Метод Гусениця-SSA

Виконав:

Студент групи СНм-51

Стодола Володимир

Тернопіль,2010

Гусениця SSA

SSA *(Singular spectrum analysis)* - метод аналізу часових рядів, заснований на перетворенні одновимірного часового ряду в багатовимірний ряд і подальшого застосування до отриманого багатовимірному тимчасовому ряду методу головних компонент. Спосіб перетворення одновимірного ряду в багатовимірний представляє собою «згортку» тимчасового в матрицю, що містить фрагменти тимчасового ряду, отримані з деяким зрушенням. Загальний вигляд сдвиговой процедури нагадує «гусеницю», тому сам метод нерідко так і називають - «Гусениця»: довжина фрагмента називається довжиною «гусениці », а величина зсуву одного фрагмента щодо іншого кроком «гусениці ».

Історія

- Broomhead і King (1986) пропонують використовувати SSA і M-SSA в контексті нелінійної динаміки в цілях відновлення атрактор системи з виміряних часових рядів.
- Ghil, Vautard і співробітники (Vautard і Ghil, 1989; Ghil і Vautard, 1991;. Vautard та ін, 1992) зауважив, аналогію між траєкторією матриця Broomhead і King, з одного боку, і Karhunen (1946)-Loève (1945) аналіз головних компонент у домені часу, з іншого. Таким чином, SSA може бути використаний як метод області часу і частоти для аналізу часових рядів - незалежно від аттрактора реконструкції і в тому числі випадків, в яких останній може дати збій.
- В даний час роботи, присвячені методологічним аспектам та застосування SSA обчислюються сотнями. Багато літератури надаються Elsner and Tsonis (1996), Danilov and Zhigljavsky (1997), Golyandina et al. (2001), and Ghil et al. (2002).

Актуальнісь використання SSA

- В даний час актуальним є аналіз і прогнозування товарних і фінансових ринків з використанням методів математичної статистики.
- Традиційні підходи, засновані на використанні класичних моделей типу "тренд + шум» або «авторегресії - ковзного середнього» призводять до задовільних результатів лише для рядів досить простої структури
- Особливістю тимчасових рядів, що відображають поведінку ринку, є те, що їх характеристики (ціни, обсяги угод, індикатори і т.д.) формуються з декількох складових: повільної - трендом, періодичної чи коливальної складової і випадкової складової описуваної випадковим процесом певного типу.
- Важливою особливістю періодичної складової, у свою чергу, є наявність періодичності зі змінним періодом і амплітудою.
- З причини розглянутих особливостей для дослідження фінансових ринків погано застосовні класичні методи аналізу, такі як аналіз Фур'є, регресійний аналіз чи вейвлет-аналіз, тому що вони використовують розкладання вихідної функції в ряд по фіксованій системі базисних функцій, що породжує властивість строгої періодичності.

Альтернативним підходом, використовуваним для аналізу та прогнозу ринків, є *Сингулярний Спектральний Аналіз SSA* (Singular Spectrum Analysis), заснований на динамічній модифікації методу головних компонент. Даний підхід заснований на дослідженні тимчасового ряду методом головних компонент і не вимагає попередньої стабілізації ряду. SSA дозволяє досліджувати структуру часового ряду, виділити окремі його складові та прогнозувати як сам ряд, так і тенденції розвитку його складових.

Особливостями методу є такі його властивості, як

- інтерактивність ;
- візуалізація результатів обчислень.

Ідеї створення:

- Першою ідеєю, що лежить в основі методу, є створення повторюваності шляхом переходу від тимчасового ряду, наприклад послідовності цін у рівновіддалені моменти часу, до послідовності векторів, що складаються з відрізків тимчасового ряду обраної довжини. Таким чином, виходить багатовимірна вибірка, іншими словами, мається на увазі, що якщо вихідний ряд мав якусь структуру, то і відрізки успадковують цю структуру.
- Другою ідеєю є аналіз отриманої багатовимірної вибірки за допомогою її сингулярного розкладання або, використовуючи статистичні аналогії, аналізу головних компонент, виділення значущих компонент і подальшому відновленні, заснованому на угрупованню і діагональному усередненні. Тим самим виходить розкладання вихідного часового ряду (його траекторної матриці) по базису, породжуваному їм самим.

