

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ  
ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ  
імені В. Н. Каразіна

**Кафедра хімічного матеріалознавства**

УДК 544.169+544.15+519.237.5



*До захисту допускаю*  
Завідувач кафедри

«12» травня 2023 р. д. х. н., проф. Олександр КОРОБОВ

**НЕЛІНІЙНІ РЕГРЕСІНІ МОДЕЛІ В КВАНТОВІЙ ХІМІЇ**

Кваліфікаційна робота магістра  
II курсу хімічного факультету  
**ДЯЧЕНКО ЮЛІЇ ВІКТОРІВНИ**

Науковий керівник

д. х. н., професор



Володимир ІВАНОВ

ХАРКІВ 2023

## РЕФЕРАТ

Робота містить 44 сторінок, 2 розділи, 8 ілюстрацій, 5 таблиць, 65 рівнянь, 1 додаток та 40 джерел.

В даній роботі розроблено програми на скриптовій мові Python. В програмах реалізовано кілька методів нелінійної апроксимації (метод найменших квадратів, метод ядерної регресії). Проведено тестування та апроксимацію поверхні потенціальної енергії двохатомних молекул  $\text{BH}$  та  $\text{HF}$ . Ефективність апроксимації описано за допомогою ряду параметрів як то коефіцієнт детермінації розрахований за процедурою *leave-one-out* (LOO) —  $Q^2$ , індекси Джина (*Gini*). Запропоновано кілька нових індексів: *curve*,  $\text{inf}(c)$ . Робота містить інформацію про отримання апроксимуючої функції при моделюванні кривої.

Дана робота може бути використана для знаходження спектральних характеристик, як то енергії коливальних станів, енергії переходів та їх інтенсивності, для яких ці характеристики ще недостатньо досліджені. Також робота може бути використана як набір тестових даних для опису потенціальної функції малих молекул.

*Об'єкт дослідження:* Методи апроксимації нелінійних функцій, нелінійний метод найменших квадратів.

*Мета роботи:*

1. Аналіз різноманітних нелінійних методів апроксимації.
2. Застосування нелінійних методів апроксимації для простих систем.
3. Розробка програмних засобів для нелінійної апроксимації поверхні потенціальної енергії для малих молекул.

КЛЮЧОВІ СЛОВА: PYTHON, LEAVE-ONE-OUT (LOO), НЕЛІНІЙНІ МЕТОДИ АПРОКСИМАЦІЇ, ЯДЕРНА РЕГРЕСІЯ.

## ABSTRACT

The work contains 44 pages, 2 chapters, 8 figures, 5 tables, 65 formulas, 1 chart and 40 references.

In this work, a program was developed in the Python scripting language. The program implements several methods of nonlinear approximation (the method of least squares, the method of nuclear regression). The approximation of the potential energy surface of diatomic BH and HF molecules was carried out using the program. The effectiveness of the approximation is described using a number of parameters such as the coefficient of determination calculated by the leave-one-out (LOO) procedure — Q2, Gini indices. Several new indices are proposed: *curve*, *Inf(c)*. The work contains information on obtaining an approximating function when modeling a curve.

This work can be used to find spectral characteristics, such as the energies of vibrational states, the energy of transitions and their intensities, for which these characteristics have not yet been sufficiently studied. Also, the work can be used as a set of test data for describing the potential energy of small molecules.

*Object of research:* Methods of approximation of nonlinear functions, nonlinear least squares method.

*The purpose of the work:*

1. Analysis of various nonlinear approximation methods.
2. Application of nonlinear approximation methods for simple systems.
3. Development of software for nonlinear approximation of the potential energy surface for small molecules.

KEY WORDS: PYTHON, LEAVE-ONE-OUT (LOO), NON-LINEAR APPROXIMATION METHODS, KERNEL REGRESSION.

## ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ ТА ТЕРМІНІВ.....	5
ВСТУП.....	6
1 ЛІТЕРАТУРНИЙ ОГЛЯД.....	8
1.1 Метод ядерної регресії .....	8
1.2 Методи розрахунку потенціальних кривих двохатомних молекул.....	10
1.3 Методи непараметричної регресії.....	13
1.4 Методи апроксимації нелінійних функцій.....	14
2 РОЗРАХУНКОВА ЧАСТИНА.....	19
2.1 Апроксимація нелінійних функцій методом ядерної регресії.....	19
2.2 Тестові розрахунки методом ядерної регресії.....	22
2.3 Апроксимація потенціальної енергії функцією Морзе.....	27
2.4 Регресійні моделі опису потенціальних кривих двохатомних молекул.....	29
2.4.1 Молекула ВН.....	29
2.4.2 Молекула FH.....	32
ВИСНОВКИ.....	35
ПЕРЕЛІК ПОСИЛАНЬ .....	36
ДОДАТКИ.....	39

## ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ ТА ТЕРМІНІВ

LOO – процедура leave-one-out

ППЕ – поверхня потенціальної енергії

pdf – функція щільності ймовірності

pmf – функція маси ймовірності

МНК – метод найменших квадратів

МЯР – метод ядерних регресій

$D$ ,  $\beta$  та  $r_e$  – параметри функції

$r$  – між'ядерна відстань

$V(r_k)$  – енергія системи на відстані  $r_k$

$\hat{V}_k$  – теоретичне значення функції

$V_k$  – експериментальне значення функції

Gini – індекс Джині

$|curve|$  – середня кривизна

$y_j$  – теоретичне значення

$y(x)$  – експериментальне значення

$\bar{y}$  – середнє теоретичне значення

$\bar{y}(x)$  – середнє експериментальна значення

## ВСТУП

Квантово-хімічні методи обчислення електронної та просторової будови молекул, їх найрізноманітніших фізико-хімічних властивостей стали в даний час невід'ємною частиною повноцінного наукового дослідження. Цьому чимало сприяв різкий сплеск комп'ютерних можливостей і розробка високоефективних алгоритмів і потужних обчислювальних програм, в основі яких лежать як напівемпіричні методи квантової хімії (реалізовані, наприклад, в програмному комплексі MOPAC), так і неемпіричні, або *ab initio* методи (програми Gaussian 94 (98, 03), GAMESS, NWChem і ін.).

Сучасні квантово хімічні методи дозволяють, зокрема, розраховувати енергію і різного роду фізико-хімічні характеристики молекулярної системи в стаціонарному стані як на основні поверхні потенційної енергії (ППЕ), так і отримати теоретичну інформацію про енергію і властивості збуджених станах молекул.

Особливої актуальності набули квантово хімічні розрахунки електронної та просторової структури реагентів, продуктів, інтермедіатів і перехідних станів в хімічних реакціях, що дозволяє пролити світло на механізми хімічних реакцій. Слід особливо відзначити, що квантово хімічні розрахунки є єдиним джерелом прямої інформації про структуру та енергетику перехідних станів на поверхні потенційної енергії системи, що, взагалі кажучи, дозволяє прогнозувати можливі напрямки хімічної реакції. В даний час інтенсивно проводяться квантово хімічні дослідження в галузі вивчення динаміки хімічних реакцій на відповідних поверхнях потенційної енергії.

Одним із найважливіших завдань будь-якого наукового аналізу є побудова моделей, щоб представити зв'язок між змінними, залученими до аналізу. Часто це співвідношення включає змінний результат, який залежить від набору параметрів в систематичний спосіб. Більшість публікацій у статистичній літературі припускають, що систематичний зв'язок є лінійним у конкретних параметрах і моделях побудови відповідно. Однак багато цікавих проблем є нелінійними за своєю природою.

*Актуальність:*

Сучасні квантово-хімічні методи дозволяють з високою точністю відтворити спектральні характеристики малих молекул. Таких розрахункових даних потребують різні хімічні та фізичні дисципліни, зокрема астрохімія. Однак, для проведення високоточних розрахунків електронно-коливальних станів потрібний точний опис поверхні потенціальної енергії молекул та її аналітичне представлення.

*Мета роботи:*

1. Розробити програмні засоби апроксимації нелінійних функцій (поверхні потенціальної енергії) зокрема на основі методу ядерної регресії та нелінійного методу найменших квадратів.
2. Проаналізувати нелінійні методи апроксимації на прикладі модельних задач та деяких двохатомних молекул.

## 1 ЛІТЕРАТУРНИЙ ОГЛЯД

### 1.1 Методи ядерної регресії

У статистиці, особливо в Баєсовій статистиці, ядро функції густини ймовірності (probability density function, PDF) грає значну роль. Зазначають, що параметри статистичного оцінювання (наприклад в побудові регресійної моделі) цілком можуть бути функціями параметрів PDF. При цьому коефіцієнти нормалізації розподілу ймовірностей в багатьох ситуаціях непотрібні. Наприклад, у вибірці псевдовипадкових чисел більшість алгоритмів вибірки ігнорують коефіцієнт нормалізації. Крім того, в Баєсовому аналізі спряжених попередніх розподілів, коефіцієнти нормалізації зазвичай ігноруються під час обчислень, і розглядається лише ядро. Наприкінці розрахунків перевіряється форма ядра, і якщо вона відповідає відомому розподілу, коефіцієнт нормалізації можна відновити.

