Об оценке параметров нелинейной регрессии специального вида

Нечаева Мария Леонидовна, 522 группа

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

Научный руководитель: д.ф-м.н., проф. Ермаков С.М.
Рецензент: ст.научн.сотр. Солнцев В.Н.

Санкт-Петербург
2007г.

- Задача нелинейной регрессии с одной независимой переменной при отсутствии систематической ошибки:

$$y_j = \eta(x_j|\theta) + \varepsilon_j, \quad x_j \in X, \quad j = 0, \ldots, N,$$

  - $\eta(x|\theta)$ — нелинейная функция на $X \times \Theta$, задана с точностью до набора параметров $\theta \in \Theta$, $(\Theta \subseteq \mathbb{R}^m)$;
  - $\varepsilon_j, \quad j = 0, \ldots, N,$ — случайные ошибки измерений.

  Требуется оценить неизвестные параметры θ.

- Рассмотрим конкретную модель:

$$\eta(x|\theta) = \eta(x|\lambda, \omega, \alpha, \beta) = \sum_{i=1}^{p} e^{\lambda_i x} \left(\alpha_i \cos(\omega_i x) + \beta_i \sin(\omega_i x)\right),$$

$$\lambda = \{\lambda_i\}_{i=1}^p, \quad \omega = \{\omega_i\}_{i=1}^p, \quad \alpha_i = \{\alpha_i\}_{i=1}^p, \quad \beta = \{\beta_i\}_{i=1}^p,$$

$\theta = \{\lambda, \omega, \alpha, \beta\}$ — наборы параметров.

- Предполагается:
  - число слагаемых для $\eta(x|\theta)$: $p = 1, 2$;
  - в основном $x_j, \quad j = 0, \ldots, N,$ — равноотстоящие точки из отрезка $[0, 1]$;
  - $\varepsilon = (\varepsilon_0, \ldots, \varepsilon_N)^T$: $\varepsilon \sim N(0, \sigma^2 I), \quad (\sigma < \infty)$;
  - $\Theta = \mathbb{R}^{4p}$.

$$\hat{\theta} = \left\{\hat{\lambda},\ \hat{\omega},\ \hat{\alpha},\ \hat{\beta}\right\} = \arg\min_{\theta \in \Theta} F(\theta) \text{ — оценка МНК,}$$

$$\text{где} \quad F(\theta) = \sum_{j=0}^{N} \left[y_j - \eta(x_j|\theta)\right]^2 \text{ — целевая функция.}$$

- Задача:
  - исследовать нахождение оценок $\hat{\theta}$ несколькими численными методами;
  - рассмотреть зависимость ошибки в оценках от погрешности в измерениях и некоторые статистические свойства этих оценок;
  - сравнить некоторые результаты оценок методом “Гусеница”-SSA и МНК при условии отсутствия систематической погрешности.

- Для проведения численных экспериментов используется среда MATLAB.
- Рассматриваются 4 метода минимизации.
- Моделирование исходных данных $(p = 1)$:
  - выбор истинного набора параметров $\tilde{\theta}$;
  - вычисление $\eta(x_j|\tilde{\theta}) = \eta(x_j|\tilde{\lambda}, \tilde{\omega}, \tilde{\alpha}, \tilde{\beta}) = e^{\tilde{\lambda}x_j}\left(\tilde{\alpha}\cos(\tilde{\omega}x_j) + \tilde{\beta}\sin(\tilde{\omega}x_j)\right)$, где $x_j = jh, \quad h = \frac{1}{N}, \quad j = 0, \ldots, N$;
  - генерация случайного вектора $\varepsilon \sim N(0, I)$ и умножение на различные $\sigma = \sigma(\varepsilon)$;
  - вычисление $y_j = \eta(x_j|\tilde{\theta}) + \varepsilon_j, \quad j = 0, \ldots, N$.
- Начальное приближение для методов: $\theta_0 = \tilde{\theta}$.
- Выбор $\tilde{\theta}$:
  - несколько наборов, взятых произвольно;
  - фиксируем $\tilde{\alpha} = 1$, для $\tilde{\lambda}$, $\tilde{\omega}$, $\tilde{\beta}$ берем значения в вершинах куба $[0.5, 5]^3$ и случайные точки внутри него.

