Анализ структуры шума при ошибках в значениях и аргументах функции

Абрамова Анастасия Николаевна, гр. 422

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
2015г.

Введение: общая постановка задачи

Временной ряд $F = (f_1, \ldots, f_N)$, N — длина ряда.

u — достаточно гладкая функция.

$$f_i = u(x_i + \varepsilon_i), \qquad (X_A)$$
$$f_i = u(x_i + \varepsilon_i) + \delta_i, \qquad (X_A Y_A)$$
$$f_i = u(x_i + \varepsilon_i)(1 + \delta_i). \qquad (X_A Y_M)$$

Случайные величины $\varepsilon_i \sim N(0, \sigma_x^2), \delta_i \sim N(0, \sigma_y^2)$ и независимы в совокупности.

Задача

Необходимо оценить параметры σ_x^2 и σ_y^2.

Приближенные модели первого порядка:

$$h_i = u(x_i) + u'(x_i)\varepsilon_i, \qquad (X_A)_1$$
$$h_i = u(x_i) + u'(x_i)\varepsilon_i + \delta_i, \qquad (X_A Y_A)_1$$
$$h_i = (1+\delta_i)u(x_i) + \varepsilon_i(1+\delta_i)u'(x_i). \qquad (X_A Y_M)_1$$

Приближенные модели второго порядка:

$$g_i = u(x_i) + u'(x_i)\varepsilon_i + \frac{\varepsilon_i^2}{2}u''(x_i), \qquad (X_A)_2$$
$$g_i = u(x_i) + u'(x_i)\varepsilon_i + \delta_i + \frac{\varepsilon_i^2}{2}u''(x_i), \qquad (X_A Y_A)_2$$
$$g_i = (1+\delta_i)\left(u(x_i) + \varepsilon_i u'(x_i) + \frac{\varepsilon_i^2 u''(x_i)}{2}\right). \qquad (X_A Y_M)_2$$

Шум

$$\varepsilon_i \sim N(0, \sigma_x^2), \ \ \delta_i \sim N(0, \sigma_y^2), \text{ независимы.}$$

Нужно: оценить параметры σ_x^2 и σ_y^2.
Методы: метод максимального правдоподобия (ММП), взвешенный метод наименьших квадратов (ВМНК).

Было получено Федоренко (2013):

- Оценки по ММП для моделей $(X_A)_1, (X_A Y_A)_1$,
- Оценки по ВМНК для модели $(X_A)_1, (X_A Y_A)_1$и$(X_A Y_M)_1$.

Задачи данной работы:

- Построить оценки σ_x^2 и σ_y^2 по ММП в моделях $(X_A)_2, (X_A Y_A)_2$ и $(X_A Y_M)_1$.
- Для случая $x_i = i/N$ получить асимптотические (при $N \to \infty$) характеристики оценок по ММП и ВМНК.

Определение

Рассмотрим регрессионное уравнение

$$Y = \mathbf{X}B + R.$$

$\mathbf{X}$ *— известная матрица* $N \times m$, $m < N$, $\mathrm{rk}\mathbf{X} = \mathrm{m}$.

$Y = (y_1, \ldots, y_N)^{\mathrm{T}} \in \mathbb{R}^k$ *— случайный вектор.*

$R = (r_1, \ldots, r_N)^{\mathrm{T}} \in \mathbb{R}^k$ *— случайный вектор,* $\mathbb{E}r_i = 0$, $\mathbb{E}r_i^2 < \infty$,

$B = (b_1, \ldots, b_m)^{\mathrm{T}}$ *— вектор неизвестных параметров, которые необходимо оценить.*

Оценкой по взвешенному методу наименьших квадратов $\widehat{B}$ *называется*

$$\widehat{B} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}^{-1}Y,$$

где $\mathbf{W}$ *— симметричная, положительно определенная матрица.*

Известно, что если $\mathbf{W} = \Sigma_R$, то оценка по ВМНК — наилучшая в классе линейных несмещенных оценок.

Проблема: Распределение ошибок r_i зависит от параметров B и поэтому $\mathbf{W} = \mathbf{W}_B$. Формула $\widehat{B} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}_B^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}_B^{-1}Y$ нереализуема.