Прикладом є нормальний розподіл. Його функція щільності ймовірності:

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (1.1)$$

Та пов'язане ядро:

$$p(x|\mu, \sigma^2) \propto e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (1.2)$$

Множник перед експонентою пропускається, навіть якщо він містить параметр  $\sigma^2$ . У непараметричній статистиці ядро — це вагова функція, яка використовується в методах оцінювання. Ядра використовуються в оцінці щільності, для оцінки функцій щільності випадкових змінних або в регресії ядра для оцінки умовного очікування випадкової змінної. Ядра також використовуються в часових рядах під час використання періодограми для оцінки спектральної щільності, де вони відомі як віконні функції. Додаткове використання полягає в оцінці змінної в часі інтенсивності для точкового процесу де функції (ядра) згорнуті з даними часових рядів.

Зазвичай ширину ядра також необхідно вказувати під час виконання непараметричної оцінки.



Отже Ядро — це невід’ємна дійсна інтегрована функція  $K$ . Для більшості застосувань бажано визначити функцію, яка могла б задовольняла двом додатковим вимогам:

1) Нормалізація:

$$\int_{-\infty}^{+\infty} K(u) du = 1; \quad (1.3)$$

2) Симетрія:

$$K(-u) = K(u) \text{ для всіх значень } u. \quad (1.4)$$

Перша вимога гарантує, що результатом методу оцінки щільності ядра є функція щільності ймовірності. Друга вимога гарантує, що середнє значення відповідного розподілу дорівнює використаній вибірці.

Якщо  $K$  є ядром, то функція  $K^*$  також визначається як

$$K^*(u) = \lambda K(\lambda u), \text{ де } \lambda > 0. \quad (1.5)$$

Це можна використовувати для вибору масштабу, який підходить для даних.

Використовують кілька типів ядерних функцій: рівномірну, трикутну, Епанечнікова, Четвертинну (двоточкову), кубічну (треточкову), Гауса та косинусну. [1-8]

У таблиці нижче, якщо  $K$  задано з обмеженим носієм, тоді  $K(u) = 0$  для значень  $u$ , що лежать поза опорою (Табл. 1.1). [9] У цій же таблиці вказано ефективність ядра відносно ядра Епанечнікова з оптимізованою «шириною вікна»  $u$ .

Таблиця 1.1 Ядерні функції

Функції ядра, $K(u)$		$\int u^2 K(u) du$	$\int K(u)^2 du$	Ефективність відносно ядра Епанечнікова
Рівномірна	$K(u) = \frac{1}{2}$ при $ u  \leq 1$	$\frac{1}{3}$	$\frac{1}{2}$	92,9%
Трикутна	$K(u) = (1 -  u )$ при $ u  \leq 1$	$\frac{1}{6}$	$\frac{2}{3}$	98,6%

Епанечнікова	$K(u) = \frac{3}{4}(1 - u^2)$ при $ u  \leq 1$	$\frac{1}{5}$	$\frac{3}{5}$	100%
Четвертинна	$K(u) = \frac{15}{16}(1 - u^2)^2$ при $ u  \leq 1$	$\frac{1}{7}$	$\frac{5}{7}$	99,4%
Кубічна	$K(u) = \frac{35}{32}(1 - u^2)^3$ при $ u  \leq 1$	$\frac{1}{9}$	$\frac{350}{429}$	98,7%
Гауса	$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$	1	$\frac{1}{2\sqrt{\pi}}$	95,1%
Косинус	$K(u) = \frac{\pi}{4} \cos(\frac{\pi}{2}u)$ при $ u  \leq 1$	$1 - \frac{8}{\pi^2}$	$\frac{\pi^2}{16}$	99,9%

## 1.2 Деякі аспекти розрахунку потенційних кривих двохатомних молекул

Дослідження та побудова потенційних кривих для двохатомних молекул має принципове значення при пошуку активних середовищ в квантовій електроніці, спектроскопії, плазмохімії, хімічній фізиці, теорії конденсованих середовищ, теоретичній біології та ін. [10-14] Потенціальні криві двохатомних молекул мають основне значення для побудови ППЕ хімічно реагуючих систем. Простота і точність аналітичних виразів для двоатомних потенціалів визначає ефективність застосування ППЕ для констант швидкостей, динамічних розрахунків перерізів та енергообміну у хімічних реакціях. [11-15]

Велике значення парні потенціали отримують при побудові еталонних багаточасткових завдань квантової механіки, заснованих на математично коректних рівняннях квантової теорії розсіювання в системі багатьох тіл, для вирішення яких як вихідні дані необхідні заряди, маси і парні потенціали взаємодії між частинками, що визначають в основному точність одержуваних рішень. [16-17]

При виборі методів та засобів визначення залежностей потенційної енергії від відстані між ядрами молекули необхідно враховувати, що з одного боку ці методи мають бути достатньо математично строгим для отримання кількісних результатів, а з іншого боку, мати практичне застосування для розрахунків реальних молекул.

