l = \sup_{\theta < -\widetilde{\theta}} \{h / \theta + lb(\theta) / \theta\},$$

де $\widetilde{\theta}$ – величина допускового інтервалу відхилення навколо θ.

На відміну від АКС Пейджа, ефективність якого падає при відхиленні від передбачуваного значення θ_2, алгоритм Лордена рівномірно ефективний для деякої множини значень θ_2. Проте, недоліком методу є складність отримання рекурентного запису, що приводить до великої ресурсоємності рішення за допомогою ЕОМ.

Методи, засновані на лемі Неймана-Пірсона. Дана група методів заснована на перевірці гіпотези $\theta = \theta_1$ проти $\theta = \theta_2$, що виконується на кожному кроці для допоміжної вибірки обсягом $\widetilde{N}$: $\{y_t^{t+\widetilde{N}-1}\}$ згідно критерію максимальної правдоподібності. Для цієї вибірки обчислюється кумулятивна сума і порівнюється з порогом h, як і у попередньому випадку. У разі нормального розподілу кумулятивна сума може бути обчислена за формулою:

$$S_{\widetilde{N}}^t = ((\theta_2 - \theta_1) / \sigma^2)(\sum_{i=t}^{t+\widetilde{N}-1} y_i - \widetilde{N}(\theta_2 + \theta_1) / 2)$$

Для даного випадку правило виявлення є еквівалентним використанню карт Шухарта.

Карти Шухарта, запропоновані Уолтером Шухартом [7], є одним із найстаріших відомих методів контролю виробництва. В основу цього методу покладена побудова показника, що відповідає еталонному значенню характеристики. Як правило, еталонним служить середнє арифметичне значення спостережень на певному інтервалі (допоміжній вибірці).

Карта Шухарта має дві статистично визначувані контрольні межі щодо центральної лінії, що проводяться на відстані $k\sigma$ від центральної лінії, де σ – дисперсія випадкової величини, k – деяка

контрольна межа (найчастіше використовується значення *k=3*). Таким чином, сигнал про розладнання подається, якщо

$$y_{\widetilde{N}} > m + k\sigma$$

$$y_{\widetilde{N}} < m - k\sigma$$

$$y_{\widetilde{N}} = \frac{1}{\widetilde{N}} \sum_{i=t}^{t+\widetilde{N}-1} y_i$$

де $\widetilde{N}$ – обсяг допоміжної вибірки;

m – математичне очікування.

Частковий випадок, коли в якості центральної лінії використовується ковзне середнє, а допустиме відхилення дорівнює 2σ, являє собою лінії Болліндже́ра – індикатор, що широко використовується в технічному аналізі фінансових ринків.

Даний алгоритм призначений тільки для знаходження зміни математичного очікування послідовності. Це не є значним недоліком, оскільки будь-яка задача про розладнання може бути зведена до зміни математичного очікування, якщо розглядати замість початкової послідовності сформовані з неї нові послідовності.

Алгоритми, що засновані на експоненціальному згладжуванні. Метод, що заснований на експоненціальному згладжуванні, описаний в роботі [8].

На кожному кроці спостереження накопичується сума

$$S_t = (1-k)S_{t-1} + k(y_t - m),$$

де m – математичне очікування.

Замість y може бути використане середнє значення $y_{\widetilde{N}}$, отримане з допоміжної вибірки. Також можливий варіант застосування алгоритму, коли величина k залежить від попередніх значень S_t, і, таким чином, здійснюється пристосування до зміни статистичних властивостей послідовності.

У разі збільшення середнього сигнал про розладнання подається згідно правила:

$$\tau = \inf(t \geq 1 : S_t > h)$$

Записавши рівняння у вигляді

$$\tau = \inf(t \geq 1 :| S_t |> h),$$

тобто, додавши нижню межу, отримаємо правило для двобічної процедури.

Існує модифікація даного методу, в якій накопичуються дві суми [9]:

$$S_t = (1-k)S_{t-1} + k(y_t - m),$$
$$R_t = (1-k)R_{t-1} + k\,|\,y_t - m\,|,$$

і на підставі їх обчислюється параметр:

$$G(n) = \frac{S(n)}{R(n)}.$$

Верхній і нижній пороги спрацьовування алгоритму:

$$-1 < h_1 < h_2 < 1.$$

Сигнал про розладнання подається, якщо $G \geq h_2$ або $G \leq h_1$.

Непараметричні алгоритми виявлення розладнання. Алгоритми, що вимагають інформації про розподіл до і після розладнання, є точними, але, в той же час, можуть бути застосовані не завжди. Часто інформацію про розподіл після розладнання неможливо отримати заздалегідь. В цьому випадку можливе налаштування моделі на підставі деякого попереднього діапазону даних, але це породжує замкнене коло: метод, що призначений для перевірки однорідності даних, налаштовується на підставі фактично того ж часового ряду, однорідність якого перевіряється. У такому разі кращим рішенням є використання непараметричних методів, що не вимагають апріорної інформації про розподіл.

Приведена вище форма АКС потребує інформації про значення параметра θ після розладнання. Існує декілька способів ослабити ці вимоги. Перший з них – використання ковзаючого вікна.

Метод з ковзаючим вікном полягає в тому, що значення характеристик розподілу обчислюються на інтервалах [t-*l*;y] [t;t+*l*] заданої довжини l до і після точки передбачуваного розладнання. Для обчислень потрібно l значень часового ряду після спостережуваного (рисунок 2).

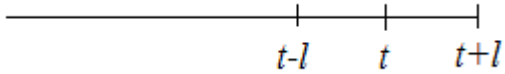

Рисунок 2 – Інтервали ковзаючого вікна

Інший підхід полягає в налаштуванні алгоритму кумулятивних сум так, щоб він реагував на будь-які зміни заданої характеристики розподілу.

Нехай $\breve{\theta}_2$ – невідома величина, але відомий напрям її зміни, наприклад $\theta_2 > \theta_1$. АКС зберігає здатність виявляти розладнання, якщо $\breve{\theta}_2 > \dfrac{\theta_2 + \theta_1}{2}$. Якщо відомо, що величина параметра після розладнання збільшується, то за умови нормального розподілу досліджуваної величини, для розрахунку доданків g може бути застосований наступний підхід [10]:

$$g_t = y_t - m_1 - k \quad при\ m_2 > m_1 \tag{13}$$

Для випадку зменшення m:

$$g_t = -y_t + m_1 + k \quad при\ m_2 < m_1 \tag{14}$$

де $k \geq 0$ – поріг чутливості методу для відхилення від θ_1;

m_1, m_2 – математичне очікування величини до і після розладнання

У роботі [11] була доведена еквівалентність процедур (13), (14), вживаних одночасно з нульовим допуском, до наступної процедури:

$$R_t = \max \sum_{i=1}^{m} (y_i - \theta_i) - \min_{m \leq t} \sum_{i=1}^{m} (y_i - \theta_i).$$

Також Хінклі [12] були запропоновані апостеріорні обчислювальні схеми на основі даного підходу, в тому числі для випадку, коли розподіл до і після розладнання невідомі:

$$t_0 = \arg\max_{1 \leq t \leq N-1} (t(N-t)(\bar{y}_t - y^*_t)^2 / N),$$

$$\bar{y}_t = t^{-1} \sum_{i=1}^{t} y_i,$$

$$\bar{y}^*_t = (N-t)^{-1} \sum_{i=1}^{N} y_i.$$

Існує наступний метод виявлення зміни медіани випадковій послідовності, що не вимагає знання функції розподілу:

$$g_t = \max(0, g_{t-1} + \Delta g(y_t - k)),$$

$$\Delta g(y) = \begin{cases} 1 & при\ x \geq 0, \\ -1 & при\ x < 0. \end{cases}$$

де k – значення медіани;

g – функція простого одновимірного блукання з екраном, що відображає.

