O metodach identyfikacji składowych okresowych w wielowymiarowych szeregach czasowych

Maciej Kawecki

Instytut Matematyki i Informatyki
Politechnika Wrocławska

„Statystyka Matematyczna Wisła 2009"

7-11 grudnia 2009

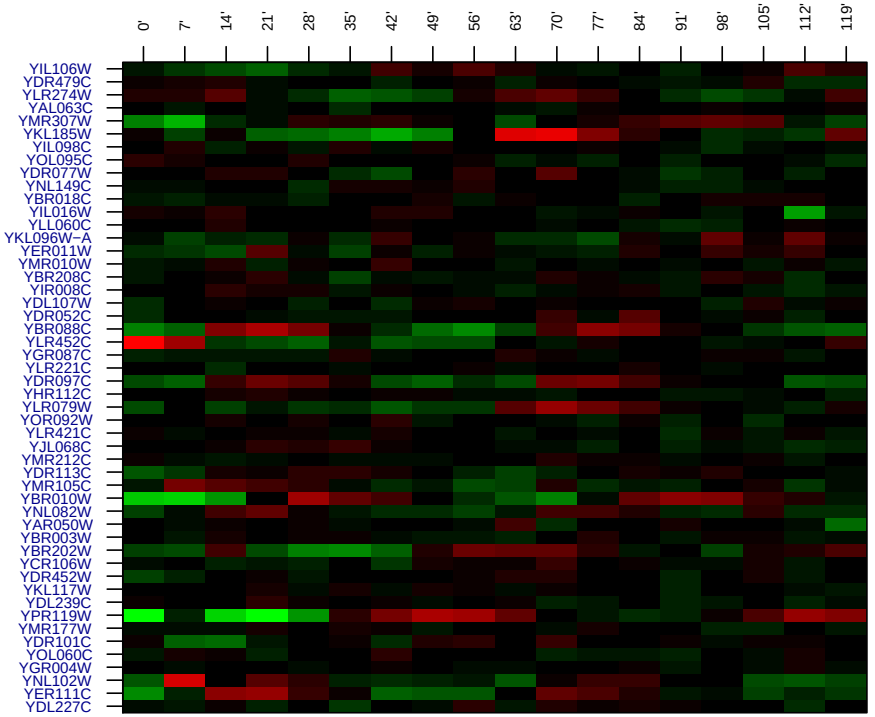
Plan prezentacji

1. Wprowadzenie - geneza i opis problemu
2. Przegląd stosowanych metod
3. Testy klasycznej analizy spektralnej wraz z modyfikacjami
4. Zastosowanie modelu RRRVAR
5. Porównanie efektywności testów
6. Analiza danych rzeczywistych

Geneza i opis problemu

Charakterystyka analizowanych szeregów czasowych

- Długość szeregu czasowego $T \leq 50$
- Liczba składowych $M > 400$
- Składowe silnie skorelowane
- Dwie grupy składowych: okresowe i nieokresowe



Rysunek: Przykładowa *heatmapa* ekspresji genów komórki drożdży (Spellman, 1998)

Zaproponowane podejście

Cel: klasyfikacja

Każdej składowej chcemy przyporządkować wartość:
0 – nieokresowy komponent
1 – okresowy komponent

Przeprowadzamy klasyfikację scoringową

$$Y = \left[\begin{array}{c} Y_1(t) \\ Y_2(t) \\ \vdots \\ Y_M(t) \end{array}\right] \xrightarrow[\text{scoringi p-wartości}]{\text{metoda}} \left\{\begin{array}{c} s_1 \\ s_2 \\ \vdots \\ s_M \end{array}\right\} \xrightarrow[\text{zbiór uczący testowanie wielokrotne}]{\text{klasyfikacja binarna}} \left\{\begin{array}{c} \{0,1\} \\ \{0,1\} \\ \vdots \\ \{0,1\} \end{array}\right\}$$

Jaką wiedzę posiadamy?

Ogólny model dla i-tej składowej

$$Y_i(t) = a_{i,0} + a_{i,1} \cdot f_i(t + \tau_i) + \varepsilon_i(t)$$

Dodatkowe założenia

funkcja okresowa	$f_i(t) = f(t)$
okres	$s_i = s$
wariancja błędów	$\sigma_i^2 = \sigma^2$
wyraz wolny	$a_{i,0} = 0$

Dodatkowe informacje a priori

- postać funkcji okresowej, jej częstotliwość
- wariancja błędów, struktura korelacji składowych
- zbiór uczący
  – podzbiór składowych, o których wiadamo czy są okresowe

Zestawienie metod

- jednowymiarowe metody niewykorzystujące wiedzy a-priori
  a) klasyczna analiza spektralna (Brockwell i in., Wichert i in., 2004)
- wielowymiarowe metody niewykorzystujące wiedzy a-priori
  a) dopasowanie modelu RRRVAR (Kustra, Zagdański, 2006)
  b) połączony test regularności i okresowości (Lichtenberg, 2005)
- metody wykorzystujące wiedzę a-priori
  a) klasyczna analiza spektralna
  b) podejście bayesowskie (Andersson i in., 2006)
  c) test korelacyjny (Spellman i in., 1998)
  d) interpolacja współdzielonej funkcji okresowej za pomocą B-splajnów (Luan, 2004)

TESTY KLASYCZNEJ ANALIZY SPEKTRALNEJ

Testy klasycznej analizy spektralnej

Idea

Model:

$$Y_i(t) = a_{i,0} + a_{i,1} \cdot f(t + \tau_i) + \varepsilon_i(t)$$

Weryfikowane hipotezy:

$$H_0 : a_{i,1} = 0 \qquad \text{(składowa nieokresowa)}$$
$$H_1 : a_{i,1} \neq 0 \qquad \text{(składowa okresowa)}$$