Роботи з дослідження методів розрахунку потенційних кривих реальних молекул проводять за напрямками

- Розробка принципово нових методів розрахунку; [11, 12, 19]
- Удосконалення існуючих методів, алгоритмів та програмного забезпечення; [11, 12, 17, 18]
- Розробка спеціалізованих обчислювальних систем. [20-23]

Для визначення залежності потенційної енергії від відстані між ядрами в загальному випадку необхідно вирішити рівняння Шредінгера для системи з  $M$  ядер і  $N$  електронів, яка має вигляд:

$$H\Psi = E\Psi \quad (1.6)$$

де  $H$  – оператор енергії;

$E$  – повна енергія системи;

$\Psi$  – хвильова функція.

Практично всі важливі результати в теоріях молекул отримуються на основі застосування різних наближених методів рішень рівняння (1.6):

1) Адіабатичне наближення. Ця апроксимація є основним підходом для розрахунків різних властивості молекул. [11, 12, 19, 24-26] Спрощена формула цього наближення полягає в тому, що стани молекули визначаються з конфігурації рівноваги ядер. Даний підхід базується на тому, що відношення ваги електрона до ваги ядра становить порядок  $10^{-3} - 10^{-5}$ , тому можна дослідити рух електронів, враховує, що рух ядер відбувається нескінченно повільно;

2) Методи, при яких різні властивості молекул визначаються без рішення рівняння Шредінгера та визначення хвильових функцій. [19-27] З цією метою застосовується метод теореми віріала та масштабні перетворення, які дозволяють використовувати загальні квантові закони при змінах ваги та зарядів частинки, що надходять у гамільтоніан;

3) Побудова емпіричних потенціалів за допомогою метода еталонного молекулярного потенціалу, тої чи іншої апроксимації таблично заданих парних потенціалів даних, побудови парних потенціалів, визначаєм характеристиками ізольованих атомів, складаючих молекулу. [10, 27-36] Особливість такого підходу визначається тим, що атом не втрачає своєї індивідуальності в молекулі при взаємодії з іншими атомами та ефективні розміри атомів у молекулі завжди менші, від відстані між ними, саме це дозволяє ввести параметр малості, необхідний для розрахунку потенціальної кривої (потенціальної функції) для всіх між'ядерних відстаней.

Загалом отримані потенціальні функції використовуються в розв'язку широкого ряду квантово-хімічних задач. Серед них важливе місце посідають розрахунки електронно-коливально-обертальних станів і переходів. Для малих молекул, такі розрахунки, при наявності достатньо точних потенціальних функцій, дозволяють відтворити переходи із експериментальною точністю. Так для двохатомних молекул стає можливим розв'язок радіального рівняння Шредингера.

$$\Psi''_{v,J}(r) - \frac{J(J+1)\Psi_{v,J}(r)}{r^2} + 2\mu(E_{v,J} - V(r))\Psi_{v,J}(r) = 0 \quad (1.7)$$

де  $\Psi_{v,J}$  – хвильова функція ядерних рухів в молекулі;

$v$  та  $J$  – квантові числа коливальних та обертальних рухів;

$\mu$  - приведена маса атомів молекули;

$E_{v,J}$  - енергія обертально-коливальних станів;

$V(r)$  – потенціальна функція.

Отже для точного розв'язку рівняння (1.7) необхідно мати функцію  $V(r)$  в точному аналітичному вигляді. Такі функції можуть бути отримані за допомогою безпосередніх квантово-хімічних розрахунків із послідовною їх апроксимацією. Останньому власне і присвячена дана робота.

### 1.3 Методи непараметричної регресії

Непараметричні методи являють собою статистичні прийоми, які не потребують специфікації функцій розподілу форм оцінюваних об'єктів. Замість цього данні самі певним способом формують модель.

Ядерне згладжування це один з найпростіших методів. Він простий у застосуванні, зрозумілий на інтуїтивному рівні та не вимагає додаткових математичних відомостей.

Ядерна регресія - непараметричний статистичний метод, що дозволяє оцінити умовне математичне прогнозування випадкової величини. Основа регресії в тому, щоб шукати нелінійні зв'язки між двома випадковими змінними  $X$  і  $Y$ .