Бродським і Дарховським [13] був запропонований непараметричний метод послідовного визначення, заснований на обчисленні наступної статистики:

$$Z(n) = \max_{\lfloor \alpha M \rfloor \le k \le \lfloor (1-\alpha) M \rfloor} \left| Y_M(k,n) \right|,$$

$$Y_M(k,n) = \frac{1}{k} \sum_{n-M+1}^{n-M+k} y(t) - \frac{1}{M-k} \sum_{n-M+k+1}^{n} y(t),$$

$$0 < \alpha < \frac{1}{2},$$

де n – довжина часового ряду.

Сигнал про розладнання подається, якщо значення $Z(n)$ перевищує поріг визначення c.

Алгоритм має три параметри налаштування: α, M, c. Із збільшенням об'єму пам'яті M поліпшується якість виявлення, тому його можна вибирати виходячи з обчислювальних можливостей. Значення $\alpha < 2M$ вибирається з урахуванням того, що величина затримки визначення має порядок αM. Значення c настроюється експериментальним шляхом для конкретного числового ряду.

Подібний алгоритм був запропонований авторами для апостеріорного виявлення:

$$S(n) = \left[\frac{n}{N} \left(1 - \frac{n}{N} \right) \right]^{\nu} \left(\frac{1}{n} \sum_{t=1}^{n} y(t) - \frac{1}{N-n} \sum_{t=n+1}^{N} y(t) \right),$$

$$0 \le \nu \le 1,$$

$$t_0 = \arg \max_{[aN] \le n \le [bN]} \left| S(n) \right|,$$

$$0 < a < \frac{1}{2} < b < 1,$$

де t_0 – момент розладнання.

Асимптотично найкращий метод щодо ймовірності помилкового сигналу: $\nu = 1$, щодо ймовірності помилкового спокою: $\nu = 0$. Асимптотично мінімаксний метод: $\nu = \frac{1}{2}$.

Тими ж авторами були запропоновані непараметричні модифікації раніше згаданих методів послідовного виявлення.

1. Алгоритм кумулятивних сум. Авторами запропоновано використовувати значення ряду x як прирощення суми, а накопичену суму порівнювати з деяким «великим» числом c.

$$S_t = \max(0, (y(t-1) + y(t))),$$

$$\tau = \inf(t : y(t) \geq c).$$

Даний метод призначений для виявлення зміни середнього значення ряду (у бік збільшення). Двобічна процедура може бути отримана шляхом узяття значення *y* по абсолютній величині:

$$\tau = \inf(t :| S_t | \geq c).$$

У початковому вигляді метод призначений для роботи з рядами з нульовим середнім. Якщо даний ряд не є центрованим, необхідне додаткове центрування значень відносно математичного очікування ряду, підрахованого на діапазоні *(t-l;t)* заданої довжини *l*.

2. Алгоритм ГРШ.

У даній модифікації методу замість відношення щільності ймовірності використовується експоненціальна функція:

$$W_t = (1 + W_{t-1})\, e^{y(t)}.$$

Як і в базовому методі ГРШ, вирішальне правило має вигляд:

$$\tau = \inf(t : W_t \geq b).$$

3. Метод кумулятивних сум, заснований на експоненціальному згладжуванні.

$$S_t = (1-k)S_{t-1} + ky(t).$$

В роботі [14] запропоновано наступний непараметричний метод визначення зміни середнього:

$$S_t = \max((p+q), (S_{t-1} + q\ sign(y(t) - y(t-m)) - p)),$$

де $q > p$ – натуральні нескорочувані числа;

sign(x) – функція, що повертає знак числа x;

$p + q$ – поріг чутливості алгоритму.

Метод виявляє зміну середнього в сторону збільшення. Для визначення змін в сторону зменшення перепишемо рівняння у вигляді:

$$S_t = \max((p+q), (S_{t-1} - q\ sign(y(t) - y(t-m)) + p)).$$

Вирішальна функція алгоритму, заснованого на принципі нев'язок, формується як нев'язка (розбіжність) між моделлю випадкового процесу, що спостерігається, і прийнятою раніше моделлю [15]:

$$S(t) = \frac{1}{\sqrt{2n}} \sum_{t=n-M}^{n} \left(\left(\frac{y(t) - m_t}{\sigma_t^2} \right)^2 - 1 \right), \qquad (15)$$

де M – глибина пам'яті;

m_i – математичне очікування процесу до розладнання;

σ_i^2 – дисперсія до розладнання.

Формула (15) може бути перетворена до рекуррентного вигляду:

$$S(t) = \sqrt{\frac{n-1}{n}} \cdot G(n-1) + \frac{1}{\sqrt{2n}} \left(\left(\frac{y_n - m_n}{\sigma_n^2} \right)^2 - 1 \right),$$

$$S(1) = \frac{1}{\sqrt{2}} \left(\left(\frac{y_1 - m_1}{\sigma_1^2} \right)^2 - 1 \right).$$

Алгоритм, заснований на перевірці узагальненої дисперсії може бути застосований для виявлення розладнання в оновлюючих послідовностях фільтра Калмана [16]. Запис алгоритму в матричному вигляді:

$$G(n) = \begin{cases} 0, n < M \\ \det(S_n), \ n \geq M \end{cases},$$

де $S_n = \frac{1}{M-1} \sum_{i=n-M+1}^{n} (y(t_i) - m_n)(y(t_i) - m_n)^T$, $n \geq M$;

$\det(S_n)$ – визначник матриці S_n;

M – глибина пам'яті;

m – математичне очікування послідовності до розладнання:

$$m_n = \frac{1}{M} \sum_{i=n-M+1}^{n} y(t_i), n \geq M .$$

Незважаючи на те, що алгоритм призначений для роботи з матричними параметрами y, він може бути застосований і для аналізу одновимірних послідовностей. Прийнявши розмірність y за одиничну, приходимо до рекурентних формул:

$$S_n = S_{n-1} + \frac{y_n - y_{n-M}}{M-1}\left(y_n + y_{n-M} - 2m_{n-1} - \frac{1}{M}\right),$$

$$m_n = m_{n-1} + \frac{y_n - y_{n-M}}{M}, n > M .$$

Початкові умови:

$$S_M = \frac{1}{M-1}\sum_{i=1}^{M}\left(y(t_i) - m_M\right)^2 .$$

Метод визначення мінімуму інформаційної неузгодженості також не потребує для роботи даних про характеристики розподілу після розладнання. Алгоритм заснований на знаходженні інформаційної неузгодженості автоковаріаційних матриць, побудованих на основі двох вибірок: $Y_1 = (y_{1,1}, y_{1,1}, ..., y_{1,M_1})$ і $Y_1 = (y_{2,1}, y_{2,1}, ..., y_{2,M_2})$ об'ємом M_1 і M_2 відповідно.

$$S_1 = \frac{1}{M_1}\sum_{i=1}^{M_1} y_{1,i}\ y_{1,i}^T ,$$

$$S_2 = \frac{1}{M_2}\sum_{i=1}^{M_2} y_{2,i}\ y_{2,i}^T ,$$

$$S_0 = (M_1 / M_0)S_1 + (M_2 / M_0)S_2 ,$$

де $M_0 = M_1 + M_2$.

