Анализ структуры шума при разных типах ошибок измерений

Федоренко Кристина Андреевна, 522-я группа

Санкт-Петербургский Государственный Университет
Математико-механический факультет
Кафедра статистического моделирования

Научный руководитель — к.ф.-м.н. Н.Э. Голяндина
Рецензент — к.ф.-м.н. В.В. Некруткин

Санкт-Петербург
2013г.

Постановка задачи

Модели наблюдений

Рассмотрим ряд $F = (f_0, \ldots, f_{N-1})$, такой что

$$
\begin{aligned}
f_i &= u(x_i) + \delta_i, && (Y_A),\\
f_i &= u(x_i + \varepsilon_i), && (X_A),\\
f_i &= u(x_i + \varepsilon_i) + \delta_i, && (X_A Y_A),\\
f_i &= u(x_i)(1 + \delta_i), && (Y_M),\\
f_i &= u(x_i + \varepsilon_i)(1 + \delta_i), && (\mathbf{X_A Y_M}),
\end{aligned}
$$

где $u \in \mathcal{C}^2_{\mathcal{B}}$ – неизвестная функция на $\mathbb{R}$,
$\bar{x} = (x_0, \ldots, x_{N-1})$ — произвольный дискретный набор точек,
$\varepsilon_i \sim \mathrm{N}(0, \sigma_x^2)$, $\delta_i \sim \mathrm{N}(0, \sigma_y^2)$ — независимы в совокупности.

Задача: в каждой модели оценить параметры σ_x^2 и σ_y^2.
В данном случае сигнал неизвестен, но представляет интерес не он, а именно параметры σ_x^2 и σ_y^2.

План доклада

1. Сначала получим оценки, предположив, что значения функций $u(x)$ и $u'(x)$ известны.
2. Получим два вида оценок неизвестных параметров σ_x^2 и σ_y^2:
   - оценки, полученные с помощью сведения задачи к линейной регрессии;
   - оценки максимального правдоподобия.
3. Построим оценки значений функций $u(x)$ и $u'(x)$.
4. Исследуем свойства оценок параметров σ_x^2 и σ_y^2 при условии, что значения функций $u(x)$ и $u'(x)$ неизвестны.
5. Рассмотрим работу методов на реальных данных.

Используемые аппроксимации

Для моделей (X_A), $(X_A Y_A)$ и $(X_A Y_M)$ с ошибками в аргументе используем линеаризацию с помощью представления:

$$u(x+\gamma) = u(x) + \gamma u'(x) + \frac{\gamma^2}{2} u''(x+\theta\gamma),\ 0 < \theta < 1.$$

Например, для модели $(X_A Y_M)$ вместо

$$f_i = u(x_i + \varepsilon_i)(1+\delta_i),$$

будем рассматривать

$$g_i = (1+\delta_i)u(x_i) + \varepsilon_i(1+\delta_i)u'(x_i).$$

Замечание. Оценки строятся в модели ряда $G = (g_0, \ldots, g_{N-1})$, но для реальных данных в качестве исходного ряда будет использоваться ряд $F = (f_0, \ldots, f_{N-1})$.

Задача оценивания дисперсий ошибок в регрессионной постановке

Обозначим $w(x) = u'(x)$.
Представим вектор $(g_i - u(x_i))^2$ в виде:

$$(g_i - u(x_i))^2 = \mathbb{E}(g_i - u(x_i))^2 + r_i,$$
$$\mathbb{E}r_i = 0.$$

Модели (Y_A), (X_A), (Y_M) — простые регрессионные модели (оценки выписываются явно).
Модели (X_AY_A), (X_AY_M) — сложные регрессионные модели.