Поиск оценок:

$$\widehat{B}^{(i+1)} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}_{\widehat{B}^{(i)}}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}_{\widehat{B}^{(i)}}^{-1}Y,$$

$$\lim_{i\to\infty}\widehat{B}^{(i)} = \widehat{B}^{(\infty)}.$$

Теорема (например, Демиденко (1981))

Рассмотрим последовательность уравнений $Y_N = \mathbf{X}_N B + R_N$, *где* $r_i = r_{iN}$ *независимы в каждой серии и* $\mathbb{E}r_i = 0$, $\mathbb{E}r_i^2 < \infty$,
$\mathbf{W}_N = \mathrm{diag}(\mathbb{D}\mathrm{r}_1, \ldots, \mathbb{D}\mathrm{r}_\mathrm{N})$, $\det \mathbf{W}_N \neq 0$. *Пусть* $\overline{\mathbf{X}}_N = \mathbf{W}_N^{-1/2}\mathbf{X}_N$ *сильно регулярны, то есть*

$$\lim_{N\to\infty} \frac{1}{N} \mathbf{X}_N^{\mathrm{T}} \mathbf{W}_N^{-1} \mathbf{X}_N = \mathbf{A}, \tag{1}$$

где $\det \mathbf{A} \neq 0$. *Если* $\forall i = 1, 2, \ldots, m$ *выполняется*

$$\lim_{N\to\infty} \frac{1}{N} \max_{t=1,\ldots,N} \overline{x}_{ti}^2 = 0, \tag{2}$$

то оценка ВМНК $\widehat{B}_N$ *асимптотически нормальна. Более того,*

$$\lim_{N\to\infty} \sqrt{N}(\widehat{B}_N - B) \sim \mathcal{N}(0, \mathbf{A}^{-1}).$$

Рассмотрим на примере модели $(X_A Y_A)_1$.

Модель имеет вид $h_i = u(x_i) + u'(x_i)\varepsilon_i + \delta_i$.

Математическое ожидание $\mathbb{E}(h_i - u(x_i))^2 = \sigma_x^2 (u'(x_i))^2 + \sigma_y^2$.

$$(h_i - u(x_i))^2 = \mathbb{E}(h_i - u(x_i))^2 + r_i,$$

$r_i = (\varepsilon_i^2 - \sigma_x^2)(u(x_i))^2 + (\delta_i^2 - \sigma_y^2) + 2\varepsilon_i \delta_i u(x_i)$.

$\mathbb{E} r_i = 0$, $\mathbb{D} r_i = 2(\sigma_x^2 (u(x_i))^2 + \sigma_y^2)^2$, независимы.

Таким образом, в регрессионном уравнении для модели $(X_A Y_A)_1$ параметры имеют вид:

$$B = (\sigma_x^2, \sigma_y^2),$$

$$Y = ((h_1 - u(x_1))^2, \ldots, (h_N - u(x_N))^2)^{\mathrm{T}},$$

$$\mathbf{X} = [X_1, X_2],\ X_1 = ((u'(x_1))^2, \ldots, (u(x_N))^2)^{\mathrm{T}},\ X_2 = (1, \ldots, 1)^{\mathrm{T}}.$$

Весовая матрица имеет вид

$$\mathbf{W} = \mathrm{diag}(\mathbb{D} r_1, \ldots, \mathbb{D} r_N).$$

Асимптотическая нормальность:

$$\lim_{N\to\infty} \sqrt{N}(\widehat{B}_N - B) \sim N(0, \mathbf{A}^{-1}).$$

Для приближенных моделей первого порядка были проверены условия асимптотической нормальности (1) и (2). Выпишем вид матрицы $\mathbf{A}$ для каждой модели.

Модель $(X_A)_1$

$B = \sigma_x^2$, $\mathbf{A} = 2\sigma_x^4$.

Модель $(X_A Y_A)_1$

$$B = (\sigma_x^2, \sigma_y^2)^{\mathrm{T}},\ \mathbf{A} = \begin{pmatrix} \int\limits_0^1 \frac{(u'(t))^4}{\mathbb{D}r(t)}\,dt & \int\limits_0^1 \frac{(u'(t))^2}{\mathbb{D}r(t)}\,dt \\ \int\limits_0^1 \frac{(u'(t))^2}{\mathbb{D}r(t)}\,dt & \int\limits_0^1 \frac{1}{\mathbb{D}r(t)}\,dt \end{pmatrix},$$

где $\mathbb{D}r(t) = 2(\sigma_x^2(u'(t))^2 + \sigma_y^2)^2$.

Модель $(X_A Y_M)_1$

$B = (\sigma_y^2, \sigma_x^2(1+\sigma_y^2))^{\mathrm{T}}$,

$$\mathbf{A} = \begin{pmatrix} \int\limits_0^1 \frac{u^4(t)}{\mathbb{D}r(t)}\,dt & \int\limits_0^1 \frac{u^2(t)(u'(t))^2}{\mathbb{D}r(t)}\,dt \\ \int\limits_0^1 \frac{u^2(t)(u'(t))^2}{\mathbb{D}r(t)}\,dt & \int\limits_0^1 \frac{(u'(t))^4}{\mathbb{D}r(t)}\,dt \end{pmatrix},$$

$\mathbb{D}r(t) = 2\sigma_x^2(1+2\sigma_y^2)^2 u'^4(t) + 2\sigma_y^2 u^4(t) + 4\sigma_x^2\sigma_y^2(1+3\sigma_y^2)u^2(t)u'^2(t)$.