Сигнал про розладнання подається за умови:

$$\lambda(X_0) = M_1\ \gamma_{1,0} + M_2\ \gamma_{2,0} - 0{,}5(M_0 n) \geq \ln(\lambda_0) ,$$

де $\gamma_{k,0} = 0{,}5(tr(S_k\ S_0^{-1}) - \ln|\ S_k\ S_0^{-1}|)$ – величина інформаційної неузгодженості,

|·| – визначник матриці;

tr – слід квадратної матриці;

n – константа.

Метод передбачає, що процес X є центрованим з нульовим математичним очікуванням.

4 Ймовірнісне прогнозування часових рядів

Імовірнісний характер моделі забезпечує максимальну прозорість для користувача, який приймає рішення, що є особливо актуальним в умовах великої кількості факторів, які впливають на рух динамічного випадкового процесу.

В основі моделі [17], яку розглянемо, лежить припущення, що час появи розладнання в статистичних характеристиках ряду є випадковою величиною, функція розподілу якої повільно змінюється в часі. Відтворивши щільність розподілу, отримаємо можливість на кожному кроці спостережень судити про ймовірність зміни напрямку тренду процесу.

Модель заснована на математичних методиках визначення точок розладнання статистичних характеристик ряду і методі відтворення щільності імовірнісного розподілу з використанням сплайнів. На першій стадії обробки даних вхідний числовий ряд розбивається на відрізки, напрям тренду на яких незмінний. Після цього відбувається відтворення двовимірної функції розподілу тривалості T таких ділянок і різниці Y між кінцевим і початковим значенням ряду на відрізку. На підставі отриманої функції розподілу обчислюється функція розподілу ймовірності виникнення розладнання.

Як відомо, функцією розподілу випадкової величини X називається функція $F(x)$, що дорівнює ймовірності того, що випадкова величина X в результаті випробування прийме значення, яке менше за x:

$$F(x)= P(X<x).$$

$F(x)$ – універсальна характеристика випадкової величини і є однією з форм закону розподілу. Функція розподілу може бути задана як для дискретної, так і для безперервної випадкової величини.

Геометричний зміст функції розподілу такий: $F(x)$ – є ймовірність того, що випадкова величина X в результаті випробування прийме значення, яке на числовій осі лежить лівіше точки x.

$F(x)$ – неспадна функція, значення якої належать відрізку [0, 1]:

$$0<F(x)<1.$$

Ймовірність того, що випадкова величина прийме значення в інтервалі $(x_1<X< x_2)$ дорівнює приросту функції розподілу на цьому інтервалі:

$$P(x_1<X<x_2)=F(x_1) - F(x_2).$$

Для неперервних випадкових величин використовується функція щільності ймовірності, що є похідною від функції розподілу:

$$f(x)=F'(x),$$

$$\int_{-\infty}^{x} f(x)dx = F(x).$$

Сума всіх її значень дорівнює 1 (ймовірність всіх можливих варіантів розвитку подій):

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

Сказане поширюється на двовимірний випадок у наступному вигляді [18]:

$$F(x, y) = P(X < x, Y < y),$$

$$0 < F(x, y) < 1,$$

$$\int_{-\infty}^{x}\int_{-\infty}^{y} f(x, y)dxdy = F(x, y),$$

$$\int_{-\infty}^{\infty} f(x, y) = 1,$$

$$F(x_1 < X < x_2, y_1 < Y < y_2) = \int_{x_1}^{x_2}\int_{y_1}^{y_2} f(x, y)dydx.$$

Сказане справедливо, якщо між величинами x і y існує ймовірнісна залежність. Якщо між двома величинами існує така залежність, то, знаючи значення однієї величини, неможна вказати точно значення іншої, натомість можна вказати закон її розподілу, який залежить від того, яке значення прийняла інша величина.

Випадкові величини є незалежними, якщо закон розподілу кожної з них не залежить від того, яке значення прийняла інша. Якщо величини незалежні, можна записати щільність розподілу системи як добуток щільностей розподілу окремих величин, які входять в систему, що суттєво спростило б задачу:

$$f(T, Y) = f(T)f(Y) \ ,$$

Якщо ж результати перевірки гіпотези про незалежність випадкових величин T і Y (де T – час від розладнання до розладнання, Y – різниця на початку і в кінці відрізку між двома сусідніми подіями розладнання) з обчисленням критеріїв χ^2 і точного критерію Фішера вказують на наявність ймовірнісної залежності між функціями розподілу величин T і Y, то спрощення

форми розподілу розподілу вектора до добутку двох одновимірних розподілів є неможливим. Тоді шуканий розподіл є двовимірним і описується вектором випадкових величин (T, Y), типові графік і гістограма розподілу якого показані на рисунках 3 і 4.

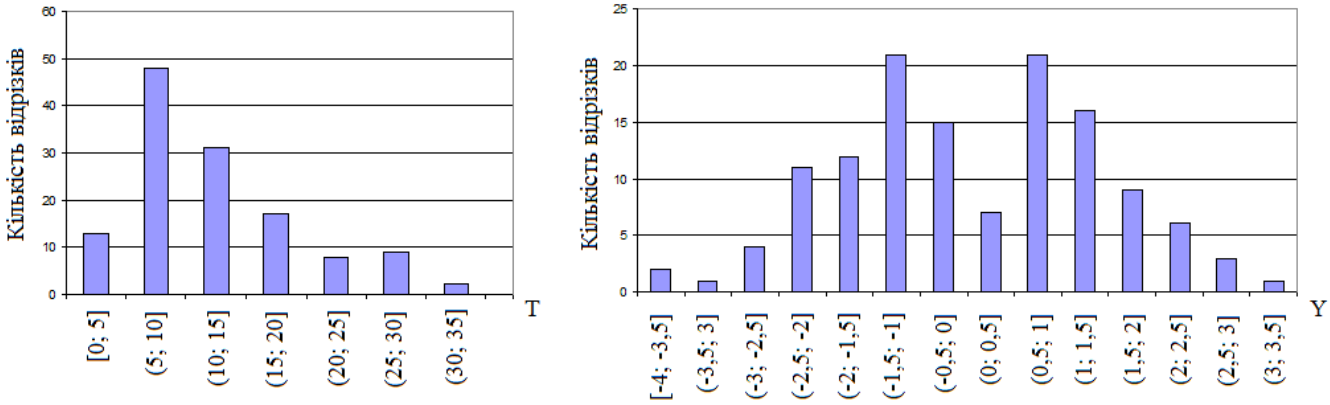

Рисунок 3 – Гістограми розподілів величин T і Y

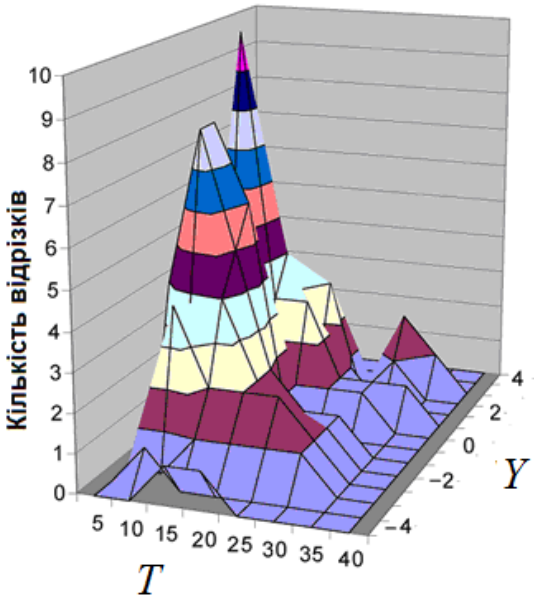

Рисунок 4 – Гістограма розподілу вектору (T, Y)

Щоб вирішити задачу відтворення щільності, необхідно знайти клас функцій, який дозволить апроксимувати емпіричні дані про розподіл з таким ступенем точності, щоб отриманий результат апроксимації був максимально близький до передбачуваної теоретичної залежності.