Пример простой модели (X_A):

$$\mathbb{E}(g_i - u(x_i))^2 = \sigma_x^2 w^2(x_i),$$
$$r_i = (\varepsilon_i^2 - \sigma_x^2)w^2(x_i).$$

Пример сложной модели (X_AY_M):

$$\mathbb{E}(g_i - u(x_i))^2 = \sigma_x^2(1 + \sigma_y^2)w^2(x_i) + \sigma_y^2 u^2(x_i),$$
$$r_i = (\varepsilon_i^2(1 + \delta_i)^2 - \sigma_x^2(1 + \sigma_y^2))w^2(x_i) + (\delta_i^2 - \sigma_y^2)u^2(x_i) + 2\varepsilon_i\delta_i(1 + \delta_i)w(x_i)u(x_i).$$

Задача оценивания дисперсий ошибок в регрессионной постановке

Рассмотрим регрессионное уравнение

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

где $Y = (y_0, \ldots, y_{N-1})$, $y_i = (g_i - u_i)^2$, $R = (r_0, \ldots, r_{N-1})$.

- Для модели (X_A)
  $\mathbf{X} = [X_1]$, $X_1 = (w^2(x_0), \ldots, w^2(x_{N-1}))^{\mathrm{T}}$, $B = \sigma_x^2$,
  $\mathbb{D}r_i = 2\sigma_x^4 w^4(x_i)$;
  Условие невырожденности матрицы $\mathbf{X}$:
  вектор из производных не нулевой. Например, не подходит $u(x) = c$.
- Для модели $(X_A Y_M)$
  $\mathbf{X} = [X_1 : X_2]$,
  $X_1 = (u^2(x_0), \ldots, u^2(x_{N-1}))^{\mathrm{T}}$, $X_2 = (w^2(x_0), \ldots, w^2(x_{N-1}))^{\mathrm{T}}$,
  $B = (\sigma_y^2, \sigma_x^2(1 + \sigma_y^2))^{\mathrm{T}}$,
  $\mathbb{D}r_i = 2\sigma_x^4(1 + 2\sigma_y^2)^2 w^4(x_i) + 2\sigma_y^4 u^4(x_i) + 4\sigma_x^2\sigma_y^2(1 + 3\sigma_y^2)u^2(x_i)w^2(x_i)$.
  Условие невырожденности матрицы $\mathbf{X}$:
  вектор из производных функции не пропорционален вектору из её значений. Например, не подходит $u(x) = \exp(\alpha x)$.

Отметим, что дисперсия компонент вектора R непостоянна и зависит от i.

Способы решения регрессионного уравнения

Рассмотрим способы решения регрессионного уравнения

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

Пусть $\mathbf{W} = \mathrm{C}\,\mathrm{diag}(\mathbb{D}r_0, \ldots, \mathbb{D}r_{N-1})$, $\mathrm{C} > 0$.

- Если $\mathbb{D}r_i > 0, i = 0, \ldots, N-1$, то следующая формула дает BLUE оценки параметров B

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

- Если $\mathbb{D}r_{k_0} = 0$ при каком-то k_0, то
  - можно рассмотреть регуляризацию матрицы $\mathbf{W}$: $\mathbf{W}_\lambda = \mathbf{W} + \lambda\mathbf{I}$;
  - или удалить наблюдения с нулевой дисперсией шума.

Проблема: $\mathbf{W} = \mathbf{W}_B$. Тогда для нахождения оценки получаем систему уравнений

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

Способы решения регрессионного уравнения: итерационная процедура

Если $\mathbf{W}$ зависит от параметров B, рассмотрим итерационное решение. Пусть τ и M — параметры остановки.

- Шаг 1. Выбор начального значения
  Выберем некоторое начальное значение $\widehat{B}_0$, $i = 0$.
- Шаг 2. Вычисление приближенных значений матрицы $\mathbf{W}$
  $\mathbf{W}_{(i)} = \mathbf{W}(\widehat{B}_{(i)})$.
- Шаг 3. Нахождение оценок

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

- Шаг 4. Критерий остановки
  Если $||\widehat{B}_{(i)} - \widehat{B}_{(i+1)}|| < \tau$ или $i + 1 = M$, процедура заканчивается и результатом ее работы является $\widehat{B}_{(i+1)}$.
  Иначе $i = i + 1$, переход к шагу 2.

