Два метода прогнозирования временных рядов

Попов Сергей Альбертович, гр. 522

Санкт-Петербургский государственный университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель: к.ф.-м.н., доц. Голяндина Н.Э.
Рецензент: к.ф.-м.н., доц. Некруткин В.В.

Санкт-Петербург
2010г.

Наблюдается ряд $F_N = (f_0, \ldots, f_{N-1})$.
$F_N = S_N + R_N$, где S_N — «сигнал», R_N — «шум».
Предполагаем:

- S_N управляется ЛРФ. $s_n = \sum_{i=1}^{r} a_i s_{n-i}$, $n = r, \ldots, N-1$
- R_N — белый гауссовский шум

Задача: прогноз S_N, наблюдая F_N

- отделение сигнала от шума
- идентификация структуры сигнала, т. е. оценивание r и $(a_1, \ldots, a_r)$

$$\widehat{s}_N = \sum_{i=1}^{r} \widehat{a}_i \widehat{s}_{N-i}$$

Точность: $\mathrm{MSE} = \mathrm{E}(\widehat{s}_N - s_N)^2$

Задачи отделения сигнала и идентификации структуры сигнала решаются одновременно методом Singular Spectrum Analysis.

- Выбирается параметр $1 < L < N$ («длина окна»).

$$F_N \xrightarrow{L} \mathbf{X} = \begin{pmatrix} f_0 & f_1 & \dots & f_{K-1} \\ f_1 & f_2 & \dots & f_K \\ \vdots & \vdots & \ddots & \vdots \\ f_{L-1} & f_L & \dots & f_{N-1} \end{pmatrix}, \quad \mathbf{X} = \mathbf{S} + \mathbf{R}$$

Пусть размерность минимальной ЛРФ управляющей сигналом $=$ r $(= \operatorname{rank} \mathbf{S})$.
Пространство, натянутое на столбцы $\mathbf{S}$ — подпространство сигнала $(\mathcal{L}_r)$.

- По $\mathbf{X}$ оценивается базис подпространства сигнала $\{P_i\}_{i=1}^{r}$
- $\widehat{\mathcal{L}}_r = \operatorname{span}(\widehat{P}_1, \dots, \widehat{P}_r)$ — оценка подпространства сигнала.
    - Восстановление: проектирование столбцов $\mathbf{X}$ на $\widehat{\mathcal{L}}_r$
    - Оценка a_i: по $\widehat{\mathcal{L}}_r$

Два метода оценивания $\mathcal{L}_r$

- Basic SSA $\Longrightarrow$ рекуррентный прогноз
- Toeplitz SSA $\Longrightarrow$ теплицев прогноз

Задачи:

- Сравнить точность прогнозирования двумя методами
- Выработать рекомендации по выбору оптимальных параметров для прогнозирования

Примеры: $f_n = s_n + r_n$, где

- $s_n = \cos(2\pi n/10)$
- $s_n = (0.999)^n \cos(2\pi n/10)$
- $s_n = (0.9999)^n \cos(2\pi n/10)$

$r_n = \varepsilon_n$, $\varepsilon_n \approx N(0, \sigma^2)$ с $\sigma = 0$ или $\sigma = 0.1$.
$N = 400$

Построение оценки базиса подпространства сигнала:

- Выбор длины окна L. Построение траекторной матрицы $\mathbf{X}$.
- Ортогональное разложение $\mathbf{X} = \sum\limits_{i=1}^{L} \sqrt{\lambda_i} P_i Q_i^{\mathrm{T}}$, $P_i \perp P_j$
    - Рекуррентный прогноз P_i — с. в. матрицы $\mathbf{S} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$
    - Теплицев прогноз P_i — с. в. оцененной ковариационной матрицы $\mathbf{C}$ с элементами

$$c_{kj} = \frac{1}{N - |k-j|} \sum_{m=0}^{N-|k-j|-1} f_m f_{m+|k-j|}, \quad 1 \leq k, j \leq L,$$

- Оценивание подпространства сигнала $\widehat{\mathcal{L}}_r = \mathrm{span}(P_1, \ldots, P_r)$

В базовом варианте прогнозов берется одна и та же длина окна L для восстановления сигнала и для оценивания коэффициентов ЛРФ.