Для вектора $\widehat{B}_N^* = (\widehat{\sigma}_{xN}^2, \widehat{\sigma}_{yN}^2)^{\mathrm{T}}$, $B^* = (\sigma_x^2, \sigma_y^2)^{\mathrm{T}}$ получаем

$$\lim_{N\to\infty} \sqrt{N}(\widehat{B}_N^* - B^*) \sim N(0, \mathbf{G}^{-1}),$$

где

$$\mathbf{G} = \begin{pmatrix} 0 & 1 \\ 1+\sigma_y^2 & \sigma_x^2 \end{pmatrix} \mathbf{A}^{-1} \begin{pmatrix} 0 & 1 \\ 1+\sigma_y^2 & \sigma_x^2 \end{pmatrix}^{\mathrm{T}}.$$

Определение

Пусть $\xi = (\xi_1, \dots, \xi_N)$ — случайный вектор в евклидовом пространстве с распределением $\mathcal{P}_\theta$ и плотностью p_θ, где $\theta = (\theta_1, \dots, \theta_k)^{\mathrm{T}} \in \Theta$ — вектор неизвестных параметров, Θ — открытое подмножество $\mathbb{R}^k$. Функцией правдоподобия для вектора ξ называется функция N переменных

$$\mathcal{L}(z_1, \dots, z_N \mid \theta) = p_\theta(z_1, \dots, z_N).$$

Оценкой максимального правдоподобия характеристики θ называется

$$\widehat{\theta}_{\text{омп}} = \underset{\theta \in \Theta}{\arg\,\max}\, \mathcal{L}(\xi_1, \dots, \xi_N \mid \theta).$$

Если в данном определении случайные величины ξ_i, $i = 1, \dots, N$, независимы в совокупности, и плотность случайной величины ξ_i обозначить за $p_{\theta,i}$, то функция правдоподобия будет иметь вид

$$\mathcal{L}(z_1, \dots, z_N \mid \theta) = \prod_{i=1}^{N} p_{\theta,i}(z_i),$$

Теорема

Функция правдоподобия для g_i *в модели* $(X_A)_2$ *имеет вид*

$$\mathcal{L}(g_1, \dots g_N | \theta) = \prod_{i=1}^{N} \frac{\left(p_\varepsilon\left(\frac{u'(x_i)+D(g_i)}{u''(x_i)}\right) + p_\varepsilon\left(\frac{-u'(x_i)+D(g_i)}{u''(x_i)}\right)\right)}{D(g_i)} \mathbb{I}(g_i),$$

где

$$\mathbb{I}(g_i) = \begin{cases} 1, & \text{если } g_i \in \left(u(x_i) - \frac{u'(x_i)}{2u''(x_i)}; \infty\right), u''(x_i) > 0 \text{ или} \\ & g_i \in \left(-\infty; u(x_i) - \frac{u'(x_i)}{2u''(x_i)}\right), u''(x_i) < 0 \\ 0, & \text{иначе} \end{cases},$$

$D(g_i) = \sqrt{(u'(x_i))^2 - 2u''(x_i)(u(x_i) - g_i)} > 0$ *при* $\mathbb{I}(g_i) = 1$.

ММП: функция правдоподобия для $(X_A Y_A)_2$

Теорема

Функция правдоподобия для g_i в модели $(X_A Y_A)_2$ имеет вид:

$$\mathcal{L}(g_1, \ldots, g_N \mid \overline{\theta}) = \prod_{i=1}^{N} \int\limits_{-\infty}^{\infty} p_\delta \left(\psi \left(g_i, s; \frac{u''(x_i)}{2}, u'(x_i), u(x_i) \right) \right) p_\varepsilon(s)\, ds.$$

p_ε *— плотность распределения* $\mathcal{N}(0, \sigma_x^2)$,
p_δ *— плотность распределения* $\mathcal{N}(0, \sigma_y^2)$,

$$\psi(y, s; a, b, c) = y - as^2 - bs - c.$$

ММП: функция правдоподобия для $(X_A Y_M)_1$

Теорема

Функция правдоподобия для h_i *модели в модели* $(X_A Y_M)_1$ *имеет вид:*

$$\mathcal{L}(h_1, \ldots, h_N \mid \overline{\theta}) = \prod_{i=1}^{N} \frac{1}{|u'(x_i)|} \int\limits_{-\infty}^{\infty} \frac{p_\delta(s)}{|(1+s)|} p_\varepsilon(\varphi(h_i, s; u(x_i), u'(x_i)))\, ds.$$

p_ε *— плотность распределения* $\mathcal{N}(0, \sigma_x^2)$,
p_δ *— плотность распределения* $\mathcal{N}(0, \sigma_y^2)$,

$$\varphi(y, s; a, b) = \frac{y - a(1+s)}{|b|(1+s)}.$$

Задача: аппроксимировать интеграл

$$\int\limits_{-\infty}^{\infty} s(x)t(x)\,dx < \infty,$$

где $s(x) \underset{x\to\infty}{\longrightarrow} 0$, $t(x) \underset{x\to\infty}{\longrightarrow} 0$.