Регрессионное решение для рассматриваемых моделей

- Для простых моделей (Y_A), (X_A), (Y_M):
  $\mathbb{D}r_i = c(B)q_i$, q_i не зависит от B, тогда $\mathbf{W} = \mathrm{diag}(q_0, \ldots, q_{N-1})$ — не зависит от B, следовательно решение выписывается явно, при условии, что $\mathbf{W}$ не вырождена.

  Например, для модели (X_A):

$$\hat{\sigma}_x^2 = \frac{1}{N} \sum_{i=0}^{N-1} \frac{(f_i - u(x_i))^2}{w(x_i)}$$

  и условие невырожденности матрицы $\mathbf{W}$: $w(x_i) \neq 0$ для всех i.

- В сложных моделях $(X_A Y_A)$ и $(X_A Y_M)$ $\mathbf{W} = \mathbf{W}(B)$, поэтому оценим параметры по итерационной процедуре.
  Условие невырожденности матрицы $\mathbf{W}$ в $(X_A Y_M)$: $w(x_i)$ и $u(x_i)$ не равны нулю одновременно.

Оценки максимального правдоподобия

- В моделях (Y_A), (X_A), (Y_M) ОМП выписываются явно и полностью совпадают с регрессионными BLUE оценками.
- Рассмотрим модель $(X_A Y_A)$, $\theta = (\theta_1, \theta_2)$, где $\theta_1 = \sigma_x^2$, $\theta_2 = \sigma_y^2$,

$$g_i = u(x_i) + w(x_i)\varepsilon_i + \delta_i,$$

$$g_i \sim \mathrm{N}(u(x_i), \theta_1 w^2(x_i) + \theta_2),$$

$$\mathcal{L}(g_i \,|\, \theta) = \left(\prod_{i=0}^{N-1} \frac{1}{\sqrt{2\pi(\theta_1 w^2(x_i) + \theta_2)}} \right) \exp\left(- \sum_{i=0}^{N-1} \frac{(g_i - u(x_i))^2}{2(\theta_1 w^2(x_i) + \theta_2)} \right).$$

Функция правдоподобия выписывается, ОМП находятся численным методом.

- Рассмотрим модель зашумления $(X_A Y_M)$
$g_i = (1 + \delta_i)u(x_i) + \varepsilon_i(1 + \delta_i)u'(x_i)$.
Функцию правдоподобия не получается выписать аналитически.

Описание примера

Пример: $u(x) = x^2$, $x_i = i$, $i = 1, \ldots, 200$.

Модель $(X_A Y_M)$:

$$f_i = (x_i + \varepsilon_i)^2(1 + \delta_i), i = 1, \ldots, 200,$$
$$g_i = x_i^2(1 + \delta_i) + 2x_i\varepsilon_i(1 + \delta_i), i = 1, \ldots, 200.$$

$\varepsilon_i \sim \mathrm{N}(0, \sigma_x^2)$ и $\delta_i \sim \mathrm{N}(0, \sigma_y^2)$ — независимы в совокупности.

В модели $(X_A Y_M)$ не умеем строить ОМП, следовательно, получим оценки σ_x^2, σ_y^2 по итерационной процедуре.

Предположим, что $\sigma_x^2 \in [0, 1]$ и $\sigma_y^2 \in [0, 1]$, поэтому начальные значения возьмем равномерно распределенными на $[0, 1]$.

Характеристики оценок (смещение и стандартное отклонение) получены по 100 реализациям исходного ряда.

Точность итерационных оценок

Для модели $(X_A Y_M)$ по исходному ряду G сравним следующие оценки:

- итерационные: $\widehat{B} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}_{\widehat{B}}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}_{\widehat{B}}^{-1}Y$;
- BLUE оценки, формально предположив, что $\mathbb{D}r_i$ известны: $\widehat{B} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}^{-1}Y$.