Sposób obliczania scoringów

Scoringi to wartości statystyk testowych

Testy klasycznej analizy spektralnej

Używane testy statystyczne
Brockwell, *Time Series: Theory and Methods*

G1 test Fishera ukrytej okresowości o nieokreślonej częstotliwości (G-test Fishera)

- $f(t) = \sin(\omega t + \tau_i)$
- $G_1 = \frac{\max_k I(\omega_k)}{\frac{1}{q_T}\sum_{k=1}^{q_T} I(\omega_k)}$

G2 test ukrytej okresowości o określonej częstotliwości ω_0

- $f(t) = \sin(\omega_0 t + \tau_i)$
- $G_2 = \frac{(n-3)I(\omega_0)}{\sum_{t=1}^{T} Y_i^2(t) - I(0) - 2I(\omega_0)}$

G3 test weryfikujący obecność niesinusoidalnego komponentu okresowego o znanym okresie p

- $f_i(t) = \sum_{k=1}^{\left[\frac{p-1}{2}\right]} \left[A_{i,k}\cos\left(\frac{2\pi kt}{p}\right) + B_{i,k}\sin\left(\frac{2\pi kt}{p}\right)\right] + A_{\frac{p}{2}}(-1)^t$
- $G_3 = \frac{(T-p)}{(p-1)} \frac{2\sum_{1\le j\le p/2} I(\omega_{rj}) + \delta_p I(\pi)}{\sum_{t=1}^{T} Y_i^2(t) - I(0) - \delta_p I(\pi) - 2\sum_{1\le j\le p/2} I(\omega_{rj})}$

Testy klasycznej analizy spektralnej

Stosowane wersje estymatorów gęstości spektralnej

- standardowy periodogram

$$I(\omega_k) = \sum_{|h|<T} \hat{\gamma}(h) \exp^{-i\omega_k h}$$

- odporny estymator gęstości spektralnej (Ahdesmaki i in., 2005)

$$I_{*1}(\omega) = \sum_{k=-T-2}^{t=T-2} \tilde{\rho}(k) \exp(-i\omega k)$$

- wygładzony dyskretnie periodogram

$$I_{*2}(\omega_j) = (2\pi)^{-1} \sum_{|k|\leq m} W(k) I(\omega_{j+k})$$

- wygładzony jądrowo periodogram

$$I_{*3}(\omega_j) = \hat{f}(\omega_j, h) = \frac{1}{Th} \sum_{k=-q_T}^{q_T} K\left(\frac{\omega_j - \omega_k}{h}\right) I(\omega_k)$$

Otrzymujemy 12 wersji testów klasycznej analizy spektralnej.

ZASTOSOWANIE MODELU RRRVAR

RRRVAR – podejście uwzględniające korelacje

Definition (Model RRVAR(p, r))

Szereg czasowy $\{Y_t\}$ jest szeregiem czasowym z modelu RRVAR(p, r) jeżeli:

$$Y_t = ABX_t + \epsilon_t \ \ (t = 1, \ldots, T)$$

gdzie:

- $X_t = [Y'_{t-1}, \ldots, Y'_{t-p}]'$ — *wektor kolumnowy wymiaru:* $Kp \times 1$
- A — *macierz współczynników wymiaru* $K \times r$
- $B = [B_1, B_2, \ldots, B_p]$ — *macierz wymiaru* $r \times Kp$
- B_i — *macierz współczynników wymiaru* $r \times K$
- ϵ_t — *wektor błędu o nieosobliwej macierzy kowariancji*

Zauważmy, że macierze współczynników z wyjściowego procesu VAR(p) wyrażają się przy pomocy współczynników RRVAR(p, r) w następujący sposób $\Phi_i = AB_i$.

RRRVAR – szczegóły techniczne

Dysponując parametrami p oraz r estymujemy współczynniki macierzy, kanoniczne wagi oraz szeregi czasowe.

Kanoniczne szeregi czasowe

Dla $j = 1, \ldots, r$ definiujemy j-ty kanoniczny szereg czasowy jako:

$$w_t^j = l_j' Y_t$$

Kanoniczne wagi

Definiujemy:

$$l_j = \Sigma_\epsilon^{-\frac{1}{2}} V_j',$$

gdzie V_j jest j-tym wektorem własnym macierzy:

$$W = \Sigma_{YY}^{-1/2} \Sigma_{XY}' \Sigma_{XX}^{-1} \Sigma_{XY} \Sigma_{YY}^{-1/2},$$

odpowiadzającym j-tej największej wartoiści własnej.
Do wyznaczenia estymatorów Σ stosuje się estymatory autokorelacji $\tilde{\Gamma}(0) = \hat{\Gamma}(0) + \lambda I_{M \times M}$.
Parametr λ nazywany jest współczynnikiem kary grzbietowej.

RRRVAR – odkrywanie ukrytej wiedzy

Szergi kanoniczne mogąbyć interpretowane jako najczęsciej występujące wzorce w wielowymiarowym szeregu czasowym (również okresowe), więc w celu znalezienia komponentów okresowych możemy przeprowadzić procedurę:

Procedura wyznaczenia scoringów

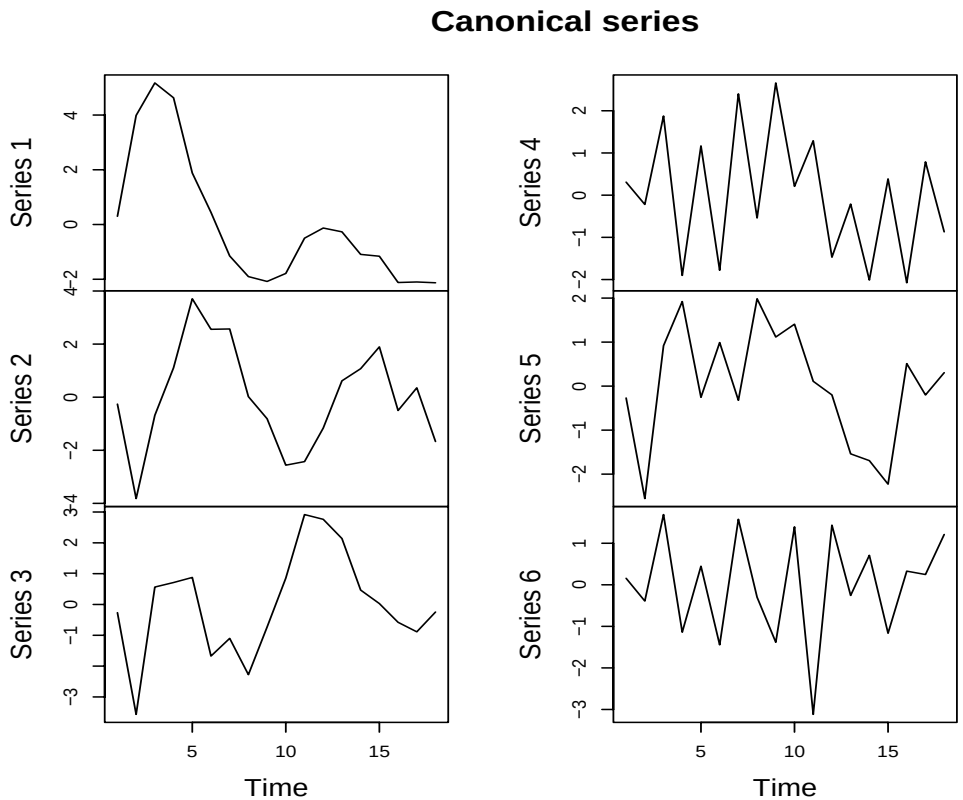
1. dopasowujemy odpowiednie parametry p oraz r modelu
2. znajdujemy współczynniki macierzy, szeregi oraz wagi kanoniczne
3. testujemy, które szeregi kanoniczne są okresowe
   (otrzymujemy r_1 szeregów)
4. wybieramy wagi kanoniczne dla tych szeregów
   (otrzymujemy macierz $M \times r_1$ $(r_1 \leq r)$)
5. dla każdego wiersza obliczamy maksimum z wartości bezwzględnych wybranych wag
   w ten sposób uzyskujemy „scoring okresowości" dla każdego szeregu

Cykl podziału komórki drożdży - eksperyment alpha

- mamy 6178 genów oraz 18 obserwacji dla każdego szeregu.
- znamy 104 nazwy genów okresowych.
- Spellman zidentyfikował 800 okresowych genów.

Dopasowaliśmy model RRRVAR(2,6) używając kryteriów wyboru modelu (CV)

Otrzymujemy następujące szeregi czasowe.



SYMULACYJNE PORÓWNANIE METOD

Jak porównać efektywność metod scoringowych? uwagi ogólne

Czy dysponujemy zbiorem uczącym?

1. tak
   - pole pod krzywymi ROC (AUC)
   - błąd klasyfikacji
     obliczony metodą k-fold cross validation
2. nie
   - metody porównujące rangi
   - klasyfikacja '*TOP N*' – diagramy Venna
     porównanie pierwszych N najwyżej sklasyfikowanych składowych

Jak porównać efektywność metod scoringowych? scenariusz symulacyjny

Ogólny model symulacyjny

$$Y(t) = [\underbrace{Y_{1,1}(t), \ldots, Y_{1,M_1}(t)}, \ldots, \underbrace{Y_{i,1}(t), \ldots, Y_{i,M_i}(t)}, \ldots, \underbrace{Y_{k,1}(t), \ldots, Y_{k,M_k}(t)}]$$

$$Y_{k,i}(t) = \mu_{k,i} + \alpha_k f_k(t - \tau_{k,i}) + \varepsilon_{k,i}(t)$$

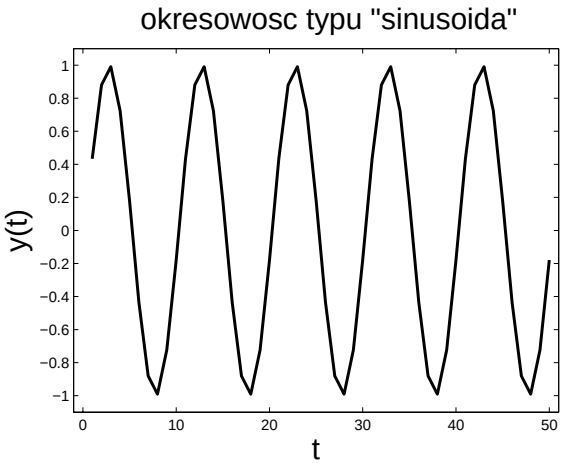
Parametry modelu symulacyjnego

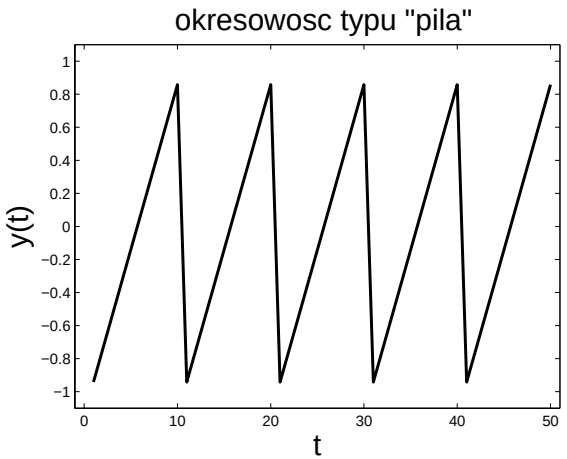
- długość szeregu czasowego i liczba składowych
- macierz kowariancji białego szumu
- postać funkcji okresowej i częstotliwość
- liczność frakcji składowych okresowych