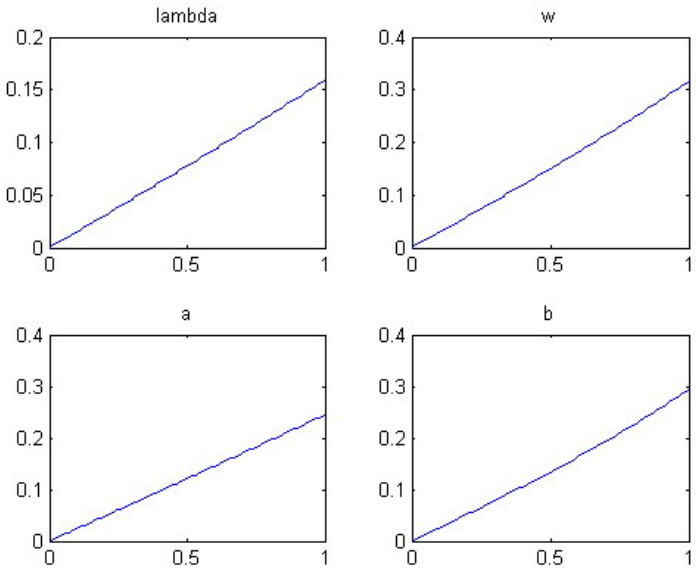
- **Эксперимент:** для каждого параметра построение зависимости оценки стандартного отклонения ошибки в оценках параметров $\hat{\sigma}(\theta) = \{\hat{\sigma}(\lambda), \hat{\sigma}(\omega), \hat{\sigma}(\alpha), \hat{\sigma}(\beta)\}$ от стандартного отклонения для случайной ошибки в измерениях $\sigma = \sigma(\varepsilon)$.

  Цель эксперимента: исследовать результаты оценки параметров несколькими методами и рассмотреть общий характер зависимости ошибки в оценках от ошибки в измерениях.

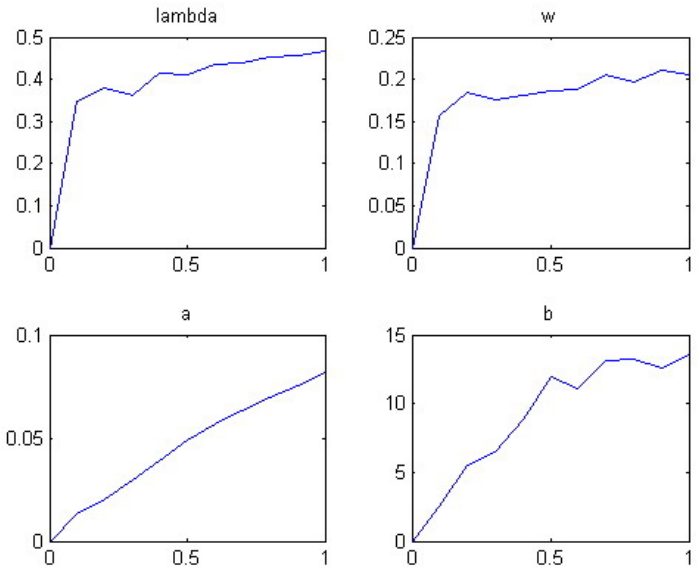
- Используемые значения:
  - фиксировано $N = 50$;
    $x_0 = 0$, $x_{50} = 1$, $x_j = jh$, $h = 1/N$, $j = 1, \ldots, N-1$;
  - значение $\sigma(\varepsilon)$ меняется от 0 до 1 с шагом $h = 0.01$, либо $h = 0.1$, и на эти значения умножается каждый смоделированный вектор случайных ошибок;
  - число повторений для вычисления оценки $\hat{\sigma}(\theta)$ $M = 20$.