1. Если $s(x)t(x) < \tau$ на $\mathbf{K}(\tau)$, то существует монотонная функция $\lambda(\tau)$:

$$\left| \int\limits_{K(\tau)} s(x)t(x)\,dx \right| < \lambda(\tau) = \epsilon.$$

2. $\mathbf{K}(\tau) = \mathbf{K}^{(s)}(\tau) \cap \mathbf{K}^{(t)}(\tau)$, где
   $s(x) < \sqrt{\tau}$ на $\mathbf{K}^{(s)}(\tau)$,
   $t(x) < \sqrt{\tau}$ на $\mathbf{K}^{(t)}(\tau)$.

Таким образом: фиксируем ϵ и по ним находим $\mathbf{K}^{(s)}(\lambda^{-1/2}(\epsilon))$, $\mathbf{K}^{(t)}(\lambda^{-1/2}(\epsilon))$ и пересекаем их.

В ходе численных экспериментов на модельных примерах были получены следующие результаты:

1. **для данных** $(X_A)_2$, (X_A)

   оценки по ММП в $(X_A)_2 <_{\text{bias}}$ оценки по ММП $(X_A)_1$;
   оценки по ММП $(X_A)_2 <_{\text{rmse}}$ оценки по ММП $(X_A)_1$;

2. **для данных** $(X_A Y_A)_2$, $(X_A Y_A)$

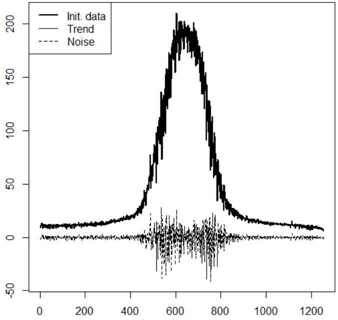
   оценки по ММП в $(X_A Y_A)_2 <_{\text{bias}}$ оценки по ММП $(X_A Y_A)_1$;
   оценки по ММП в $(X_A Y_A)_2 \approx_{\text{rmse}}$ оценки по ММП $(X_A Y_A)_1$;

3. **для данных** $(X_A Y_M)_1$, $(X_A Y_M)$

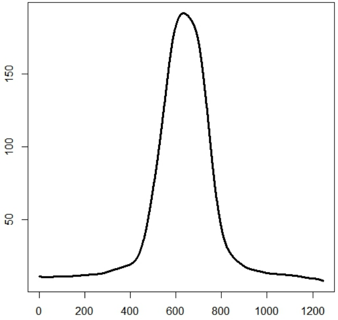
   оценки по ММП в $(X_A Y_M)_1 \approx_{\text{bias}}$ оценок по ВМНК $(X_A Y_M)_1 \approx 0$;
   оценки по ММП в $(X_A Y_M)_1 <_{\text{rmse}}$ оценок по ВМНК $(X_A Y_M)_1$.

Данные: измерения активности генов. Предполагается, что подчиняются модели $(X_A Y_M)$.

Реальные данные



Модель $(X_A Y_M)_1$
$\sigma_x^2 = 0.6$, $\sigma_y^2 = 0.02$, функция u



В приведенных ниже таблицах смещения не значимы.

Для параметра $\sigma_x^2 = 0.6$.

	bias	rmse
ОВМНК	$-1.7 \cdot 10^{-2}$	0.2
ОМП	$-4.6 \cdot 10^{-3}$	0.17
p-level		0.0004

Для параметра $\sigma_y^2 = 0.02$.

	bias	rmse
ОВМНК	$2 \cdot 10^{-4}$	0.003
ОМП	$-1.5 \cdot 10^{-5}$	0.002
p-level		0.0005

Таким образом, в работе были рассмотрены методы оценок неизвестных параметров для моделей (X_A), $(X_A Y_A)$, $(X_A Y_M)$.

1. Была доказана асимптотическая нормальность оценок по ВМНК в моделях $(X_A)_1$, $(X_A Y_A)_1$, $(X_A Y_M)_1$.
2. Были построены оценки по ММП в моделях $(X_A)_2$, $(X_A Y_A)_2$, $(X_A Y_M)_1$.
3. Построенные методы были реализованы на языке $\mathbf{R}$. На модельных примерах было продемонстрировано, что ошибки полученных оценок меньше, чем ошибки оценок, предложенных ранее.