Отримана функція розподілу має складну форму з кількома вираженими піками і, очевидно, не може бути апроксимована з прийнятним ступенем точності жодним із стандартних двовимірних розподілів. Для вирішення задачі потрібні більш гнучкі методи відтворення щільності.

Спробуємо оптимізувати отриману функцію розподілу сумішшю двомірних нормальних розподілів. Для цього застосуємо ЕМ-алгоритм, який припускає, що дані можуть бути кластеризовані і підпорядковуються лінійній комбінації (суміші) розподілів:

$$f(x) = \sum_{j=1}^{k} \omega_j f_j,$$

$$\sum_{j=1}^{k} \omega_j = 1,$$

$$\omega_j \geq 0.$$

де p_j – функція правдоподібності j-ї компоненти суміші;

ω_j – її апріорна ймовірність.

Нехай функція правдоподібності належить до параметричного сімейства $\varphi(x;\theta)$ і відрізняються лише значеннями параметра θ.

Задача розділення суміші розподілів полягає в тому, щоб, маючи вибірку X^m випадкових спостережень з суміші, знаючи число k і функцію $\varphi(x;\theta)$, оцінити вектор параметрів $\Theta = (\omega_1,...\omega_k, \theta_1,...\theta_k)$.

ЕМ-алгоритм складається з ітераційного повторення двох кроків:

1. На Е-кроці обчислюється значення вектору прихованих змінних G.

$$g_{ij} \equiv P(\theta_j \mid x_i),$$

$$\sum_{j=1}^{k} g_{ij} = 1,$$

$$g_{ij} = P(Z_i = j \mid X_i = x_i; \theta^{(t)}) = \frac{\omega_j f_j(x)}{\sum_{s=1}^{k} \omega_s f_j(x)},$$

де f_j – значення функції щільності розподілу;

g_{ij} – апостеріорна ймовірність того, що об'єкт x_i належить до j-ї компоненти суміші.

Щільність ймовірності двовимірного нормального розподілу має вигляд:

$$f(x) = \frac{1}{2\pi\sqrt{\det(C)}} \exp\left\{-\frac{1}{2}(x-\mu)^T C^{-1}(x-\mu)\right\},$$

$$C = (Cov_{ij}) = \begin{pmatrix} Cov_{11} & Cov_{12} \\ Cov_{21} & Cov_{22} \end{pmatrix},$$

де μ – вектор математичних очікувань;

C – коваріаційна матриця.

Або, у вираженні через дисперсії:

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times$$

$$\times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - \rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]\right\}.$$

Відповідно, функція логарифму максимальної правдоподібності приймає вигляд:

$$L(\theta; x, z) = \exp\left\{\sum_{i=1}^{n}\sum_{j=1}^{2} \log\omega_j - \frac{1}{2}\log\left|\sigma_j\right| - \frac{1}{2}(x_i-\mu_j)^T\sigma_j^{-1}(x_i-\mu_j) - \right.$$

$$\left. - - \frac{d}{2}\log(2\pi)\right\}$$

Отже, функція обчислення логарифму правдоподібності, що очікується, (Е-крок алгоритму) може бути записана як:

$$Q(\theta \mid \theta^{(t)}) = E[\log L(\theta; x, Z)] = E[\log \prod_{i=1}^{n} L(\theta; x_i, Z_i)] =$$

$$= E[\sum_{i=1}^{n} \log L(\theta; x_i, Z_i)] = \sum_{i=1}^{n} E[\log L(\theta; x_i, Z_i)] =$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{2} g_{ji}^{(t)}\left[\log\omega_j - \frac{1}{2}\log\left|\sigma_j\right| - \frac{1}{2}(x_i-\mu_j)^T\sigma_j^{-1}(x_i-\mu_j) - \frac{d}{2}\log(2\pi)\right].$$

На М-кроці вирішимо задачу максимізації значення параметра θ у функції логарифма правдоподібності і знайдемо наступне значення вектору Θ за поточними значеннями вектору прихованих змінних.

$$\omega^{(t+1)} = \arg\max_{\omega} Q(\theta \mid \theta^{(t)}) = \arg\max_{\omega} \left\{ \sum_{k=1}^{2} \left[\sum_{i=1}^{n} g_{k,i}^{(t)} \right] \log \omega_k \right\} =$$

$$= \frac{\sum_{i=1}^{n} g_{i,j}^{(t)}}{\sum_{k=1}^{2} \sum_{i=1}^{n} g_{k,i}^{(t)}} = \frac{1}{n} \sum_{i=1}^{n} g_{j,i}^{(t)},$$

$$(\mu_k^{(t+1)}, \sigma_k^{(t+1)}) = \arg\max_{\mu_k, \sigma_k} Q(\theta \mid \theta^{(t)}) =$$

$$= \arg\max_{\mu_k, \sigma_k} \sum_{i=1}^{n} g_{k,i}^{(t)} \left[\log \omega_j - \frac{1}{2} \log \left| \sigma_j \right| - \frac{1}{2} (x_i - \mu_k)^T \sigma_j^{-1} (x_i - \mu_k) \right],$$

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{n} g_{k,i}^{(t)} x_i}{\sum_{i=1}^{n} g_{k,i}^{(t)}},$$

$$\sigma_k^{(t+1)} = \frac{\sum_{i=1}^{n} g_{k,i}^{(t)} (x_i - \mu_k^{(t+1)})^T (x_i - \mu_k^{(t+1)})}{\sum_{i=1}^{n} g_{k,i}^{(t)}}.$$

Метод не є стійким до початкового наближення, і кінцевий результат залежить від даних, якими ініціалізується алгоритм. Під час рішення задачі ініціалізація проводилася двома типами даних:

1) результатом кластеризації вибірки, наприклад за методом *k*-середніх;

2) заданням початкового наближення в ручному режимі.

Результат роботи алгоритму наведений на рис. 5. Незалежно від методів ініціалізації алгоритм виявився не чутливим до спаду значень щільності розподілу біля нульових значень по осі *Y*.

Слід врахувати, що під час вирішення задачі відтворення щільності розподілу оцінка міри якості апроксимації з використанням, наприклад, середньоквадратичної помилки далеко не завжди є прийнятною. Важливішим є відповідність результуючої функції властивостям вхідних даних, таким як: кількість мод, асиметрії, поведінка хвостів. У зв'язку з цим, зазвичай доцільним є

перехід до непараметричних методів відтворення щільності розподілу.

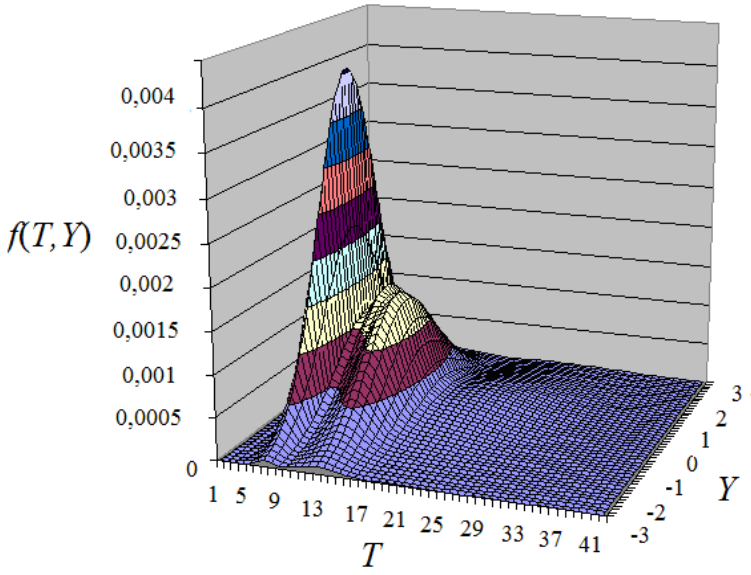

Рисунок 5 – Результат аппроксимації розподілу сумою гаусіан за допомогою ЕМ-алгоритму

Задача непараметричного відтворення щільності розподілу за гістограмою полягає в побудові апроксимуючої функції з прямокутної сітці значень з відтворенням проміжних значень. В якості апроксимуючої функції були обрані B-сплайни. Вибір пояснюється наступними їх властивостями:

1) гладкість одержуваної на виході поверхні відповідає природним формам функцій розподілу;

2) B-сплайни швидкі в побудові, а їх властивість локальності дозволяє з меншими витратами оновлювати розподіл при отриманні нових даних про розладнання .