Зависимость $\hat{\sigma}(\theta)$ от $\sigma(\varepsilon)$ для “хороших” и “плохих” параметров:

$\{\tilde{\lambda} = 3, \tilde{\omega} = 3, \tilde{\alpha} = 1, \tilde{\beta} = 1\}$



$\{\tilde{\lambda} = 5, \tilde{\omega} = 0.5, \tilde{\alpha} = 1, \tilde{\beta} = 0.5\}$



- H — матрица вторых производных целевой функции $F(\theta)$ по параметрам при $\sigma(\varepsilon) = 0$ в точке истинных значений $\tilde{\theta}$;
- λ_{min}, λ_{max} — минимальное и максимальное собственные числа матрицы H;
- $d = \lambda_{min}/\lambda_{max}$ — число обусловленности матрицы H.

Результат: у "плохих" наборов параметров число обусловленности много меньше по сравнению с "хорошими" наборами.

"хорошие" наборы					"плохие" наборы				
$\tilde{\lambda}$	$\tilde{\omega}$	$\tilde{\alpha}$	$\tilde{\beta}$	d	$\tilde{\lambda}$	$\tilde{\omega}$	$\tilde{\alpha}$	$\tilde{\beta}$	d
2	3	1	4	$0.34 \cdot 10^{-3}$	1	1	1	1	$0.12 \cdot 10^{-5}$
3	3	1	1	$0.39 \cdot 10^{-3}$	0.5	0.5	1	5	$0.53 \cdot 10^{-7}$
5	5	1	0.5	$0.27 \cdot 10^{-3}$	0.3	2	1	3	$0.32 \cdot 10^{-5}$
0.5	5	1	5	$0.72 \cdot 10^{-2}$	4.3	1.4	1	1.1	$0.45 \cdot 10^{-5}$
0.8	4.4	1	2.6	$0.13 \cdot 10^{-1}$	5	0.5	1	5	$0.15 \cdot 10^{-8}$
0.5	5	1	5	$0.72 \cdot 10^{-3}$	1.7	0.8	1	0.9	$0.29 \cdot 10^{-5}$

- При $\varepsilon \backsim N(0, \sigma^2 I)$ оценки МНК являются эффективными, асимптотически состоятельными и нормальными.
- **Эксперимент:** при различных значениях $\sigma(\varepsilon)$ и N для некоторых параметров
  - построены гистограммы одномерных распределений ошибок в оценках параметров;
  - вычислена оценка смещения оценок МНК;
  - оценена матрица ковариаций вектора погрешности в оценках.
- Используемые значения:
  - рассмотрено $N = 5,\ 10,\ 50$;
  - значение $\sigma(\varepsilon) = 0.01,\ 0.1,\ 0.2,\ 0.5,\ 1$;
  - фиксировано $M = 150$ — объем выборки оценок.

- В случае равноотстоящих точек x_j, $j = 0, \ldots, N$ функция $\eta(x|\theta)$ удовлетворяет уравнению:

$$\eta(x_j|\theta) = a\eta(x_{j-1}|\theta) + b\eta(x_{j-2}|\theta), \ j = 2, \ldots, N.$$