Таблица: $\widehat{B} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}^{-1}Y$ для модели $(X_A Y_M)$ по G

σ_x^2	σ_y^2	bias of $\widehat{\sigma}_x^2$	sd $\widehat{\sigma}_x^2$	bias of $\widehat{\sigma}_y^2$	sd $\widehat{\sigma}_y^2$
0.25	1e-04	1.65e-03	4.8e-02	0e+00	2.1e-05
0.49	4e-04	-1.01e-04	1.1e-01	1e-05	8.1e-05
1.00	9e-04	-2.82e-02	2.3e-01	1e-05	1.5e-04

Таблица: $\widehat{B} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}_{\widehat{B}}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}_{\widehat{B}}^{-1}Y$ для модели $(X_A Y_M)$ по G

σ_x^2	σ_y^2	bias of $\widehat{\sigma}_x^2$	sd $\widehat{\sigma}_x^2$	bias of $\widehat{\sigma}_y^2$	sd $\widehat{\sigma}_y^2$
0.25	1e-04	1.63e-03	4.9e-02	0e+00	2.2e-05
0.49	4e-04	2.2e-04	1.04e-01	1e-05	8.3e-05
1.00	9e-04	-2.4e-02	2.3e-01	1e-05	1.5e-04

Результаты итерационной процедуры близки к оптимальным оценкам.

Влияние линеаризации модели

Сравним итерационные оценки, полученные по исходным данным и по линеаризованной модели.

Таблица: $\widehat{B} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}_{\widehat{B}}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}_{\widehat{B}}^{-1}Y$ для модели $(X_A Y_M)$ по G

σ_x^2	σ_y^2	bias of $\widehat{\sigma}_x^2$	sd $\widehat{\sigma}_x^2$	bias of $\widehat{\sigma}_y^2$	sd $\widehat{\sigma}_y^2$
0.25	1e-04	1.63e-03	4.9e-02	0e+00	2.2e-05
0.49	4e-04	2.2e-04	1.04e-01	1e-05	8.3e-05
1.00	9e-04	-2.4e-02	2.3e-01	1e-05	1.5e-04

Таблица: $\widehat{B} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}_{\widehat{B}}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}_{\widehat{B}}^{-1}Y$ для модели $(X_A Y_M)$ по F

σ_x^2	σ_y^2	bias of $\widehat{\sigma}_x^2$	sd $\widehat{\sigma}_x^2$	bias of $\widehat{\sigma}_y^2$	sd $\widehat{\sigma}_y^2$
0.25	1e-04	5.7e-03	5.2e-02	0e+00	3.3e-05
0.49	4e-04	-1.8e-02	1.0e-01	1e-05	7.3e-05
1.00	9e-04	-2.7e-02	2.1e-01	-3e-05	1.4e-04

Таблицы демонстрируют, что линеаризация незначительно влияет на результат.

Оценивание тренда и его производной

Пусть $x_i = i$. Так как $\mathbb{E}g_i = u(x_i)$, следовательно $(u_0, \ldots, u_{N-1})$ можно рассмотреть как тренд ряда F.

Оценивание тренда $u(x_i)$ с помощью метода «Гусеница».

Рассмотрим ряд $F = (f_0, \ldots, f_{N-1})$ длины $N > 2$.

- **Шаг 1: Вложение**
  Выберем длину окна $L : 1 < L < N$. Процедура вложения образует $K = N - L + 1$ векторов вложения
  $Z_i = (f_{i-1}, \ldots, f_{i+L-2})^{\mathrm{T}}, \quad 1 \le i \le K.$
  $\mathbf{Z} = [Z_1 : \ldots : Z_K]$ — траекторная матрица ряда F.
- **Шаг 2: Сингулярное разложение**
  $\mathbf{Z} = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}}$.
- **Шаг 3: Группировка и диагональное усреднение**
  Пусть $F = T + N$, ранг T равен $r < d$, $I = \{1, \ldots, r\}$.
  $Z_I = \sum_{i=1}^{r} \sqrt{\lambda_i} U_i V_i^{\mathrm{T}} \longrightarrow \widetilde{F}$, $\widetilde{F}$ оценивает T.

Оценивание производной $u'(x_i)$ по формуле

$$\tilde{w}(x_i) = \frac{-\tilde{f}_{i+2} + 8\tilde{f}_{i+1} - 8\tilde{f}_{i-1} + \tilde{f}_{i-2}}{12h}, i = 2, \ldots, N-3, \ h = 1.$$

Описание примера

Пример: $u(x) = x^2$, $x_i = i$, $i = 1, \ldots, 200$.