Розділимо графік розподілу функції (T, Y) на прямокутні ділянки розміром dT на dY. Порахуємо кількість попадань в кожну ділянку і на підставі отриманих значень побудуємо гістограму. Точки гістограми будуть служити вузлами сплайна. Отримуємо двовимірний масив вузлових точок величиною $m \times n$. Двовимірну B-сплайн функцію запишемо у вигляді тензорного добутку одновимірних B-сплайнів, побудованих на кожній з двох осей координат. У такому випадку значення B-сплайн поверхні в довільній точці буде рівним [1]:

$$f(T,Y) = \sum_{k1=0}^{n} \sum_{k2=0}^{m} P_{k1,k2} B_{k1}(Y) B_{k2}(T),$$

$$B_{k,1} = \begin{cases} 1, & t_k < t < t_{k+1} \\ 0, & \text{інакше} \end{cases},$$

$$B_{k,d} = \left(\frac{u - u_k}{u_{k+d-1} - u_k} \right) B_{k,d-1}(u) + \left(\frac{u_{k+d} - u}{u_{k+d} - u_{k+1}} \right) B_{k+1,d-1}(u),$$

де Y, T – координати точки;

$P_{k1,k2} B_{k1}$ – значення гістограми в точці $k1$, $k2$;

$B_{k,d}(u)$ – значення стикувальної функції порядку d в точці u, що належить до інтервалу $(u_k ; u_{k+1})$;

u – вузловий вектор сплайну.

При рішенні задачі використовувалися кубічні стикувальні функції і відкриті однорідні вузлові вектори

$$U=[-3,-2,-1,0,1,..,n+1],$$
$$V=[-3,-2,-1,0,1,..,m+1].$$

Типовий результат відтворення щільності розподілу (T, Y) та маргінальні щільності розподілів T і Y наведені на рис. 6 та 7.

Отримана щільність є щільністю розподілу моментів розвороту тренду в залежності від прирощення значення курсу валют з моменту минулого розвороту і часу. Знаючи цю щільність, а також поточне (з моменту останнього зафіксованого розладнання до теперішнього моменту) значення (T, Y), шляхом інтегрування функції щільності можемо обчислити функцію ризику, що виражає ймовірність розвороту тенденції на наступному кроці спостережень .

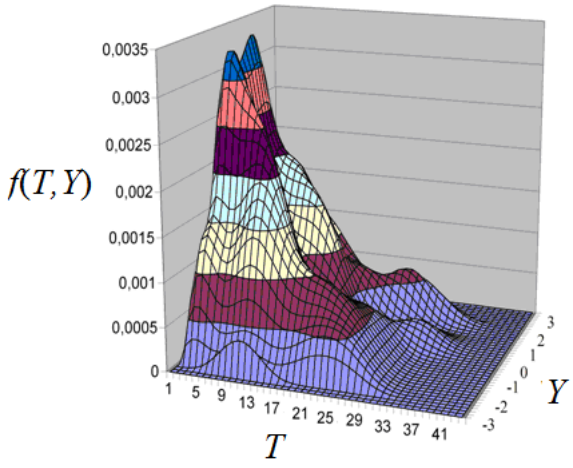

Рисунок 6 – Відтворена функція щільності розподілу

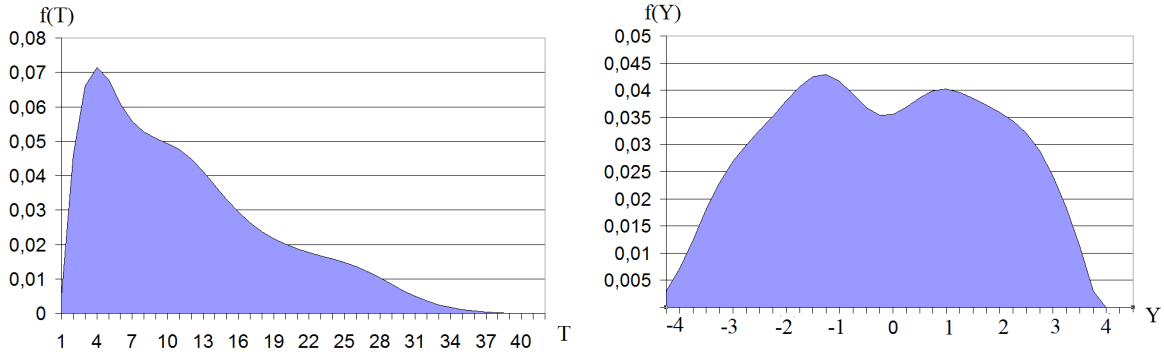

Рисунок 7 – Діаграми щільності маргінальних розподілів $f(T)$ і $f(Y)$

Отримана функція щільності розподілу не є стаціонарною і змінює своє значення у часі (рис. 8). Для забезпечення адекватності прогнозу в обчислення функції щільності бере участь обмежена кількість подій розладнання. Оновлення функції щільності відбувається щоразу у момент виявлення нового розладнання.

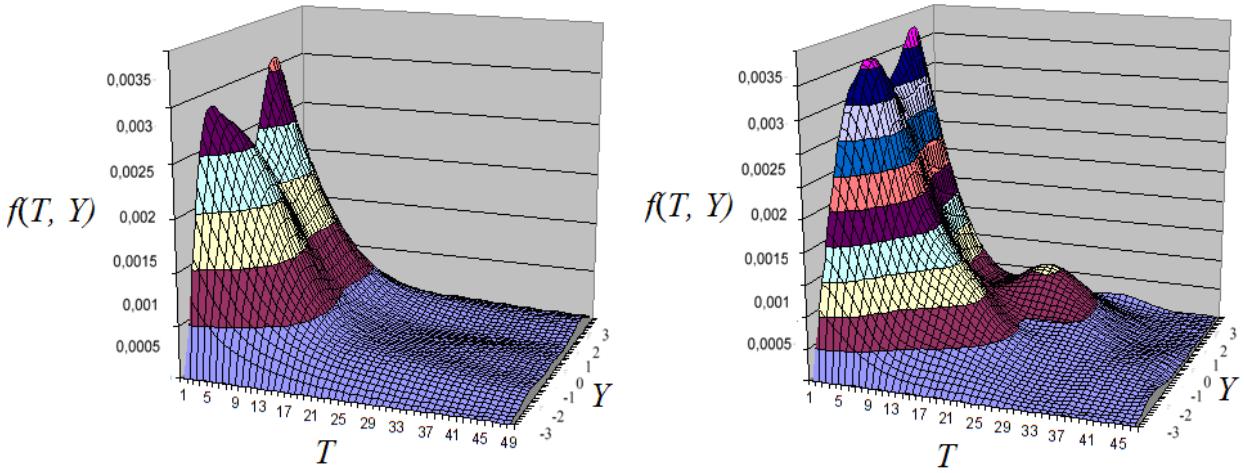

Рисунок 8 – Зміна функції щільності розподілу з часом

5 Обчислювальні схеми розрахунку функцій ризиків торгівлі, засновані на пропонованій моделі

Таким чином, з експериментальних даних було отримане чисельне вираження функції щільності розподілу $f(T,Y)$, що дає можливість обчислювати функцію ризику події розладнання як ймовірність, що змінюється у часі.