- Перевагою методу «Гусениця»-SSA є відсутність вимоги апріорного завдання моделі ряду, а також можливість виділення гармонійних складових до мінливих амплітудами і частотами, що вигідно відрізняє його від методів, в основі яких лежить метод Фур'є.
- Недоліком методу, що обмежує можливості його застосування, є припущення про лінійність моделі досліджуваного ряду. На перший план висувається завдання вибору достатньо універсальної моделі часового ряду, що дозволяє відобразити суттєві особливості його нелінійної динаміки, найчастіше носить хаотичний характер. Для вирішення подібних завдань ефективні методи, засновані на ядерних методах (kernel methods), що забезпечують можливість моделювання нелінійних зв'язків у фінансових часових рядах при порівняно малому обсязі апріорної інформації.

модифікація методу для аналізу рядів з пропусками

Нехай вихідний часовий ряд

$$F_N = (f_0 \dots f_{N-1})$$

складається з N елементів, частина яких невідома. Опишемо схему алгоритму для випадку відновлення першої складової ряду на $F_N^{(1)}$ові суми двох:

$$F_N = F_N^{(1)} + F_N^{(2)}$$

Перший етап: **розкладання**

1. **Вкладення.** Зафіксуємо довжину вікна L:1<L<N. Процедура вкладення переводить вихідний тимчасовий ряд у послідовність L - вимірних векторів вкладення $\{X_i\}_{i=1}^{K}$ де K = N- L +1. Частина векторів вкладення може мати пропуски. З векторів вкладення без пропусків $X_i, i \in C$ утворюємо матрицю X`ка при відсутності перепусток збігається з траекторной матрицею ряду F_N

2. **Знаходження базису.** Нехай $S = X` \cdot X`^T$

$\lambda_{1,} \dots \lambda_L$ - власні числа матраци, взяті в незростаючими порядку $\lambda_{1,} \geq \cdots \lambda_L \geq 0$

$U_{1,} \dots U_L$ - ортонормованна система власних векторів матриці $S`$

відповідних власним числам, d = max $\lambda_{i,}$ > 0}.

Задамо два вектори

$A = (a_1 \dots a_n)^T$ $B = (b_1 \dots b_n)^T$

Якщо ввести операцію "*" таким чином:

$$(\mathrm{A},\mathrm{B})^* = \mathrm{A}^{\mathrm{T}} * \mathrm{B} = \frac{\mathrm{n}}{\mathrm{n} - \left|\mathrm{A} \cup \mathrm{B}\right|} \sum_{\mathrm{k}:\mathrm{k} \notin \mathrm{A} \cup \mathrm{B}} \mathrm{a_k b_k},$$

то при множенні векторів без пропусків результат виконання операції збігається зі скалярним добутком, а для векторів з пропусками буде чисельно замінювати скалярний твір.

В якості матриці можна взяти матрицю $S`$

яка о$\tilde{\mathrm{S}} = \mathrm{X} * \mathrm{X}^{\mathrm{T}}$

де X - траєкторна матриця ряду F_N, яка містить пропуски.

Далі утворюємо матрицю $\tilde{\mathrm{X}}_{(\tau)}$ що складається з векторів вкладення, утримуючиих не більше τ пропущених компонент, і $\tilde{\mathrm{S}} = \tilde{\mathrm{X}}_{(\tau)} * \tilde{\mathrm{X}}_{(\tau)}^{\mathrm{T}}$

Другий етап: відновлення
3. Проекція векторів вкладення

На початку проводиться вибір підпростору
проекти векторів вкладення без пропусків

Вибирається набір номерів $I_r = \{i_1, \ldots, i_r\} \subset \{1, \ldots, d\}$
з допомогою яких утворюється
підпространство $M_r = Sp(U_{i_1}, \ldots, U_{i_r})$
відповідне виділеній компоненті.

відбувається проектування векторів вкладення без пропусків на вибраний підпростір M_r

$$\hat{X}_i = \sum_{k \in I_r} (X_i, U_k) U_k , \quad i \in C .$$

строється проекція векторів вкладення з пропусками Для кожного вектора вкладення з пропусками на місцях з P (своє для кожного вектора)

$X^{\sim} [X_1^{\sim} \ldots X_K^{\sim}]$ апроксимація траекторної матриці ряду при правильному виборі множини I_r.

4. Діагональне усереднення.

На останньому кроці базового алгоритму матриця $X^{`}$ переводиться в новий ряд $F_N^{\sim(1)}$ (відновлений ряд) за допомогою операції діагонального усереднення.