Идея: можно рассмотреть отдельно L_V и L_R.

Применение прогноза к ряду без шума

Рассматривается ряд $s_n = c^n \cos(2\pi\omega n)$.

Basic SSA

- Для $\mathbf{S} = \mathbf{X}\mathbf{X}^{\mathrm{T}}$ известно, что ранг матрицы $\mathbf{S}$ равен 2.
- Собственные векторы $\mathbf{X}$ — экспоненциально модулированные гармоники с частотой ω.

Toeplitz SSA

- Можно доказать, что в случае $c = 1$ ранг оцененной ковариационной матрицы асимптотически равен двум.
- Вычислительные эксперименты показывают, первые два собственных вектора ковариационной матрицы асимптотически будут гармониками для рассматриваемых рядов.

$f_n = c^n \cos(2\pi\omega n)$

Предложение

- $c = 1 :$ $\|\mathbf{C} - \widetilde{\mathbf{C}}\|_2 \xrightarrow{K \to \infty} 0$, *где*

$$\widetilde{c}_{kj} = \frac{1}{2}\cos(2\pi\omega(k-j))$$

и матрица $\widetilde{\mathbf{C}}$ *имеет ранг два*

- $c < 1 :$ $\|N\mathbf{C} - \widetilde{\mathbf{C}}\|_2 \xrightarrow{K \to \infty} 0$, *где*

$$\widetilde{c}_{kj} = \frac{1}{2}\left(\frac{1}{1-c^2} + \frac{1}{1-c^2\cos(4\pi\omega)}\right) c^{|k-j|}\cos(2\pi\omega(k-j))$$

План дальнейшего иследования

- рассматриваются два варианта прогнозирования: теплицев и рекуррентный
- L_R, L_V — параметры прогнозов
- ищутся оптимальные параметры, прогнозы сравниваются при оптимальных параметрах
- рассматривается структура ошибки: $\mathrm{MSE}(\widehat{S}_N, S_N) = \mathrm{Var}(S_N) + (\mathrm{Bias}(\widehat{S}_N, S_N))^2$
- рассматриваются источники ошибок

Ошибки прогнозов без шума. $s_n = c^n \cos(2\pi n/10)$.

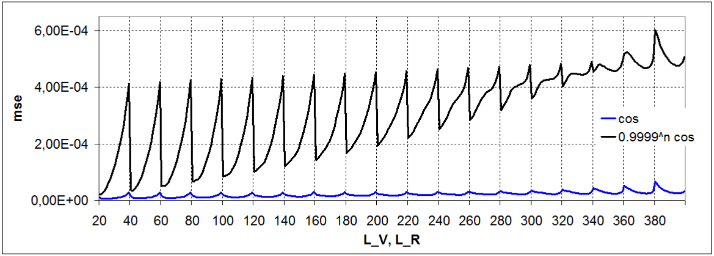

Рис.: Теплицев прогноз. $c = 1$ и $c = 0.9999$.

- MSE тепл. $> \mathrm{MSE}$ рекурр. $(= 0)$
- Для теплицева прогноза ошибка тем больше, чем c дальше от единицы
- $c = 1$. L_V — минимально возможная, $L_R \approx 40$
- $c \neq 1$. L_V и L_R — минимально возможные

Ошибки прогнозов с шумом. $s_n = c^n \cos(2\pi n/10)$.

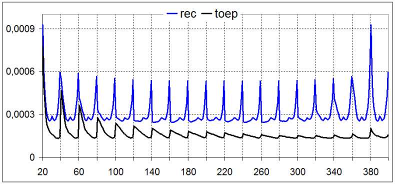

Рис.: Ошибки теплицева и рекуррентного прогнозов $c = 1$.

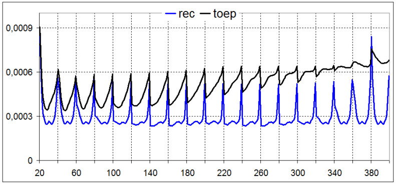

Рис.: Ошибки теплицева и рекуррентного прогнозов $c = 0.9999$.