Ризик розладнання визначимо як відношення ймовірності F_1

розладнання на наступному кроку спостережень до ймовірності F_2 виникнення розладнання на всіх інших інтервалах, до яких величина, що спостерігається, може належати на наступних кроках спостереження з урахуванням не спадаючої природи T.

$$R_{cp1}(t'_{ij}) = \frac{F_1(t'_{ij})}{F_2(t'_{ij})}, \tag{16}$$

$$F_1(t'_{ij}) = \int_{t'_{ij}}^{t'_{(i+h)j}} f(T)dT \ ,$$

$$F_2(t'_{ij}) = \int_{t'_{ij}}^{\infty} f(T)dT \ ,$$

де T – вісь часу, що пройшов з останнього розладнання до поточного моменту часу. Враховуючи, що в кожний момент часу t_i цю величину можна розрахувати шляхом віднімання від поточного моменту останнього моменту розладнання τ_j, для конкретного моменту часу її можна представити у формі функції $t'_{ij} = t_i - \tau_j$. Надалі з метою розмежування понять «поточний час», «час з моменту розладнання як елемент векторної випадкової величини» і «конкретний час з моменту розладнання, взятий у заданий момент поточного часу» будемо використовувати вказані позначення;

h – горизонт пронозування;

$f(T) = \int_{-\infty}^{\infty} f(T,Y)dY$ – відтворена щільність розподілу розладнань на часовому ряді.

Отримана ймовірність є ймовірністю розладнання на інтервалі $(t'_{ij}\,; t'_{(i+h)j})$, але без урахування ймовірнісної залежності між T і Y, так, як якщо б прирощення Δy по осі Y мали рівномірний розподіл і будь-яке значення котирування на наступному кроці спостережень було б рівноймовірним (рис. 9).

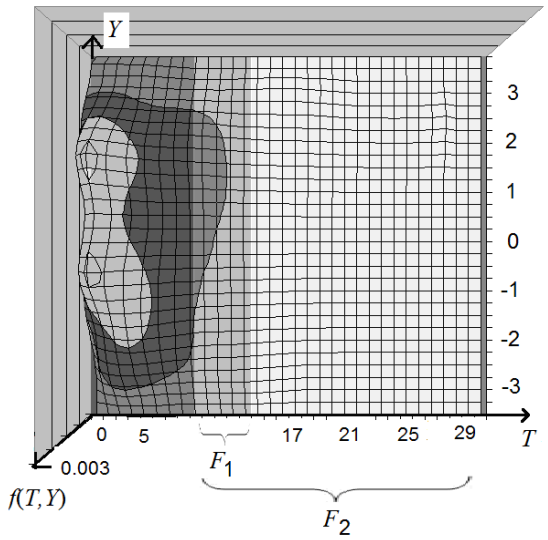

Рисунок 9 – Геометричний зміст формул (16)

Спостереження за курсами валют на різних історичних ділянках показує, що припущення про рівномірність розподілу прирощень цін Δy не відповідає дійсності, а між часом T і різницею цін Y існує стохастична залежність, що підтверджується перевіркою гіпотез із застосуванням критеріїв χ^2 і Фішера. Отже, необхідне уточнення моделі з відображенням у ній стохастичної залежності між T та Y. Це може бути досягнуте шляхом введення в (16) наступної умови:

$$R_{cp2}(t'_{ij}, Y_c) = \frac{F_1(t'_{ij}, Y_c)}{F_2(t'_{ij}, Y_c)} \qquad (17)$$

$$F_1(t'_{ij}, Y_c) = \int_{t'_{ij}}^{t'_{(i+h)j}} f(T \mid Y = Y_c)\, dT,$$

$$F_2(t'_{ij}, Y_c) = \int_{t'_{ij}}^{\infty} f(T \mid Y = Y_c)\, dT,$$

де Y_c – поточне значення приросту ціни валютної пари, рахуючи від останнього моменту розладнання.

Геометричний зміст формул (17) наведений на рисунку 10.

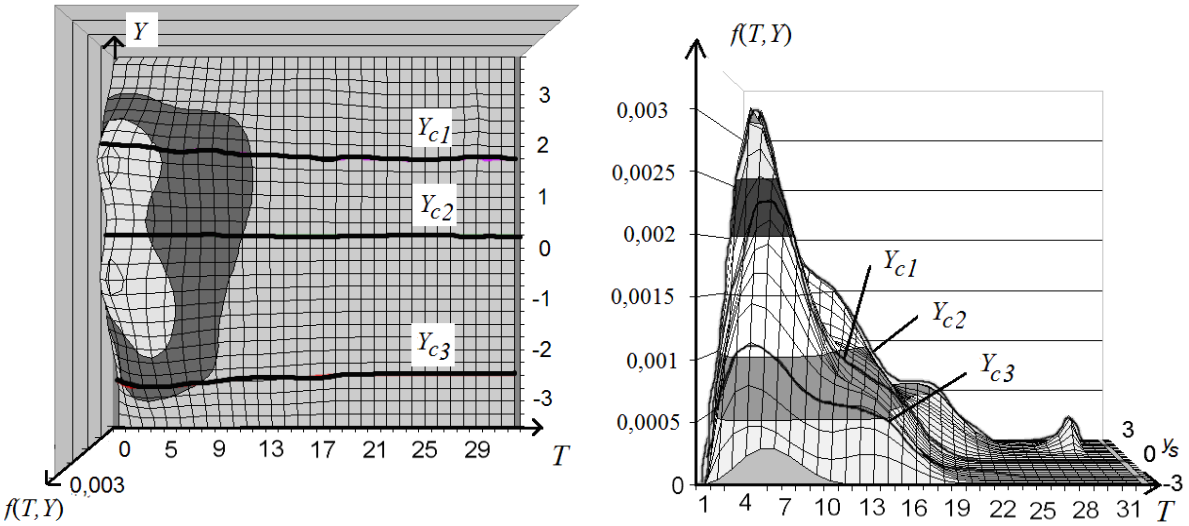

Рисунок 10 – Геометричний зміст формул (17): Y_{c1}, Y_{c2}, Y_{c3} – різні можливі значення Y_c , вид зверху і вид збоку

Таким чином, розглядаємо щільність не маргінального розподілу по T, а умовного розподілу при фіксованому значенні Y_c. Але і така модель не позбавлена недоліку: значення від розладнання до разладнання залишається порівняно постійним не для будь-якого процесу. З цих міркувань, запропоновано наступне уточнення до моделі: замість постійного значення ціни Y_c введено функцію $Y_c = Y(\Delta y, t)$, де Δy – прирощення ціни на одиничному інтервалі часу:

$$R_{cp3}(t'_{ij}, Y_c, \Delta y) = \frac{F_1(t'_{ij}, Y_c, \Delta y)}{F_2(t'_{ij}, Y_c, \Delta y)}, \tag{18}$$

$$F_1(t'_{ij}, Y_c, \Delta y) = \int_{t'_{ij}}^{t'_{(i+h)j}} f(T \mid Y = Y(\Delta y, t_i))\, dT,$$

$$F_2(t'_{ij}, Y_c, \Delta y) = \int_{t'_{ij}}^{\infty} f(T \mid Y = Y(\Delta y, t_i))\, dT.$$

При цьому для збереження фізичного змісту ймовірності, додатково введемо енергетичну міру:

$$\int_{-\infty}^{\infty} f(T \mid Y = Y(\Delta y, t_i))\, dT = 1. \tag{19}$$

Геометричний зміст функції (18) зображений на рис 11.