Модель $(X_A Y_M)$:

$$f_i = (x_i + \varepsilon_i)^2 (1 + \delta_i), i = 1, \ldots, 200,$$
$$g_i = x_i^2 (1 + \delta_i) + 2 x_i \varepsilon_i (1 + \delta_i), i = 1, \ldots, 200.$$

$\varepsilon_i \sim \mathrm{N}(0, \sigma_x^2)$ и $\delta_i \sim \mathrm{N}(0, \sigma_y^2)$ — независимы в совокупности.

Значения функции $u(x)$ оценим методом «Гусеница» с длиной окна $L = 100$ и $r = 3$.

В модели $(X_A Y_M)$ не умеем строить ОМП, следовательно, получим оценки σ_x^2, σ_y^2 по итерационной процедуре.

Предположим, что $\sigma_x^2 \in [0, 1]$ и $\sigma_y^2 \in [0, 1]$, поэтому начальные значения возьмем равномерно распределенными на $[0, 1]$.

Характеристики оценок (смещение и стандартное отклонение) получены по 100 реализациям исходного ряда.

Сравнение результатов

Сравним результаты для случаев известных и неизвестных значений тренда и его производной. Исходный ряд — F.

Таблица: Итерационное решение для модели $(X_A Y_M)$ при известном $u(x)$

σ_x^2	σ_y^2	bias of $\widehat{\sigma}_x^2$	sd $\widehat{\sigma}_x^2$	bias of $\widehat{\sigma}_y^2$	sd $\widehat{\sigma}_y^2$
0.25	1e-04	5.7e-03	5.2e-02	-2.5e-06	2.1e-05
0.49	4e-04	-6.8e-03	1.3e-01	9.6e-06	7.2e-05
1.00	9e-04	2.1e-02	2.3e-01	-1.3e-05	1.8e-04

Таблица: Итерационное решение для модели $(X_A Y_M)$ при неизвестном $u(x)$

σ_x^2	σ_y^2	bias of $\widehat{\sigma}_x^2$	sd $\widehat{\sigma^2}_x$	bias of $\widehat{\sigma^2}_y$	sd $\widehat{\sigma}_y^2$
0.25	1e-04	-1.1e-01	5.6e-02	2.4e-05	2.6e-05
0.49	4e-04	7.2e-02	2.0e-01	-6.5e-05	7.7e-05
1.00	9e-04	-6.1e-02	2.8e-01	-9.7e-05	1.7e-04

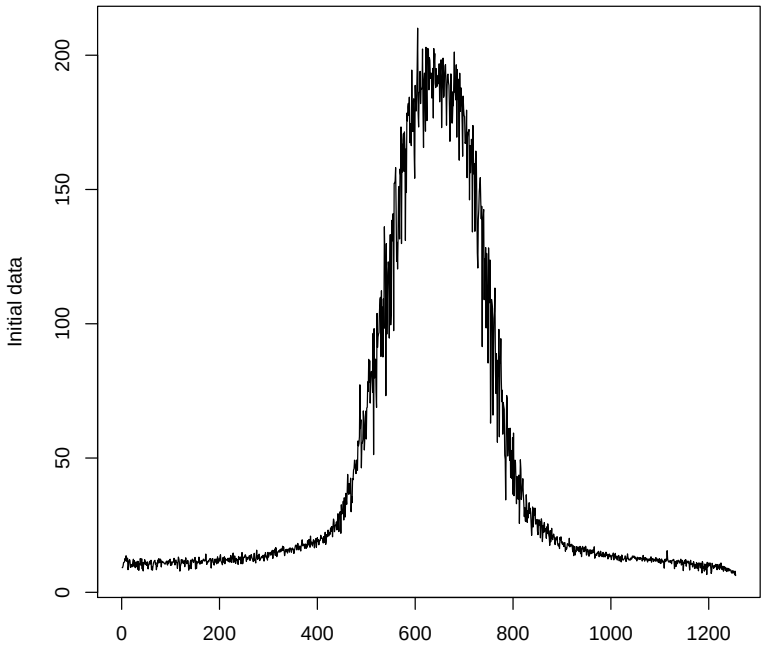
Из таблиц следует, что оценивание $u(x)$ и $w(x)$ увеличивает смещение оценок, однако оценки остаются приемлемыми.