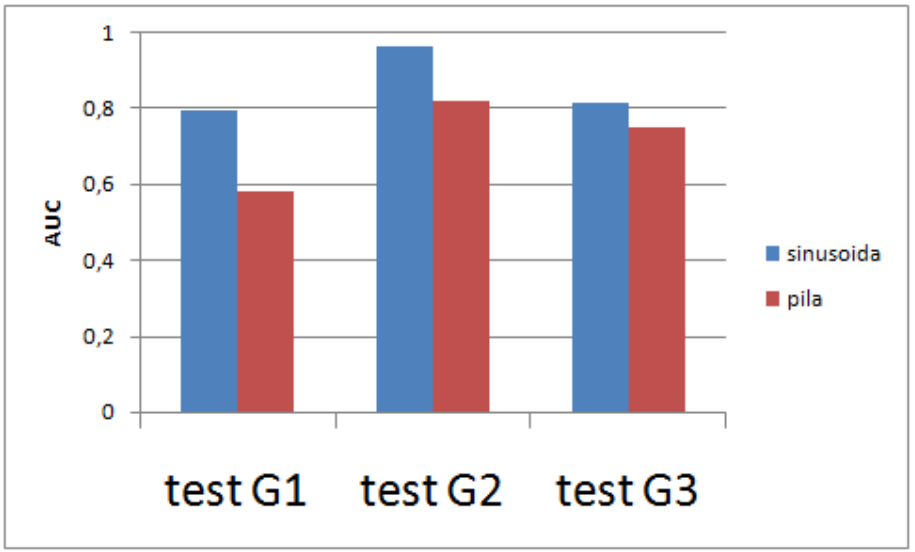
Porównanie testów - wpływ typu funkcji okresowej

wyniki testów dla różnych funkcji okresowych

rodzaj funkcji okresowej ma wpływ na efektywność metody



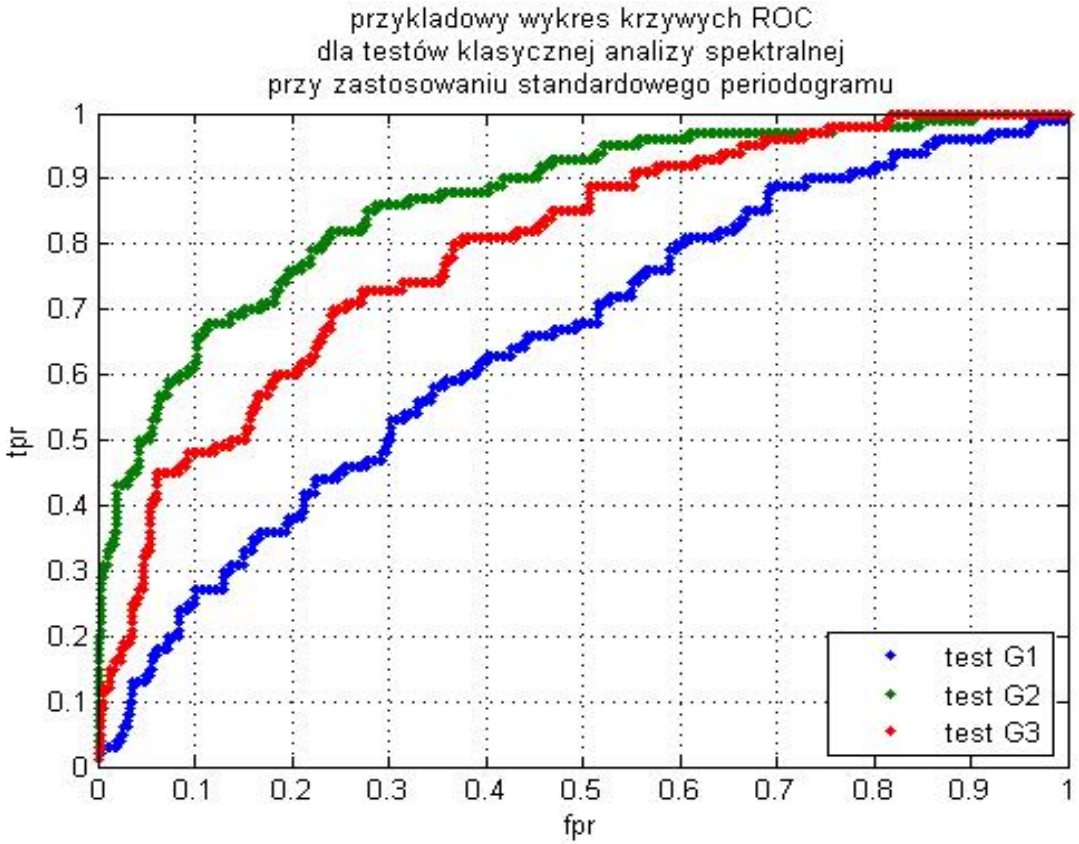




Porównanie testów - krzywe ROC

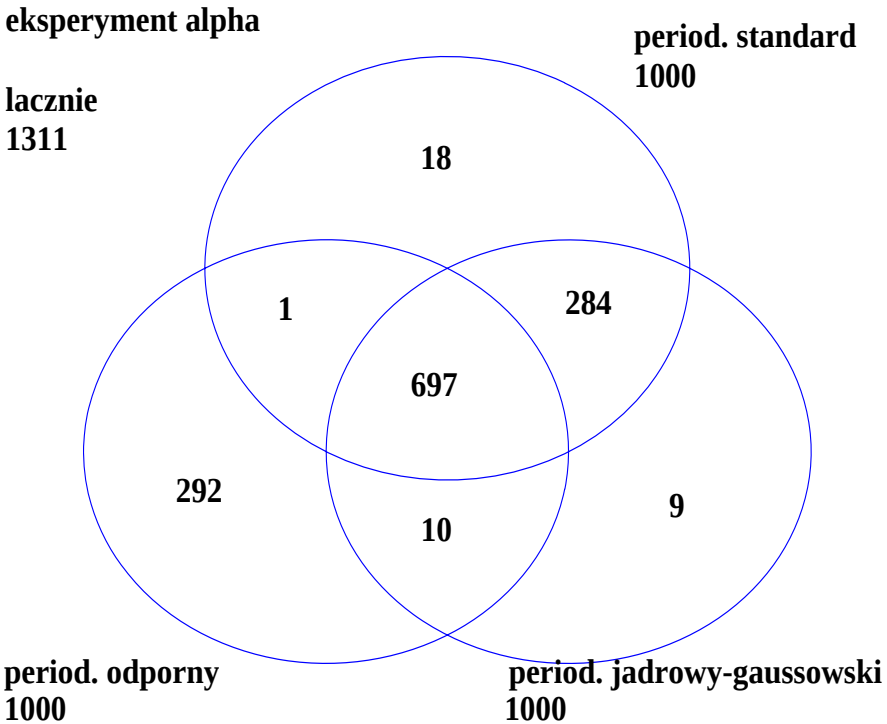
metody wykorzystujące wiedzę a-priori vs metody niewykorzystujące wiedzy a-priori