Ядерна оцінка визначається як зважене середнє змінних відгуку у фіксованій окружності точки  $x$ , причому ваги визначалися ядром  $K$  і шириною вікна  $h$ .

Основа якісного проведення непараметричного оцінювання є вибір відповідної ширини вікна для наявної задачі. Головна роль ядерної функції  $K$  полягає в гладкості отриманої оцінки і забезпеченні диференційованості. Ширина вікна  $h$  визначає поведінку оцінки в кінцевих вибірках, що ядерна функція зробити просто не в змозі.  
[37]

Конструкція оцінок найближчих сусідів ( $K$ -NN оцінка) не відповідає ядерним оцінкам. Оцінка  $k$ -найближчих сусідів є середнім, зваженим в змінній окружності. Дана окружність визначається тими значеннями змінної  $X$ , які є  $k$  найближчими до  $x$  по евклідовій відстані.

Параметр згладжування  $k$  дає ступінь гладкості оцінки кривої. Він зазнає ту ж роль, що і ширина вікна для ядерних згладжувань. Вплив змінного  $k$  на якісні характеристики оцінки аналогічні випадку ядерних оцінок з прямокутним ядром.

Метод оцінювання ортогональних розкладань допускає, що функція регресії може бути представлена в вигляді ряду Фур'є:

$$y(x) = \sum_{j=0}^{\infty} \beta_j \varphi_j(x) \quad (1.8)$$

де  $\{\varphi_j\}_{j=0}^{\infty}$  – відома система базисних функцій;

$\{\beta_j\}_{j=0}^{\infty}$  – невідомі коефіцієнти Фур'є.

Якщо зафіксувати базис функцій, проблема оцінювання функції регресії може бути зведена до оцінювання коефіцієнтів Фур'є. Однією із складностей є те, що може бути нескінченно багато ненульових коефіцієнтів  $\beta_j$ . Таким чином, при заданому кінцевому обсязі вибірки можна ефективно оцінити лише підмножину коефіцієнтів.

Загальною мірою близькості до даних для деякої кривої  $g$  є сума квадратів нев'язок:

$$\sum_{i=1}^n (Y_i - g(X_i))^2 \quad (1.9)$$

Якщо  $g$  може буди кривою - необмеженою в функціональному сенсі - то ця міра, що має сенс відстані, дорівнює нулю для всякої кривої  $g$ , що інтерполює дані.

Підхід, заснований на згладжуванні сплайнами, виключає цю небажану інтерполяцію даних за рахунок досягнення компромісу між двома суперечливими цілями: отримати гарну апроксимацію даних і отримати криву, яка не має надто швидких локальних змін.

Відомі різні способи кількісної оцінки локальних змін. Можна визначити міру плавності кривої, засновану, наприклад, на першій, другій, і більш старших похідних. Для успішного розкриття основної ідеї найзручніше ввести інтеграл від квадрата другої похідної, тобто для кількісної оцінки локального зміни використовувати штраф за порушення плавності:

$$\int (g''(x))^2 dx \quad (1.10)$$

Рекурентні методи розглядають данні не як фіксований об'єм, а як послідовність пар  $(X_1, Y_1), (X_2, Y_2), \dots$  що надходить з виходу деякого пристрою спостереження. У загальному випадку можна розглядати дані як часовий ряд. Оскільки непараметричні оцінки зазвичай визначаються по всій вибірці, її доводиться перераховувати при надходженні нових даних.

Отже, з обчислювальної точки зору краще, щоб оцінка регресії, заснована на  $(n+1)$  точках, будувалася виходячи з  $(n+1)$ -го спостереження  $(X_{n+1}, Y_{n+1})$  та оцінки, отриманої за першими  $n$  точками.

#### 1.4 Методи апроксимації нелінійних функцій

Якщо залежність спостережуваних значень  $y$  і від параметрів нелінійна, тобто:

$$y_i = h(x_i; \theta) + \varepsilon_i, \quad i = 1, \dots, n \quad (1.11)$$

де  $x_i$  – значення (скалярного або векторного) аргументу в  $i$ -му спостереженні;

$h$  – функція заданого виду, що залежить від параметрів  $\theta = (\theta_1, \dots, \theta_k)^T$ ;

$\varepsilon_i$  – похибка вимірювань.

Властивості оптимальної оцінки найменших квадратів, які доведені в лінійному випадку, ґрунтуються на лінійній моделі по  $\theta$  і на лінійності оцінок:

$$\theta = \text{Arg min } \|y - h(\theta)\|^2 \quad (1.12)$$

де  $h(\theta) = (h_1(\theta), h_2(\theta), \dots, h_n(\theta))^T \in R^n$ .