Функція $Y(\Delta y, t_i)$ не є точно детермінованою через

ймовірнісний характер параметру Δy, а отже її значення в майбутньому у момент часу t можна оцінювати лише з імовірнісної точки зору (рис. 12). Ймовірності справдження різноманітних сценаріїв щодо значення $Y(\Delta y, t_i)$ напряму залежать від ймовірностей набуття того чи іншого значення параметром Δy, звідси, оцінивши щільність розподілу одиничних прирощень цін Δy на певному інтервалі часу в минулому, отримаємо набір значень функції $Y(\Delta y, t_i)$ для всієї неперервної області визначення параметру Δy.

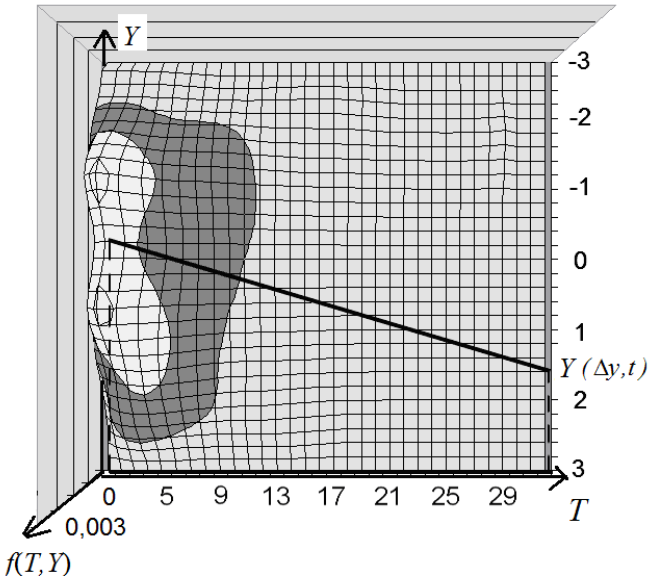

Рисунок 11 – Геометричний зміст формул (18): функція $Y(\Delta y, t_i)$ і результат введення енергетичної міри (19) – проекція $Y(\Delta y, t_i)$ на площину $(T, f(T, Y))$

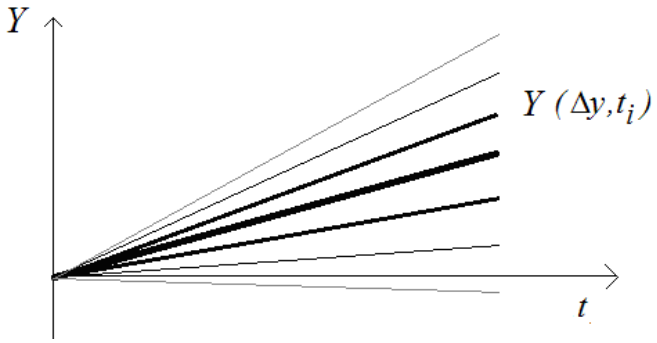

Рисунок 12 – Приклад функцій $Y(\Delta y, t_i)$ при зростаючому тренді (чим товща лінія, тим більше ймовірність сценарію)

Щоб екстраполювати значення отриманого набору функцій на

інтервали часу більше Δt, приймемо припущення, що функція щільності розподілу одиничних прирощень Δy не змінюється на протязі дії одного тренду, тобто від розладнання до розладнання (це припущення не вступає у протиріччя з визначенням тренда). Сума ймовірностей набуття функцією $Y(\Delta y,t)$ значень по кожному значенню осі T дорівнює 1. Геометричний зміст сказаного наведений на рис. 13.

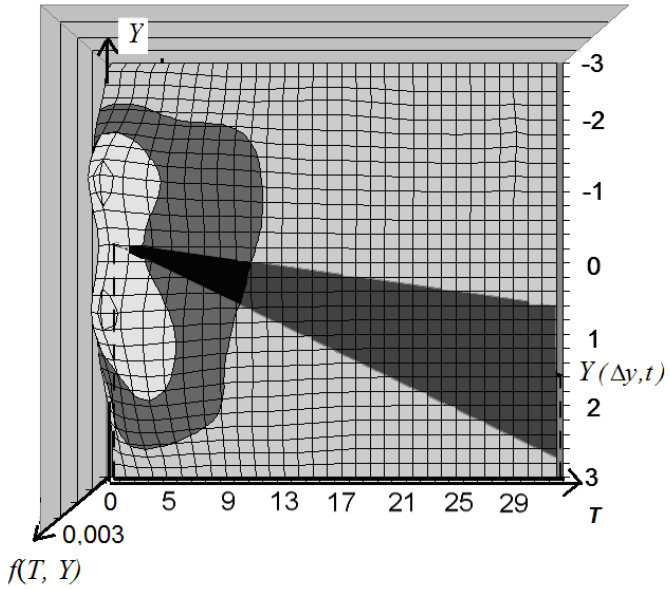

Рисунок 13 – Геометричний зміст формул (18) при ймовірнісному характері $Y(\Delta y, t_i)$

Підсумкове значення $Y(\Delta y, t_i)$, яке використовуватимемо при розрахунках, визначимо як математичне очікування $Y(\Delta y, t)$:

$$M[Y] = \int_{-\infty}^{\infty} Y(\Delta y, t_i) f(\Delta y) d\Delta y,$$

де $f(\Delta y)$ – щільність ймовірності прийняття функцією $Y(\Delta y, t)$ певного значення.

Звідси, маємо:

$$R_{cp4}(t'_{ij}, \Delta y) = \frac{F_1(t'_{ij}, \Delta y)}{F_2(t'_{ij}, \Delta y)}, \qquad (19)$$

$$F_1(t', \Delta y) = \int_{t'_{ij}}^{t'_{(i+h)j}} f(T \mid Y = \int_{-\infty}^{\infty} Y(\Delta y, t_i) f(\Delta y) d\Delta y)\, dT,$$

$$F_2(t', \Delta y) = \int_{t'}^{\infty} f(T \mid Y = \int_{-\infty}^{\infty} Y(\Delta y, t_i) f(\Delta y) d\Delta y)\, dT\,.$$

На рис. 14 наведено типовий вигляд залежності значення функції (19) від горизонту прогнозування (а) та від часу при фіксованому горизонті прогнозування (б).