metody wykorzystujące wiedzę a-priori są skuteczniejsze



Porównanie testów – dane rzeczywiste 1

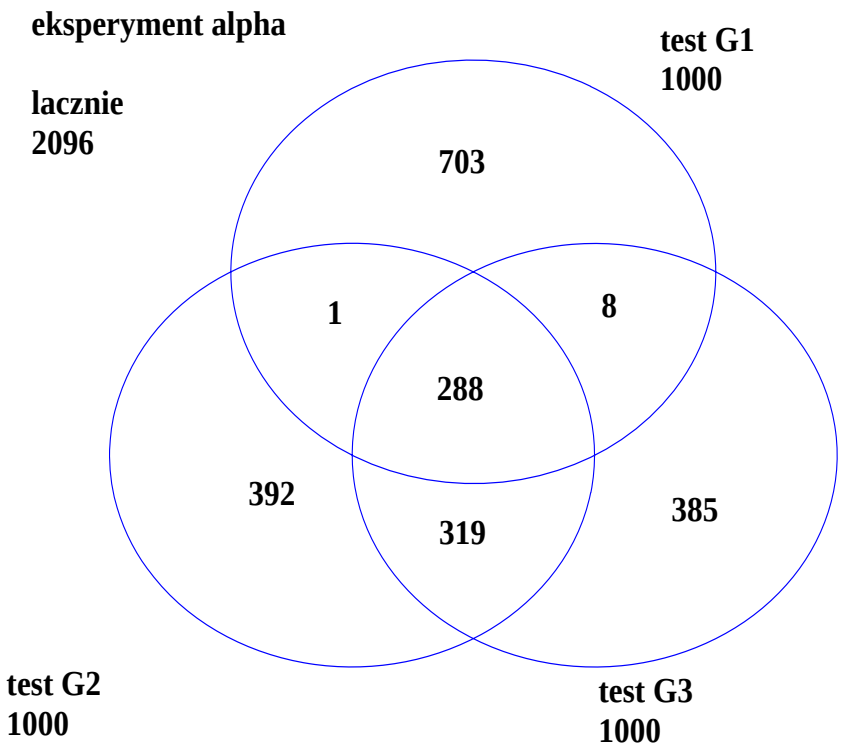
Stabilność testu w zależności od typu periodogramu



Rysunek: Diagram Venna – liczba genów z klasyfikacji '*TOP* 1000' dla eksperymentu 'alpha' dla testu G-Fishera w zależności od zastosowanego periodogramu

Porównanie testów – dane rzeczywiste 2

Rozbieżność wyników w zależności od zastosowanego testu

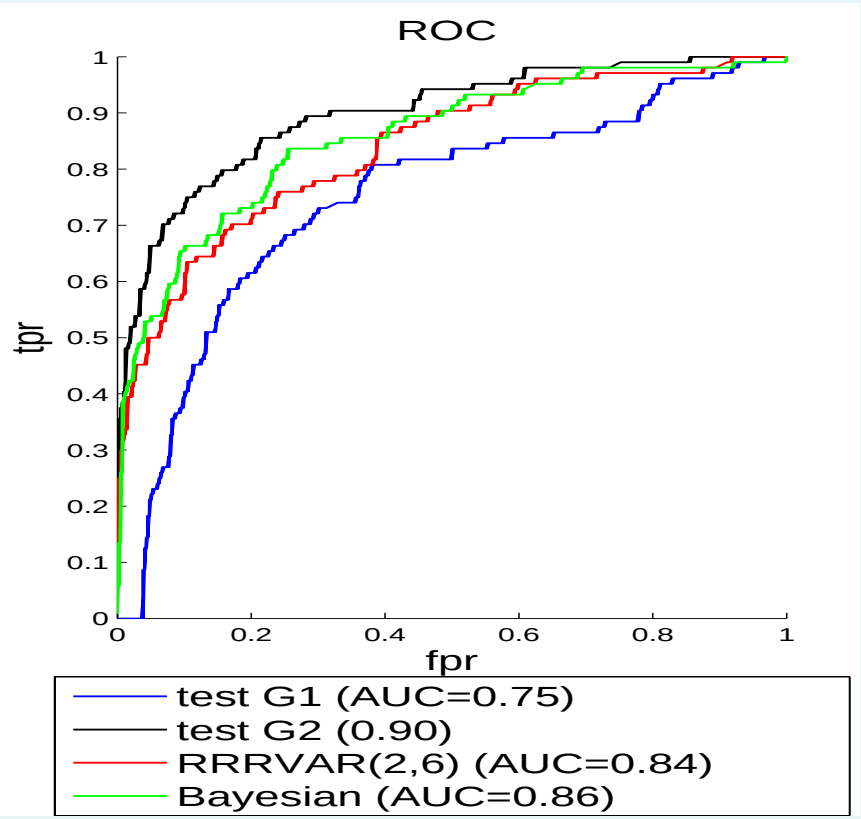


Rysunek: Diagram Venna – liczba genów z klasyfikacji '*TOP* 1000' dla eksperymentu 'alpha' przy zastosowaniu standardowego periodogramu