- $c = 1$. MSE тепл. $<$ MSE рекурр
- $c \neq 1$. MSE тепл. $>$ MSE рекурр
- Рекуррентный прогноз. $L_V \approx 200$, $L_R \in (100, 200)$
- Теплицев прогноз.
  - $c = 1$: L_V — любая, $L_R \in (120, 140)$
  - $c = 0.9999$: L_V — любая, $L_R \in (40, 120)$
  - $c = 0.999$: L_V — минимальная, $L_R \approx 40$
- Устойчивые: L_V — минимальная, $L_R \approx 200$

Структура ошибки. $s_n = c^n \cos(2\pi n/10)$

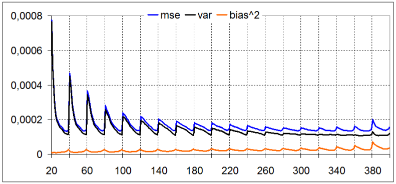

Рис.: Теплицев прогноз. $c = 1$.

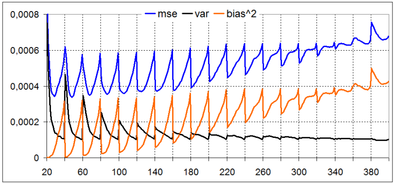

Рис.: Теплицев прогноз. $c = 0.9999$.

- Теплицев ($c = 1$) и рекуррентный прогнозы (c — любое), $\mathrm{Bias} \approx 0, \mathrm{MSE} \approx \mathrm{Var}$
- Теплицев прогноз. $c \neq 1$, $\mathrm{MSE} \approx \mathrm{Bias}$
- Смещение объясняется тем, что для незашумленного модулированного косинуса ошибка прогноза не равна нулю
.

Пусть:

- $V = (s_{N-L}, \dots, s_{N-1})$
- $V + \Delta V = (\widehat{s}_{N-L}, \dots, \widehat{s}_{N-1})$
- $A = (a_1, \dots a_{L-1})$
- $A + \Delta A = (\widehat{a}_1, \dots \widehat{a}_{L-1})$

Тогда:

- $s_N = A^{\mathrm{T}} V$
- $\widehat{s}_N - s_N = (A + \Delta A)^{\mathrm{T}} (V + \Delta V) - A^{\mathrm{T}} V$

Выделяются два источника ошибок:

- Реальный восстановленный ряд прогнозируется по идеальной ЛРФ ($A^{\mathrm{T}} \Delta V$).
- Идеальный восстановленный ряд прогнозируется по реальной ЛРФ ($\Delta A^{\mathrm{T}} V$).

Два источника ошибок. $s_n = c^n \cos(2\pi n/10)$ $L_V = 140$

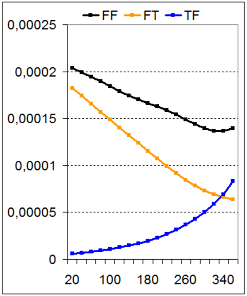

Рис.: Рекуррентный прогноз $c = 1$.

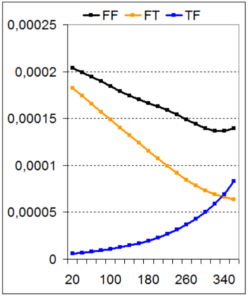

Рис.: Теплицев прогноз $c = 1$.

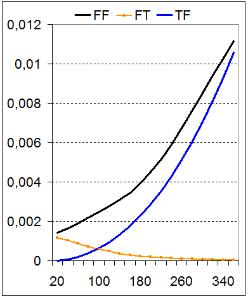

Рис.: Теплицев прогноз $c = 0.9999$.

- Поведение при разных L_V одинаково
- TF возрастает. Поведение ошибки проектора
- FT убывает. Поведение побочных корней характеристического полинома

Полученные результаты:

- Были получены результаты относительно асимптотических рангов ковариационных матриц
- При помощи моделирования было проведено сравнение ошибок теплицева и рекуррентного прогнозов
- Сделаны предположения по подбору оптимальных параметров для прогнозов для рассматриваемых рядов.