Вони не мають значення для  $h(x; \theta)$  нелінійного виду. Оцінка (1.12) зазвичай не може бути записана в явній формі, а отримується чисельно в ітераційному процесі мінімізації:

$$Q(\theta) = \|y - h(\theta)\|^2 \quad (1.13)$$

Тому в загальному випадку метод найменших квадратів (МНК) може бути не обґрунтований. А метод максимуму правдоподібності (МП) являє собою обґрунтований метод:

$$\theta_{ML} = \text{Arg max } L(\theta) \quad (1.14)$$

Однак, оцінка (1.14) не буде мати нормального розподілу. Тому теорія нормальної регресії для нелінійних функцій не підходить. Так, незміщенність і ефективність МНК-оцінки  $\theta$  лише асимптотична.

Одним з методів апроксимації нелінійної функції це градієнтні методи. Дані методи використовують рішення з допомогою градієнта задач, що зводяться до знаходження локальних екстремумів функції. [38]

Основна ідея методів полягає в тому, щоб йти в напрямку найшвидшого спуску, а цей напрямок задається антиградієнтом.

Рішення функції знаходиться ітераційно градієнтним методом:

$$\theta^{(l+1)} = \theta^{(l)} + \Delta\theta^{(l)} \quad (1.15)$$

де  $l$  – номер шагу;

$\theta^{(l)}$  – наближення для  $\theta$  на  $l$ -й ітерації;

$\theta^{(l+1)}$  – наближення для  $\theta$  на  $l+1$ -й ітерації;

$\Delta\theta^{(l)} \in R^k$  – ітераційний шаг.

Ітераційний шаг визначається формулою:

$$\Delta\theta^{(l)} = \nu_l \rho_l \quad (1.16)$$

де  $\nu_l$  – напрямком одиничного вектору, лежачий на проміні, проведений з  $\theta^{(l)}$ ;

$\rho_l$  – величина шагу.

Метод називається допустимим, якщо він складається лише з допустимих шагів. Для цього має виконуватися умова:

$$\nu = -Rg \quad (1.17)$$

де  $R \in$  матрицею  $R^{k \times k}$ ,  $R > 0$ .

Необхідно, щоб черговий шаг наближав до рішення  $\theta^*$  (локальний мінімум). Так як  $\theta^*$  невідомий, тому шаг вважається успішним, якщо він зменшив цільову функцію (1.14).

В методі найшвидшого спуску величина шагу є довільною:

$$\rho = \text{const} > 0 \Rightarrow \theta^{(l+1)} = \theta^{(l)} - \rho_l g_l \quad (1.18)$$



Метод є допустимим, та значення цільової функції вдалині від мінімуму зменшуються доволі швидко. Однак поблизу від точки  $\theta^*$  для функцій з сильно витягнутими лініями рівня ітерації даного методу приводять до шагів від борту до борту долини, тому кількість ітерацій для досягнення  $\theta^* = \text{Arg min } Q(\theta)$  може бути велика, а збіжність дуже повільна. Даний метод зазвичай не часто використовується в практиці.

Метод Ньютона розглядає квадратичне наближення для  $Q(\theta)$ :

$$Q(\theta) \approx \Phi(\theta) = Q(\theta^{(l)}) + (\theta - \theta^{(l)})^T g_l + 1/2(\theta - \theta^{(l)})^T H_l(\theta - \theta^{(l)}) \quad (1.19)$$

де  $H_l \in R^{k \times k}$  – матриця других похідних  $Q(\theta)$ .

В якості  $\theta^{(l+1)}$  використовується стаціонарна точка  $\Phi(\theta)$ :

$$0 = \partial \Phi / \partial \theta = g_l + H_l(\theta - \theta^{(l)}) \Rightarrow \theta^{(l+1)} = \theta^{(l)} - g_l / H_l \quad (1.20)$$

Шаг є допустимим при  $H_l > 0$ . Проте достатньою умовою є  $H_l > 0$  в точці мінімуму  $\theta^*$ . Оскільки функція нерозривна, то в окружності мінімуму також виконується умова. Отже, метод Ньютона допустимий в окружності  $\theta^*$  з  $R = H_l^{-1}, \rho_l = 1$ .

Серед його достоїнств те, що для квадратичної функції  $\Phi(\theta)$  він сходиться за одну ітерацію та для не квадратичного функцій він має максимальну швидкість збіжності (квадратичну) серед загальних градієнтних методів.