Задача

- 105 студентів
- 35 спроб здачі тестів

Вводимо позначення

- Ns - кількість студентів
- Na - кількість спроб для кожного студента

$$\{f_{ij}\}, i = \overline{0, N_s - 1;} \quad j = \overline{0, N_a - 1}$$

значення тимчасового ряду оцінок тестування знань для і-го студента в j-й спробі

Тимчасові ряди з пропущеними значеннями отримуємо за допомогою видалення з вихідного $\{f_{ij}\}$ ряду n значень, $n = \overline{1,\tau}$

τ поріг кількості пропущених компонент. У нашому випадку =15

сімейство тимчасових рядів

$$\{f_{ij}^{(n,k)}\}, i = \overline{0, N_s - 1}; j = \overline{0, N_a - n - 1}; n = \overline{1, \tau}; k = \overline{1, K_n},$$

K_n кількість варіантів видалення а значенm з вихідного часовогс $K_n = C_{N_a}^n$

$N_s K_n$ загальга кількість тимчасових рядів (для всіх студентів)

- У нашому випадку при n> 4 кількість варіантів перевищує 10^6
- У цьому випадку конкретні часові ряди генеруються випадковим чином за допомогою методу Монте-Карло.

  Таким чином

$$K_n = \begin{cases} C_{N_a}^n, n \le 4, \\ 10^4, n > 4. \end{cases}$$

- Застосовуємо модифікацію методу SSA для відновлення тимчасового ряду

  Отримуємо сімейство відновлених тимчасових рядів

$$\{g_{ij}^{(n,k)}\}, i = \overline{0, N_s - 1}; j = \overline{0, N_a - 1}; n = \overline{1, \tau}; k = \overline{1, K_n}.$$

- Математичне сподівання «помилки» алгоритму відновлення визначається за формулою:

$$\bar{y}^{(n)} = \frac{1}{N_s K_n} \sum_{i=1}^{N_s} \sum_{k=1}^{K_n} y_i^{(n,k)},$$

- а стандартне відхилення дорівнює

$$\delta^{(n)} = \sqrt{\frac{1}{N_s K_n - 1} \sum_{i=1}^{N_s} \sum_{k=1}^{K_n} (y_i^{(n,k)} - \bar{y}^{(n)})^2}.$$

статистичні результати імітаційного моделювання алгоритму відновлення тимчасового ряду з пропусками. Для конкретних значень кількості пропущених значень (від 1 до 15) визначалися довірчі інтервали «помилки» алгоритму з рівнем довіри 90%. Статистичний аналіз показав, що якщо число пропущених значень не перевищує семи, то «помилка» алгоритму відновлення не більше 20%.

Кількість пропущених значень	Середнє значення $(\overline{y}^{(n)})$	Стандартне відхилення $(\delta^{(n)})$	Нижня границя довірливого інтервала	Верхня границя довірливого інтервала
1	0,012	0,006	0,002	0,022
2	0,018	0,007	0,006	0,030
3	0,032	0,010	0,016	0,048
4	0,039	0,011	0,021	0,057
5	0,068	0,013	0,047	0,089
6	0,087	0,015	0,062	0,112
7	0,171	0,018	0,141	0,201
8	0,272	0,021	0,237	0,307
9	0,319	0,022	0,283	0,355
10	0,346	0,025	0,305	0,387
11	0,382	0,028	0,336	0,428
12	0,409	0,034	0,353	0,465
13	0,453	0,041	0,386	0,520
14	0,481	0,052	0,395	0,567
15	0,516	0,059	0,419	0,613

Це означає, що

- розбіжність між оцінками не перевищує один бал за п'ятибальною шкалою.
- при великих помилках тести вже не оцінюють адекватно знання студентів.

Тому при кількості пропущених значень більше семи, алгоритм SSA не можна використовувати для відновлення тимчасового ряду результатів тестування знань студентів.

Висновок

За порівняльним аналізом ефективності SSA з класичними методами Технічного Аналізу показує, що SSA підхід, принаймні, також хороший, а в багатьох випадках перевершує класичні засоби Технічного Аналізу. При цьому часто він дозволяє виявити ефекти, які розпізнати стандартними методами не представляється можливим.

Використанні джерела

1. http://www.spectraworks.com/Help/ssatheory.html
2. http://ru.wikipedia.org/wiki/SSA_%28%D0%BC%D0%B5%D1%82%D0%BE%D0%B4%29
3. http://en.wikipedia.org/wiki/Singular_spectrum_analysis#Brief_history
4. http://www.nbuv.gov.ua/portal/natural/soi/208_3/Bochar.pdf
5. http://www.nbuv.gov.ua/portal/natural/soi/2008_3/
6. www.AnalysisFX.com

Дякую за увагу