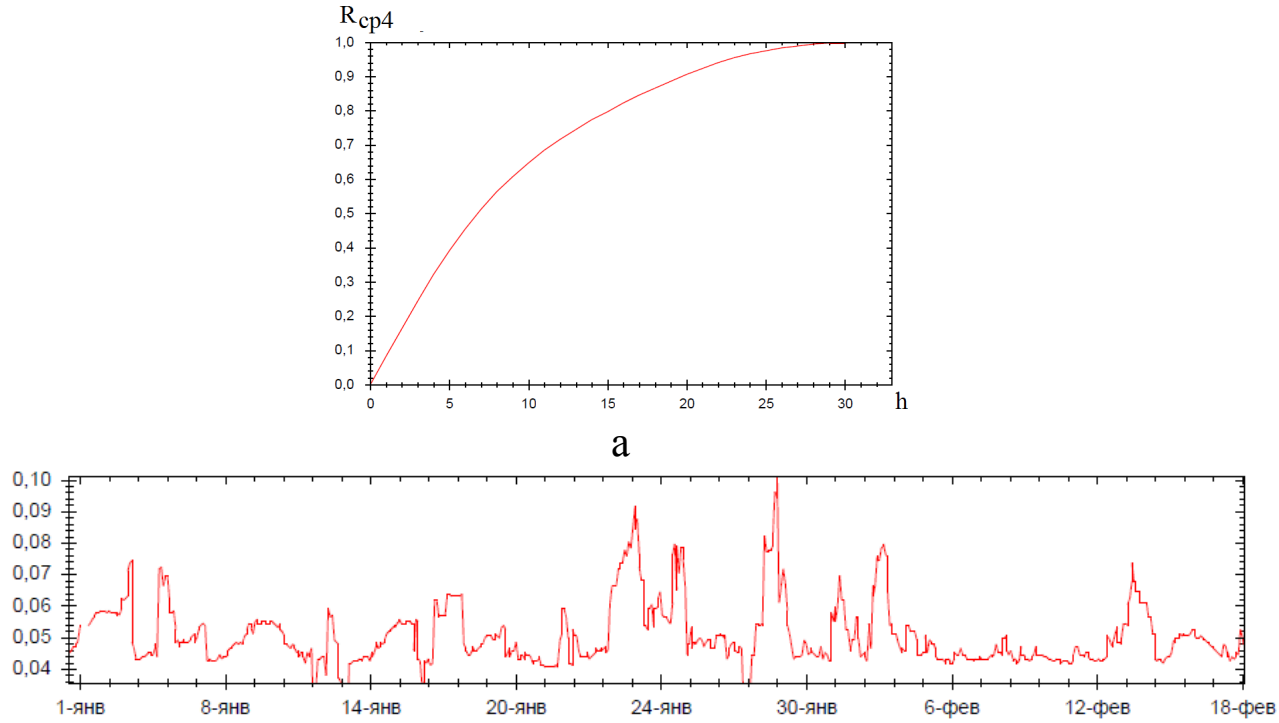

Рисунок 14 – Приклад графіку функції ризику розладнання (4.5), розрахованої на годинних інтервалах: а – залежність від горизонту прогнозування, б – залежність від часу при фіксованому горизонті прогнозування

Для окремих задач має значення не тільки ризик розладнання, але також інтервал, до якого в момент розладнання належатиме величина, що спостерігається. В цьому випадку, можливі інші інтегральні функції, зокрема, ймовірність потрапляння значення ряду при розладнанні до інтервалу $[a;b]$:

$$R_{cpa}(t'_{ij}, a, b) = \int_{t'}^{\infty} \int_{a}^{b} f(T, Y) dY\, dT\,,$$

а також ризик не перевищення певного значення при направленому тренді:

$$R_{ycp}(t'_{ij}, Y_c, \Delta y, Y_{cp}) = \frac{F_1(t'_{ij}, Y_c, \Delta y, Y_{cp})}{F_2(t'_{ij}, Y_c, \Delta y, Y_{cp})},$$

$$F_1\ (t'_{ij}\ , Y_c, \Delta y, Y_{cp}) = \begin{cases} \int\limits_{t'_{ij}}^{\infty} \int\limits_{Y_c}^{Y_{cp}} f(T,Y)dY\ dT, & Y_{cp} \geq Y_c \\ \int\limits_{t'_{ij}}^{\infty} \int\limits_{Y_{cp}}^{Y_c} f(T,Y)dY\ dT, & Y_{cp} < Y_c \end{cases},$$

$$F_2\ (t'_{ij}\ , Y_c, \Delta y, Y_{cp}) = \begin{cases} \int\limits_{t'_{ij}}^{\infty} \int\limits_{Y_c}^{\infty} f(T,Y)dY\ dT, & Y_{cp} \geq Y_c \\ \int\limits_{t'_{ij}}^{\infty} \int\limits_{-\infty}^{Y_c} f(T,Y)dY\ dT, & Y_{cp} < Y_c \end{cases},$$

де Y_{cp} – значення приросту ряду в момент прогнозованого розладнання.

Перелік посилань

1. Boor, Carl de. A Practical Guide to Splines / Carl de Boor. – Springer-Verlag, 2001 – 248 p.
2. Golyandina N. Analysis of Time Series Structure: SSA and Related Techniques / N. Golyandina, V. Nekrutkin, A. Zhigljavsky - Chapman and Hall/CRC, 2011- 320 p.
3. Никифоров, И.В. Последовательное обнаружение изменения свойств временных рядов / И.В.Никифоров. – М.: Наука, 1983. – 199 с.
4. Girshich, M.A. Bayes Approach to a Quality Control Model / M.A. Girshich, H. A. Rubin // Ann. Math. Statist. – 1952. – Vol. 23, N 1. – P. 114–125.
5. Page, E.S. Control charts for the mean of a normal population / E.S. Page // J. Roy Statist. Soc. B. – 1954. – Vol. 16, N 1. – P.131–135.
6. Lorden, G. On Excess over the boundary [Текст] / G. Lorden // Ann. Math. Statist. – 1970. – Vol. 41, N 2. – P.520–527.
7. Shewart, W.A. Economic Control of Quality of Manufactured Product [Текст] / W.A. Shewhart. – Seattle: Quality Press, 1980. – 501p.
8. Roberts, S.W. Control chart tests based on geometric moving average [Текст] / S.W. Roberts // Technometrics. – 1959. – Vol. 1, N 3. – P. 239–250.
9. Калишев, О.Н. Метод диагностирования измерительных каналов с учетом предыстории [Текст] / О.Н. Калишев // Автоматика и телемеханика. – 1988. – №6. – С. 135–143.
10. Johnson, N.L. A simple theoretical approach to cumulative sum control charts [Текст] /N.L. Johnson//J.Amer. Statist. Assoc. – 1961. – Vol. 56, N 296. – P. 835–840.
11. Nadler, J. Some characteristics of Page's twosided procedure for detecting a change in a location parameter [Текст] / J. Naddler, N.B. Robbins // Ann. Math. Statist. – 1971. – Vol. 42, N 2. – P. 538–551.
12. Hinkley, D.V. Inference about the change-point in a sequence of random variables [Текст] / D.V. Hinkley // Biometrika. – 1970. – Vol. 57, N 1, – P. 1–17.
13. Brodsky, B.E. Nonparametric Methods in Change-Point Problems [Текст] / B.E. Brodsky, B.S. Darkhovsky. – Dordrecht: Kluver Academic Publishings, 1993. – 210 p.

14. Воробейчиков, С.Э. Об обнаружении изменения среднего в последовательности случайных величин [Текст] /С.Э.Воробейчиков // Автоматика и телемеханика. – 1998. – №3. – С. 50–56.
15. Бородкин, Л.И. Алгоритм обнаружения моментов изменения параметров уравнения случайного процесса [Текст] / Л.И. Бородкин, В.В. Моттль //Автоматика и телемеханика. – 1976. – №6. – С. 23–32.
16. Гаджиев, Ч.М. Проверка обобщенной дисперсии обновляющей последовательности фильтра Калмана в задачах динамического диагностирования [Текст] / Ч.М. Гаджиев // Автоматика и телемеханика. – 1994. – №8. – С. 98–104.
17. Луценко О. П. Інформаційна система аналізу функцій ризику розладнання в процесі фінансової торгівлі / О. П. Лу¬цен¬ко, О. Г. Байбуз // Актуальні проблеми автоматизації та інформацій¬них технологій. — Д. : Ліра, 2014. — Т.18. — С. 42—51.
18. Колмогоров, А.Н. Основные понятия теории вероятностей [Текст] / А.Н.Колмогоров. – М.: Наука, 1974. – 120 с.