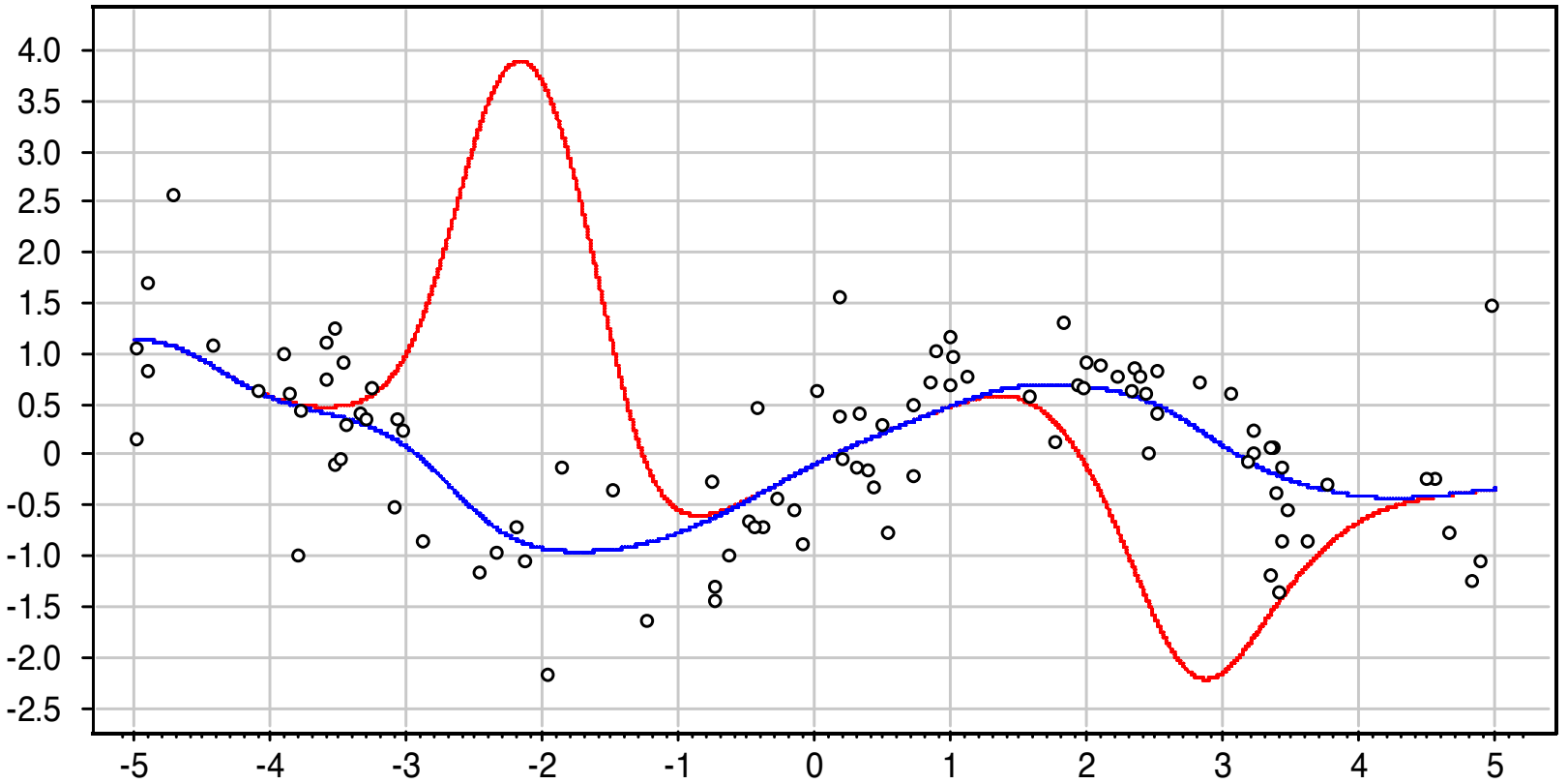
Проблема выбросов (эксперимент на синтетических данных)

$\ell = 100$, $h = 1.0$, гауссовское ядро $K(r) = \exp\left(-2r^2\right)$

Две из 100 точек — выбросы с ординатами $y_i = 40$ и -40

Синяя кривая — выбросов нет



Проблема выбросов и локально взвешенное сглаживание

Проблема выбросов: большие случайные ошибки в значениях y_i сильно искажают оценку Надарая–Ватсона

$$a_h(x; X^\ell) = \frac{\sum_{i=1}^{\ell} y_i w_i(x)}{\sum_{i=1}^{\ell} w_i(x)}, \qquad w_i(x) = K\left(\frac{\rho(x, x_i)}{h}\right).$$

Идея:
чем больше величина невязки $\varepsilon_i = \left|a_h\left(x_i; X^\ell \setminus \{x_i\}\right) - y_i\right|$,
тем меньше должен быть вес i-го объекта $w_i(x)$.

Эвристика:
домножить веса $w_i(x)$ на коэффициенты $\gamma_i = \tilde{K}(\varepsilon_i)$,
где $\tilde{K}(\varepsilon)$ — ещё одно ядро, вообще говоря, отличное от $K(r)$.

Рекомендация:
использовать квартическое ядро $\tilde{K}(\varepsilon) = K_Q\left(\frac{\varepsilon}{6\,\mathrm{med}\{\varepsilon_i\}}\right)$,
где $\mathrm{med}\{\varepsilon_i\}$ — медиана множества значений ε_i.