Проте, при розгляданні функції не в окружності точки мінімуму, умова  $H_l > 0$  не завжди виконується, тому потрібні точні початкові оцінки. Розрахунки в ітерації матриці  $H_l$  більш складні, чим її обернення, так як потрібно знаходити другу похідну.

На відміну від методу Ньютона, метод Гаусса-Ньютона може бути використаний тільки для мінімізації суми квадратів, але його перевага в тому, що метод не вимагає обчислення других похідних, що може виявитися суттєвою трудностю. [39]

Рівняння (1.19) для даного методу матиме вигляд:

$$\theta^{(l+1)} = \theta^{(l)} - g_l G_l^{-1} = \theta^{(l)} + (X^T X)^{-1} X^T e^{(l)} \quad (1.21)$$

де  $X \in R^{n \times k}$ ;

$G \square H$ .

В матричному вигляді:

$$h(\theta) \approx h(\theta^{(l)}) + X(\theta - \theta^{(l)}) \Rightarrow e = y - h(\theta) \cong y - h(\theta^{(l)}) + X(\theta - \theta^{(l)}) = e^{(l)} - X\Delta\theta \quad (1.22)$$

$G \geq 0$  при всіх  $\theta$ , і для її розрахунку потрібні лише первинні похідні, проте дана умова не гарантована вдалі від точки мінімуму  $\theta^*$ . Також збіжність більш повільна ніж в методі Ньютона. В цьому методі  $R_l = G_l^{-1}, \rho_l = 1$ .

Метод Левенберга-Марквардта є альтернативою методу Ньютона. Він може розглядатися як комбінація останнього з методом градієнтного спуску або як метод довірчих областей. [40]

Шаг методу Левенберга-Марквардта з мінімізації цільової функції:

$$Q_\lambda(\theta) = \|e^{(l)} - X\Delta\theta\|^2 + \lambda(\Delta\theta)^T P \Delta\theta \quad (1.23)$$

де  $P = \text{diag}\{X^T X\}$ .

Вводячи матрицю  $V = P^{1/2} = \text{diag}\{P_{jj}^{1/2}, \dots, P_{kk}^{1/2}\} = V^T$ , рівняння (1.23) приймає вид:

$$Q_\lambda(\theta) = \|e^{(l)} - X\Delta\theta\|^2 + \lambda(\Delta\theta)^T V^T V \Delta\theta = \|e^{(l)} - X'z\|^2 + \lambda \|z\|^2 \quad (1.24)$$

де  $z = V\Delta\theta$ ;

$X' = XV^{-1}$ .

Значення  $\lambda$  змінюється з ітераціями: на перших, вдалі від  $\theta^*$ ,  $\lambda$  велике, так що член  $\lambda P$  переважний, і ітерація близька до ітерації методу найшвидшого спуску, що забезпечує допустимість шагу і швидке зменшення значень  $Q(\theta)$ . З наближенням до  $\theta^*$  значення  $\lambda$  зменшуються, і метод Левенберга-Марквардта близький до методу Гаусса-Ньютона, що забезпечує його ефективність на останніх ітераціях.

## 2 РОЗРАХУНКОВА ЧАСТИНА

### 2.1 Апроксимація нелінійних функцій методом ядерної регресії

Отже визначаємо регресійну модель відгуку системи  $y_i$  як нелінійної функції змінних  $x_i$ :

$$y_i = g(x_i) + \varepsilon_i \quad (2.1)$$

де  $\varepsilon_i$  параметр, що визначає похибку системи. Локально-лінійна регресія оцінює регресію для підмножини спостережень для кожної точки даних і вирішує задачу мінімізації, яка задана в формі:

$$\min_{\gamma} \sum_{i=1}^n (y_i - \gamma_0 - \gamma_1(x_i - x))^2 K(x_i, x, h) \quad (2.2)$$

де  $\gamma = (\gamma_0, \gamma_1)$  і  $K(x_i, x, h)$  є параметри і ядро для кожної точки

$$K(x_i, x, h) = \prod_{j=1}^k K_j(x_{ij}, x_j, h_j) \quad (2.3)$$

$$K_j(x_{ij}, x_j, h_j) = k_j \left( \frac{x_{ij} - x_j}{h_j} \right) \quad (2.4)$$

Оцінка для усіх точок  $x$  даних веде до звичайного виразу схожого із зваженим МНК

$$\hat{\gamma} = (Z^+ W Z)^{-1} Z^+ W y, \quad (2.5)$$

де  $\hat{\gamma} = (\hat{\gamma}_0, \hat{\gamma}_1)^+$ ;

$Z$   $n(k+1)$  матриця в якій  $i$ -та строчка дається наступним чином  $\{1, (\hat{\gamma}_0, \hat{\gamma}_1)^+\}^+$ ;

$W$  –  $n \times n$  діагональна матриця в якій на  $i$  діагоналі знаходиться  $K(x_i, x, h)$ .