- Исходный временной ряд $Y = (y_0, \ldots, y_N)$.
- **Эксперимент:**
  метод "Гусеница"-SSA
  - выбираем L $(1 < L < N + 1)$; строим матрицу вложения $\mathbb{X}$;
  - λ_i — с. ч., U_i — с. в. матрицы $\mathbb{X}\mathbb{X}^{\mathrm{T}}$, $i = 1, \ldots, L$,
    $\lambda_1 \geq \ldots \geq \lambda_L \geq 0, \quad d = \max\{i : \lambda_i > 0\}$,
    $\mathbb{X} = \sum\limits_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$ — SVD-разложение матрицы $\mathbb{X}$;
  - берем матрицу $\hat{\mathbb{X}} = \sqrt{\lambda_1} U_1 V_1^{\mathrm{T}} + \sqrt{\lambda_2} U_2 V_2^{\mathrm{T}}$;
  - $\hat{Y} = (\hat{y}_0, \ldots, \hat{y}_N)$ — ряд, полученный после диагонального усреднения матрицы $\hat{\mathbb{X}}$.

  МНК
  - находим $\hat{\theta} = \{\hat{\lambda}, \hat{\omega}, \hat{\alpha}, \hat{\beta}\}$ — оценку $\theta = \{\lambda, \omega, \alpha, \beta\}$;
  - вычисляем $(\hat{\eta}_0, \ldots, \hat{\eta}_N)$:
    $\hat{\eta}_j = \eta(x_j|\hat{\theta}) = e^{\hat{\lambda} x_j} \left(\hat{\alpha} \cos(\hat{\omega} x_j) + \hat{\beta} \sin(\hat{\omega} x_j)\right)\big)$.

- Используемые значения:
  - фиксировано $N = 50$, $L = 25$;
  - значение $\sigma = \sigma(\varepsilon)$ меняется от 0 до 2 с шагом $h = 0.1$;
  - число повторений эксперимента $M = 20$.

- Для оценок $(\hat{y}_0^{i,\sigma}, \ldots, \hat{y}_N^{i,\sigma})$ и $(\hat{\eta}_0^{i,\sigma}, \ldots, \hat{\eta}_N^{i,\sigma})$ вычисляются:
  - отклонения от истинных значений в каждой точке $(j = 0, \ldots, N)$

$$res_j^\sigma S = \frac{1}{M} \sum_{i=1}^{M} \left(\hat{y}_j^{i,\sigma} - \eta(x_j|\tilde{\theta})\right),$$

$$res_j^\sigma M = \frac{1}{M} \sum_{i=1}^{M} \left(\hat{\eta}_j^{i,\sigma} - \eta(x_j|\tilde{\theta})\right);$$

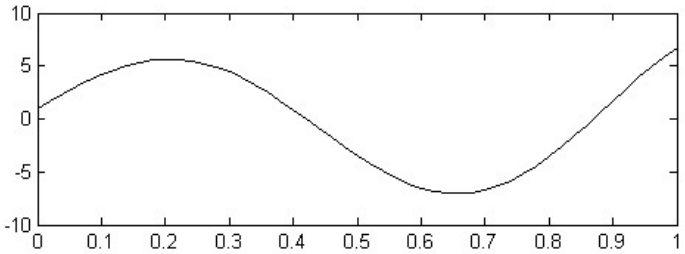
  - средняя сумма квадратов отклонений от истинных значений

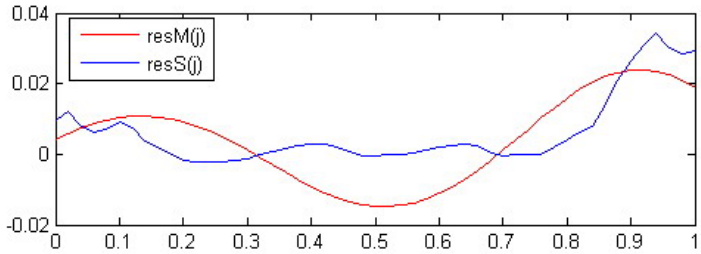
$$res^\sigma S = \frac{1}{M} \sum_{i=1}^{M} [\frac{1}{N} \sum_{j=1}^{N} \left(\hat{y}_j^{i,\sigma} - \eta(x_j|\tilde{\theta})\right)^2],$$

$$res^\sigma M = \frac{1}{M} \sum_{i=1}^{M} [\frac{1}{N} \sum_{j=1}^{N} \left(\hat{\eta}_j^{i,\sigma} - \eta(x_j|\tilde{\theta})\right)^2].$$