RRRVAR – porównanie ze standardowymi metodami

Konstrukcja porównania krzywych ROC

- zbiór testowy – 104 znane okresowe geny oraz 512 genów nie-Spellmana jako nieokresowe
- stosujemy różne metody w celu otrzymania scoringów
- dla każdej metody wyznaczamy ROC



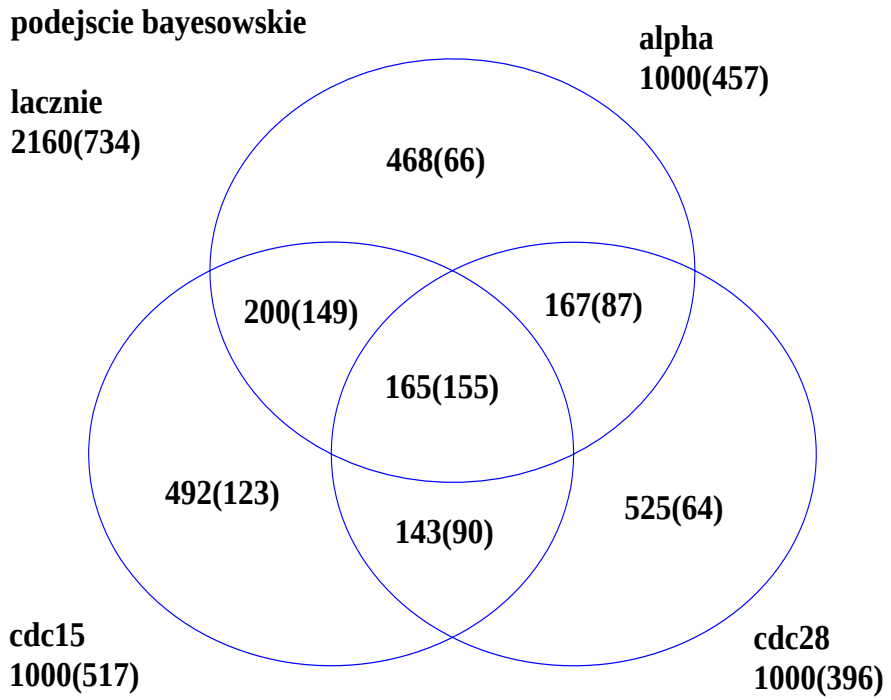
Podsumowanie

1. nową metodę identyfikacji składowych okresowych
   można sprawdzać używając opisanego schematu
2. metody wykorzystujące wiedzę a-priori mają mniejszy błąd
   klasyfikacji niż metody niewykorzystujące tej wiedzy
3. metoda RRRVAR jest porównywalna z metodami jednowymiarowymi
4. metoda RRRVAR jest lepsza od standardowego testu Fishera
   (obie nie wykorzystują żadnej wiedzy a-priori)
5. znacznie lepiej używać metod właściwych dla określonego
   typu funkcji okresowej
   – tę informację uzyskujemy używając np. RRRVAR

Dane rzeczywiste - rozbieżności między zbiorami danych

Jeden test – różne eksperymenty – brak zgodności wyników

- aktywność genów w czasie w cyklu podziału komórki drożdży (Spellman 1998)
- przeprowadzono 3 eksperymenty „cdc15", „cdc28", „alpha"
- dla każdych danych wyznaczmy „*TOP 1000*' genów w scoringu uzyskanym podejściem bayesowskim
- porównujemy listy otrzymanych genów okresowych
- w nawiasach liczba genów ze zbioru 800-Spellmana



Rysunek: Diagram Venna – liczba genów z klasyfikacji '*TOP* 1000' dla różnych eksperymentów mierzących okresową aktywność genów cyklu podziału komórki drożdży

Podsumowanie

1. nową metodę identyfikacji składowych okresowych
   można sprawdzać używając opisanego schematu
2. metody wykorzystujące wiedzę a-priori mają mniejszy błąd
   klasyfikacji niż metody niewykorzystujące tej wiedzy
3. metoda RRRVAR jest porównywalna z metodami jednowymiarowymi
4. metoda RRRVAR jest lepsza od standardowego testu Fishera
   (obie nie wykorzystują żadnej wiedzy a-priori)
5. znacznie lepiej używać metod właściwych dla określonego
   typu funkcji okresowej
   – tę informację uzyskujemy używając np. RRRVAR
6. należy zastosować kilka metod scoringowych do różnych zbiorów,
   a następnie „połączyć” rezulataty

Dziękuję za uwagę

Klasyczna analiza spektralna

Obliczanie scoringów

Wartości statystyk testowych traktujemy jako scoringi.

Obliczanie p-wartości dla różnych wersji estymatora gęstości spektralnej

1. dokładny rozkład statystyk testowych
   - standardowy periodogram
2. testowanie permutacyjne
   - odporny estymator gęstości spektralnej
   - wygładzony dyskretnie periodogram
3. metoda bootstrap w dziedzinie spektralnej
   - wygładzony jądrowo periodogram

Używane testy statystyczne (Brockwell)

G1 test Fishera ukrytej okresowości o nieokreślonej częstotliwości (G-test Fishera)

- $f(t) = \sin(\omega t + \tau_i)$
- $G_1 = \frac{\max_k I(\omega_k)}{\frac{1}{q_T}\sum_{k=1}^{q_T} I(\omega_k)}$

G2 test ukrytej okresowości o określonej częstotliwości ω_0

- $f(t) = \sin(\omega_0 t + \tau_i)$
- $G_2 = \frac{(n-3)I(\omega_0)}{\sum_{t=1}^{T} Y_i^2(t) - I(0) - 2I(\omega_0)}$

G3 test weryfikujący obecność niesinusoidalnego komponentu okresowego o znanym okresie p

- $f_i(t) = \sum_{k=1}^{\left[\frac{p-1}{2}\right]} \left[A_{i,k}\cos\left(\frac{2\pi kt}{p}\right) + B_{i,k}\sin\left(\frac{2\pi kt}{p}\right)\right] + A_{\frac{p}{2}}(-1)^t$
- $G_3 = \frac{(T-p)}{(p-1)} \frac{2\sum_{1\le j\le p/2} I(\omega_{rj}) + \delta_p I(\pi)}{\sum_{t=1}^{T} Y_i^2(t) - I(0) - \delta_p I(\pi) - 2\sum_{1\le j\le p/2} I(\omega_{rj})}$

Połączony test regularności i okresowości

Przeprowadzamy dwa testy – regularności i okresowości.
Następnie łączymy otrzymane rezultaty.