Описание задачи

Данные представляют собой измерения активности гена kruppel у зародышей дрозофил. Ошибка в данных порождена двумя причинами:

- Ошибкой определения пространственного расположения ядра гена;
- Не детерминированной природой активности гена.

Задача: оценить изменчивость активности гена при условии, что в наблюдениях присутствует ошибка определения пространственного расположения ядра гена.



Дан ряд длины $N = 1255$, модель измерений которого предположительно удовлетворяет модели $(X_A Y_M)$

$f_i = u(x_i + \varepsilon_i)(1 + \delta_i)$.

Схема исследования

- Шаг 1: Выделение сигнала
  Применим к вектору F алгоритм метода “Гусеница” с длиной окна $L = 100$, восстановим тренд по первой компоненте.
- Шаг 2: Вычисление оценки производной тренда
- Шаг 3: Вычисление оценок параметров σ_x^2 и σ_y^2 в модели $(X_A Y_M)$ по данным на интервале от 470 до 800, убрав тем самым значения с маленькими шумом для улучшения свойств оценок.
- Шаг 4: Проверка результата модели

  Сравнение модельной дисперсии шума $\widehat{\sigma}_x^2(1+\widehat{\sigma}_y^2)\widehat{w}^2(x_i) + \widehat{\sigma}_y^2\widehat{u}^2(x_i)$ и дисперсии шума как тренда квадратов остатков $(f_i - u(x_i))^2$.
- Шаг 5: Вычисление точности оценок с помощью бутстреп-метода

Оценивание тренда

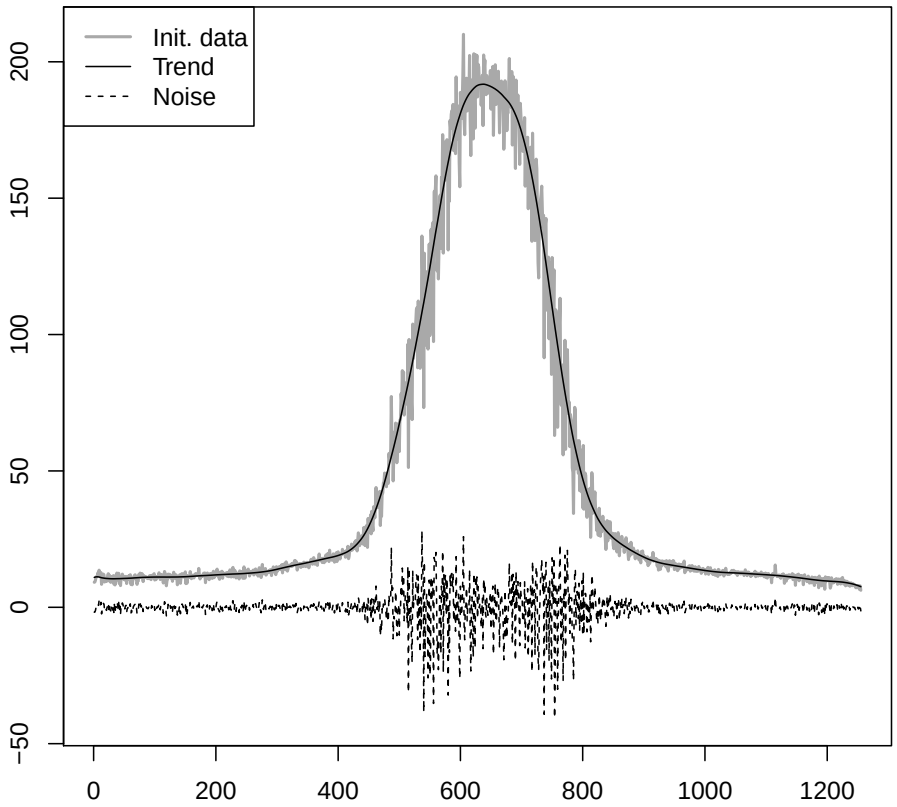


Рис.: Исходные данные, тренд, остатки $L = 100$.