Алгоритм LOWESS (LOcally WEighted Scatter plot Smoothing)

Вход: X^ℓ — обучающая выборка;
Выход: коэффициенты γ_i, $i = 1, \ldots, \ell$;

1: инициализация: $\gamma_i := 1$, $i = 1, \ldots, \ell$;
2: **повторять**
3: **для всех** объектов $i = 1, \ldots, \ell$
4: вычислить оценки скользящего контроля:

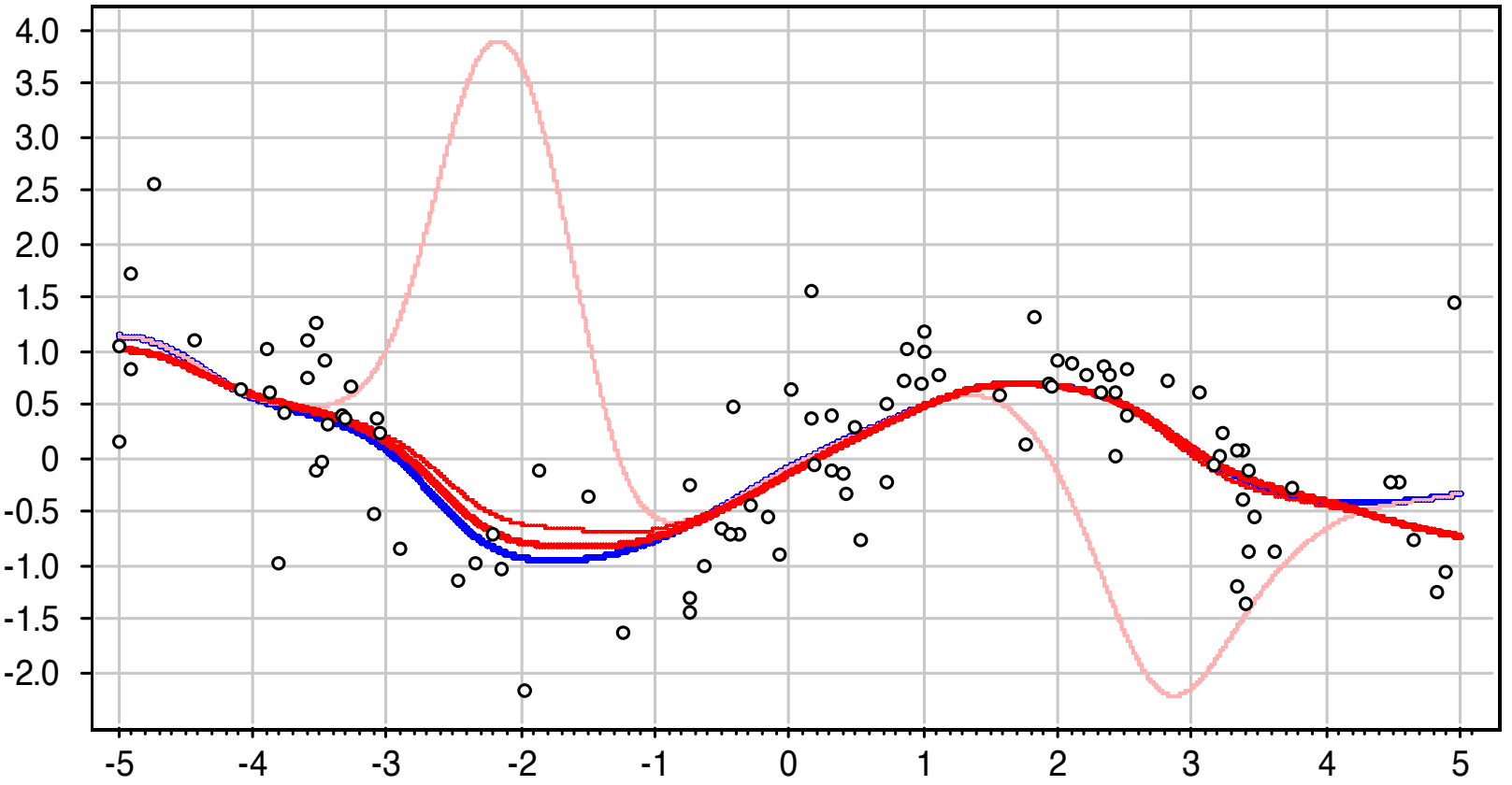
$$a_i := a_h\big(x_i; X^\ell \setminus \{x_i\}\big) = \frac{\sum\limits_{j=1,\, j\neq i}^{\ell} y_j \gamma_j K\left(\frac{\rho(x_i,x_j)}{h(x_i)}\right)}{\sum\limits_{j=1,\, j\neq i}^{\ell} \gamma_j K\left(\frac{\rho(x_i,x_j)}{h(x_i)}\right)};$$

5: **для всех** объектов $i = 1, \ldots, \ell$
6: $\gamma_i := \tilde{K}\big(|a_i - y_i|\big)$;
7: **пока** коэффициенты γ_i не стабилизируются;

Пример работы LOWESS на синтетических данных

$\ell = 100$, $h = 1.0$, гауссовское ядро $K(r) = \exp\left(-2r^2\right)$
Две из 100 точек — выбросы с ординатами $y_i = 40$ и -40

В данном случае LOWESS сошёлся за 2–3 итерации:



Резюме

- В статистике методы, устойчивые к нарушениям модельных предположений о данных, называются *робастными*. Мы рассмотрели простой робастный метод, устойчивый к наличию небольшого числа выбросов.
- В этом методе происходит обучение весов объектов.