Regularność - podejście wielowymiarowe

Składowe okresowe wielowymiarowego szeregu czasowego mają zazwyczaj wiekszą wariancję wartości niż pozostałe składowe.
p-wartość testu regulacji wyzaczamy jako:

$$p_{i,reg} = \frac{1}{B}\sum_{j=1}^{B} 1\Big\{std\big(Y_i\big) \leq std\big(X_j\big)\Big\},$$

gdzie $X_j(t)$ dla $j = 1, .., B$ – B szeregów bazowych wyznaczonych metodą bootstrap poprzez replikowanie ze zbioru $\{Y_1(t), ..., Y_M(t)\}$

Połączony test regularności i okresowości

Okresowość - podejście jednowymiarowe

Dla każdej składowej obliczamy p-wartość $p_{i,period}$ dowolnego testu weryfikującego występowanie okresowości.

Określenie scoringu

$$s_i = (p_{i,reg} \cdot p_{i,period}) \left[1 + \left(\frac{p_{i,reg}}{a}\right)^2\right] \left[1 + \left(\frac{p_{i,period}}{a}\right)^2\right],$$

gdzie $a \leq 1$ – współczynnik kary – powoduje, że składowe, które są nieistotne ani regularnie ani okresowo, otrzymują wyższy scoring
Końcowy scoring otrzymujemy poprzez przekształcenie:

$$s_{i,final} = \exp(-s_i)$$

Idea

Zakładamy model

$$Y_i(t) = a_{i,0} + a_{i,1} \cdot f(t - \tau_i) + Z_i(t)$$

Konstruujemy kompozytowy (zagregowany) szereg czasowy

$$X(t) = \beta^\star U(t)$$
$$f_X(\omega, \beta) = \beta^\star f_U(\omega)\beta$$

gdzie

$$U(t) = V^{-1/2} Y(t)$$
$$V = diag\{\sigma^2_{Y_1}, ..., \sigma^2_{Y_M}\}$$

Obwiednia spektralna

Definicja obwiedni spektralnej

$$\lambda(\omega) = \sup_{\beta \neq 0} \{ f_X(\omega, \beta) \}$$

Po przekształceniach

$$\lambda(\omega) = \sup_{b \neq 0} \left\{ \frac{b^\star f_Y(\omega) b}{b^\star V b} \right\},$$

gdzie $b = V^{-1/2}\beta$

Obwiednia spektralna

Rozwiązanie

$\lambda(\omega)$ jest największą wartością własną macierzy

$$f_U(\omega) = V^{-1/2} f_Y(\omega) V^{-1/2}, \tag{1}$$

natomiast $\beta(\omega)$ jest odpowiadającym wektorem własnym.

Zaproponowana interpretacja

1. $\beta(\omega)$ – optymalne M-wymiarowe skalowanie gęstości spektralnej f_Y w punkcie ω
2. $|\beta(\omega_0)|$ – wektor scoringów
3. $\lambda(\omega)$ – narzędzie wstępnej analizy szeregu czsowego

Stosowane wersje estymatorów gęstości spektralnej

- standardowy periodogram

$$I(\omega_k) = \sum_{|h|<T} \hat{\gamma}(h) \exp^{-i\omega_k h}$$

- odporny estymator gęstości spektralnej

$$I_{*1}(\omega) = \sum_{k=-T-2}^{t=T-2} \tilde{\rho}(k) \exp(-i\omega k)$$

- wygładzony dyskretnie periodogram

$$I_{*2}(\omega_j) = (2\pi)^{-1} \sum_{|k|\leq m} W(k) I(\omega_{j+k})$$

- wygładzony jądrowo periodogram

$$I_{*3}(\omega_j) = \hat{f}(\omega_j, h) = \frac{1}{Th} \sum_{k=-q_T}^{q_T} K\left(\frac{\omega_j - \omega_k}{h}\right) I(\omega_k)$$

Podejście bayesowskie - idea

Model:

$$H_0 : Y_i(t) = a_{i,0,0} + Z_i(t) \qquad \text{(składowa nieokresowa)}$$
$$H_1 : Y_i(t) = a_{i,1,0} + a_{i,1,1} \cdot \cos(\omega_0 t + \tau_i) + Z_i(t) \qquad \text{(składowa okresowa)}$$

Obliczamy prawdopodobieństwa a posteriori dla każdej składowej. Stosując aproksymację bayesowskiego kryterium informacyjnego (BIC) obliczmy scoringi.