Оценки дисперсий и проверка адекватности результата

Оценки параметров по итеративному регрессионному алгоритму:
$\widehat{\sigma}_x^2 = 0.6$, $\widehat{\sigma}_y^2 = 0.002$

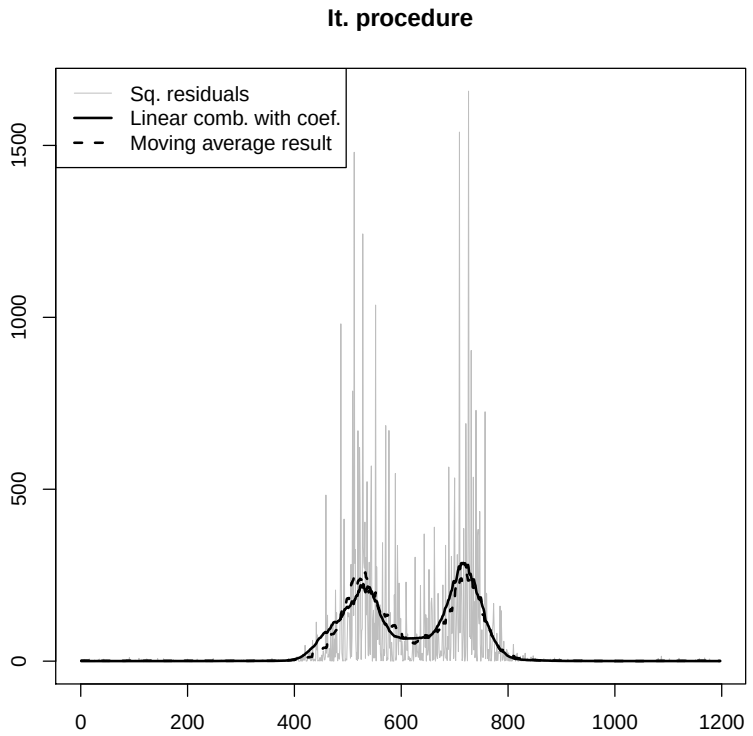


Рис.: $(f_i - \widehat{u}(x_i))^2$, скользящее среднее с окном 50 (пунктир) и модельная дисперсия шума $\widehat{\sigma}_x^2(1 + \widehat{\sigma}_y^2)\widehat{w}^2(x_i) + \widehat{\sigma}_y^2\widehat{u}^2(x_i)$

Бутстреп-процедура для оценивания точности оценок

Для вычисления характеристик полученных оценок применим к данным следующую bootstrap-процедуру:

- выделим тренд методом «Гусеница» с длиной окна $L = 100$, тренд восстановим по первой компоненте;
- получим оценки σ_x^2 и σ_y^2 в модели $(X_A Y_M)$;
- промоделируем ряд G для этой модели 100 раз, оценим σ_x^2 и σ_y^2 для каждой реализации ряда, получим выборку из оценок объема 100,
  - предполагая что тренд известен;
  - каждый раз оценивая тренд с помощью метода «Гусеница» с теми же параметрами;
- вычислим средние и стандартные отклонения оценок.

Таблица: Результаты bootstrap-процедуры

	mean of $\widehat{\sigma}_x^2$	mean of $\widehat{\sigma}_y^2$	sd $\widehat{\sigma}_x^2$	sd $\widehat{\sigma}_y^2$
тренд известен	0.59	0.002	7.0e-02	2.6e-04
тренд оценивается				

Результаты

1. Были рассмотрены пять моделей данных с ошибками разной структуры, в которых интерес представляют дисперсии шумов, входящих к модель.

2. Для каждой модели были построены оценки дисперсий шумов разными методами, методом максимального правдоподобия и с помощью сведения к регрессионной задаче.

3. На ряде примеров было проведено сравнение методов.

4. Разработанные методы были применены к реальным данным.