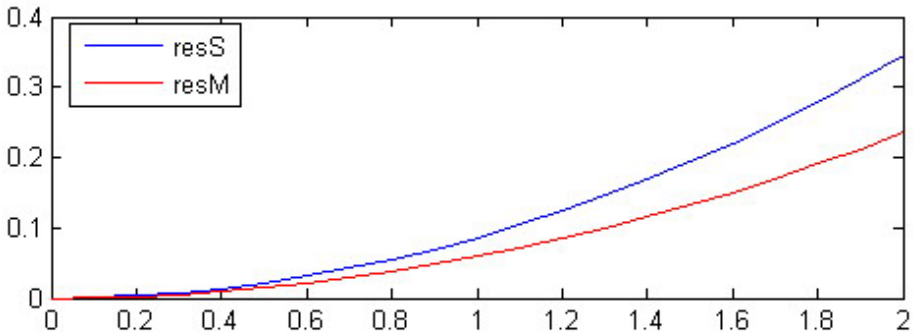
$\{\tilde{\lambda} = 0.5,\ \tilde{\omega} = 7,\ \tilde{\alpha} = 1,\ \tilde{\beta} = 5\}$

- вид ряда и отклонения в каждой точке $(\sigma(\varepsilon) = 1)$:

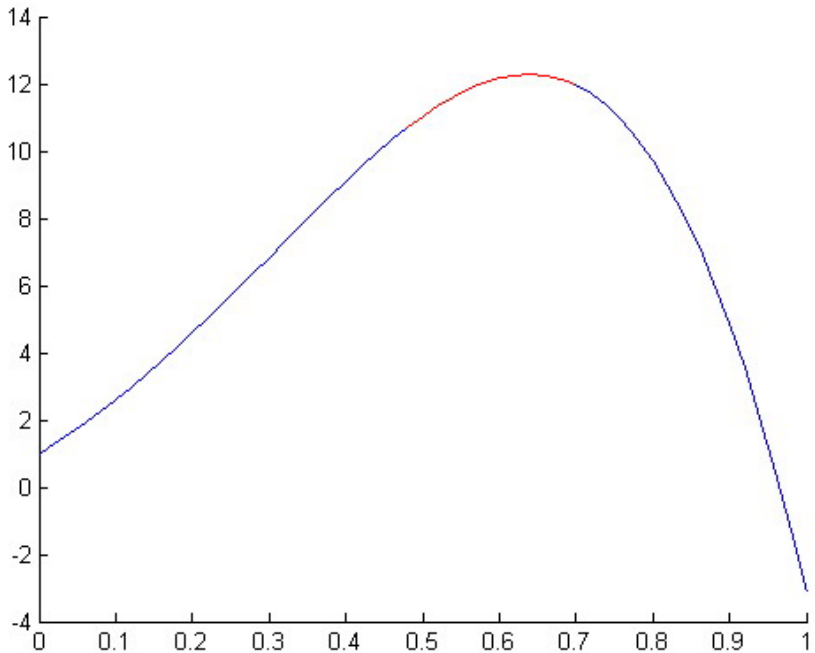




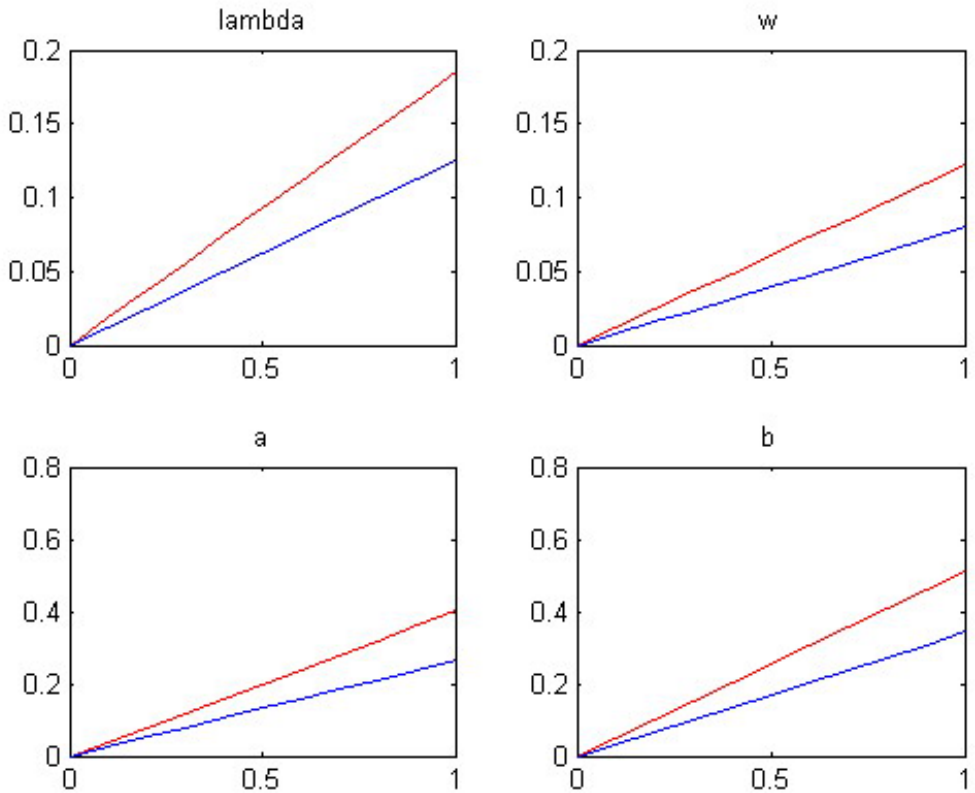
- зависимость средней суммы квадратов отклонений от истинных значений от $\sigma(\varepsilon)$:



- Оценки МНК можно строить, когда точки $x_j \in [a, b]$, $j = 0, \ldots, N$ не являются равноотстоящими.
- Исходные данные:
  на сетке из равноотстоящих точек $x_j \in [0, 1]$, $j = 0, \ldots, N$ "выкидываем" часть значений $x_p, \ldots, x_{p+k}$ (пропущенные наблюдения) и моделируем значения $y_0, \ldots, y_p, y_{p+k}, \ldots, y_N$.
- Вид ряда для $\{\tilde{\lambda} = 2, \tilde{\omega} = 3, \tilde{\alpha} = 1, \tilde{\beta} = 4\}$
  (выделен промежуток пропущенных точек, $p = 25$, $k = 10$):



Зависимость погрешности в оценке неизвестных параметров от ошибки в измерениях для модели без пропусков (синим) и при наличии пропущенных наблюдений (красным):



Результат: с увеличением числа пропущенных наблюдений ошибка растет.

Численные эксперименты в работе позволяют сделать выводы:

- методы, предлагаемые пакетом MATLAB, позволяют достаточно уверенно находить оценки МНК, когда матрица H хорошо обусловлена (число обусловленности не ниже порядка 10^{-3}), иначе требуется оптимизация методов;
- удалось проследить зависимость ошибок в оценках от ошибок в исходных данных; установлено, что оценки МНК нормальны при сравнительно небольшом числе наблюдений;
- при сравнении результатов отделения сигнала от шума методом “Гусеница”-SSA и восстановления регрессии с помощью МНК отмечено, что МНК дает меньшее отклонение от истинного значения, и это различие растет с ростом σ;
- когда из равноотстоящих точек часть наблюдений пропущены, методы также работают, но ошибка в оценках, безусловно, увеличивается и зависит от положения пропущенных точек.