Prawdopodobieństwa a posteriori

$$P(H_k|D) = \frac{P(D|H_k)P(H_k)}{\sum_{j=\{1,2\}} P(D|H_j)P(H_j)}$$
$$P(D|H_k) = \int_{\phi_{i,k} \in \Phi_{i,k}} p(D|\phi_{i,k}, H_k) p(\phi_{i,k}|H_k)\, d\phi_{i,k}$$

Podejście bayesowskie

Założenia a priori

- parametry niezależne
- znana średnia μ_{ω_0}, wariancja σ_{ω_0}
- rozkład normalny parametru ω_0
- rozkład jednostajny parametrów $a_{i,k,j}$
- rozkład normalny zaburzeń Z_i

Wyznaczenie scoringów

Aproksymacja bayesowskiego kryterium informacyjnego

$$BIC(H_k) = \log p(D|\phi_{i,k}^{MAP}, H_k) - \frac{d}{2} \log T \sim \log P(D|H_k)$$

$$g(D) = BIC(H_1) - BIC(H_0)$$

$$s_i = \frac{\exp^{g(D)}}{1 + \exp^{g(D)}}$$

Funkcje B-sklejane

P - wymiar bazy
$1 = t_0 \leq t_1 \leq ... \leq t_P = T$ - węzły
$\{b_{p,k}(t),\ p = 0, 1, .., P-1\}$ - baza funkcji B-sklejanych rzędu k

Wzory rekurencyjne Coxa-de Boora

$$b_{p,0}(t) = \begin{cases} 1 & , \quad t_p \leq t \leq t_{p+1} \\ 0 & , \quad t \notin (t_p, t_{p+1}) \end{cases}, \qquad \text{dla } p = 0, 1, .., P-1$$

$$b_{p,j}(t) = 0, \qquad \text{gdy } p \geq P \text{ i } j \geq 0,$$

$$b_{p,k}(t) = \frac{t - t_p}{t_{p+k} - t_p} b_{p,k-1}(t) + \frac{t_{p+k+1} - t}{t_{p+k+1} - t_{p+1}} b_{p+1,k-1}(t)$$

$$\text{dla } k > 0 \text{ i } p < P.$$

Funkcje B-sklejane

Baza sześciennych funkcji B-sklejanych - aproksymacja funkcji

$$B_p(t) = b_{p-1,3}(t), \qquad \text{dla } p = 1, ..., P$$

$$f(t) = \sum_{p=1}^{P} \gamma_p B_p(t)$$

Model szeregu czasowego

$$Y_i(t) = a_{i,0} + a_{i,1} f(t - \tau_i) + Z_i(t), \qquad \text{dla } i = 1, ..., M \text{ oraz } t = 1, ..., T$$

Zbiór uczący

szeregi czasowe współdzielące wspólną funkcję okresową $f(t)$

$$U_l(t) = a_{l,0} + a_{l,1} f(t - \tau_l) + Z_l(t), \qquad \text{dla } l = 1, ..., L$$

Funkcje B-sklejane
Trzykrokowa procedura estymacji współczynników

KROK 1 szukamy $a_{l,0}, a_{l,1}, \tau_l$,
przyjmujemy $f(t) = \sin(t)$

$$\min \left\{ \frac{1}{T} \sum_{t=1}^{T} [U_l(t) - a_{l,0} - a_{l,1} f(t - \tau_l)]^2 \right\}$$

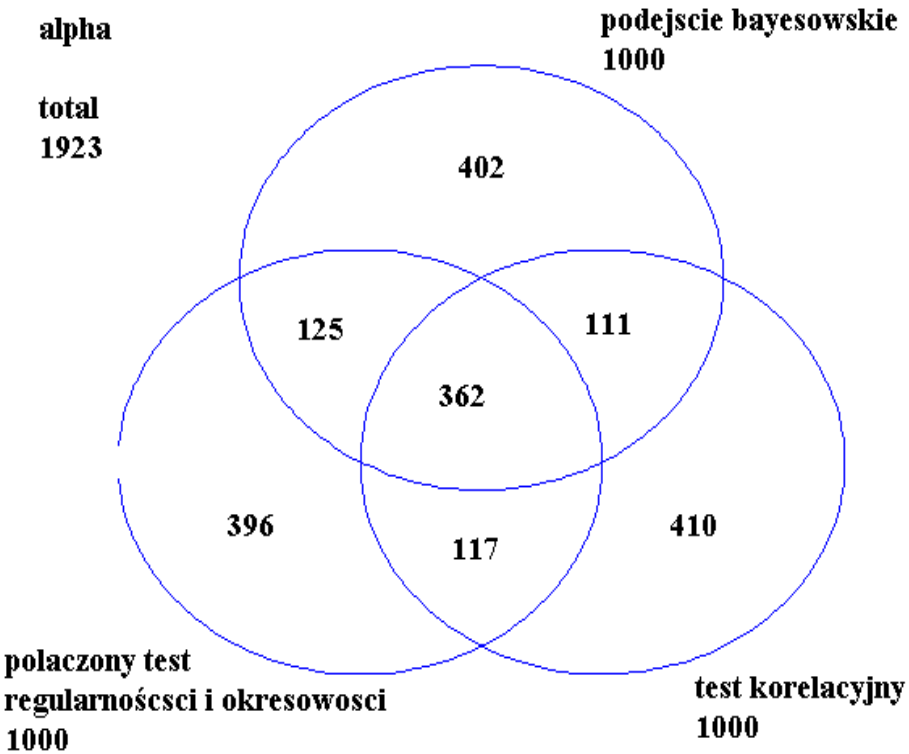
KROK 2 szukamy γ_p,
przyjmujemy $a_{l,0}, a_{l,1}, \tau_l$ obliczone w kroku 1

$$\min \left\{ \sum_{l=1}^{L} \frac{1}{T} \sum_{t=1}^{T} [u_l(t) - a_{l,0} - a_{l,1} \sum_{p=1}^{P} \gamma_k B_p(t - \tau_l)]^2 \right\}$$

KROK 3 szukamy $a_{i,0}, a_{i,1}, \tau_i$,
przyjmujemy γ_p obliczone w kroku 2

$$\min \left\{ \frac{1}{T} \sum_{t=1}^{T} [y_i(t) - a_{i,0} - a_{i,1} f(t - \tau_i)]^2 \right\}$$

Stabilność metod - diagramy Venna - dane rzeczywiste



Rysunek: Przykładowy diagram Venna – liczba genów z klasyfikacji '*TOP* 1000' dla eksperymentu 'alpha'