Для нелінійної апроксимації ми використовували Nadaraya-Watson наближення. Тоді найпростіший ядерний підхід для апроксимації нелінійних залежностей витікає з узагальненого середнього. Отже це веде до мінімізації функції в дусі МНК:

$$\min_{y(x)} \sum_{j=1}^n (y_j - y(x))^2 K_h(|x_j - x|) \quad (2.6)$$

$$y(x) = \frac{\sum_{j=1}^n K_h(|x_j - x|) y_j}{\sum_{j=1}^n K_h(|x_j - x|)} \quad (2.7)$$

Де, в якості ядра нами вибрано найбільш популярне експоненціальне:

$$K_h(|x_j - x|) = \exp(-h(x - x_j)^2). \quad (2.8)$$

Тут параметр  $h$  («ширина вікна», «вікно», bandwidth – пропускна здатність) вочевидь пов'язано із дисперсією функції розподілу  $h \approx 1/2\sigma^2$ . Саме цей параметр визначає «дальнодію» впливу точок на задану точку. При використанні ядерної регресії певної уваги потребує дослідження кривизни отриманої функції. Адже при певних значеннях параметру  $h$  апроксимація може отримати інтерполяційний характер. Тобто з'являється тенденція пройти лініям якнайближче до усіх точок.

Для контролю цієї обставини нами запропоновано розраховувати кривизну пласкої кривої. Відповідна величина (кривизна) добре відома в геометрії:

$$curve = \frac{y''}{(1 + (y')^2)^{\frac{3}{2}}}. \quad (2.9)$$

Для цього ми розраховували першу та другу похідні:

$$y' = \frac{\sum K_h' y \sum K_h - \sum K_h' \sum K_h y}{(\sum K_h)^2}, \quad (2.10)$$

$$y'' = \frac{(\sum K_h'' y \sum K_h + \sum K_h' y \sum K_h' - \sum K_h'' \sum K_h y - \sum K_h' \sum K_h' y)(\sum K_h)^2 - (\sum K_h' y \sum K_h - \sum K_h' \sum K_h y)^2 \sum K_h'}{(\sum K_h)^4} \quad (2.11)$$

Відповідні похідні ядра можна легко розрахувати:

$$K_h'(|x_j - x|) = -2h(x - x_j) \exp(-h(x - x_j)^2), \quad (2.12)$$

$$K_h''(|x_j - x|) = (-2h + 4h^2(x - x_j)^2) \exp(-h(x - x_j)^2). \quad (2.13)$$

Відповідні значення кривизни ми розраховували для кожної точки  $curve_i = curve(x_i)$ . Отримані величини для повної кривої були використані для обчислення середньої величини кривизни:

$$|curve| = \sum_j curve_j / M, \quad (2.14)$$

де  $M$  – кількість точок кривої.

Інформаційний індекс, що характеризує неоднорідність кривизни протягом всієї функції  $c_i$  оцінюється на основі функції Шеннона:

$$Inf(c) = - \sum_i \frac{|curve_i|}{\sum |curve_j|} \log_2 \frac{|curve_i|}{\sum |curve_j|}. \quad (2.15)$$

Також ми розраховували так званий індекс Джині (*Gini* коефіцієнт) який добре відомий в економетрії. Цей індекс характеризує неоднорідність розподілу кривизни або ступінь розшарування даних з кривизни:

$$Gini = \sum_j curve_j (1 - curve_j). \quad (2.16)$$

При цьому відповідні величини кривизни для усіх точок було нормовано:

$$\sum_j curve_j = 1. \quad (2.17)$$

Для оцінки якості апроксимації ми використовували відомі параметри – коефіцієнт детермінації.

$$R^2 = 1 - \frac{\sum (y_j - y_j^{calc})^2}{\sum (y_j - \bar{y})^2}. \quad (2.18)$$

де величина  $y_j^{calc}$  – результат розрахунку залежної змінної для навчаючої вибірки;

$\bar{y}$  - середнє значення величини  $y$ .

Звісно, що цей параметр ( $R^2$ ) не може бути використаний для непараметричної (ядерної) регресії тому, що строго кажучи регресійні моделі такого роду не роблять оцінок величин навчаючої вибірки. Тому принаймні необхідно провести оцінку за процедурою Leave-One-Out (LOO).

$$Q^2 = 1 - \frac{\sum (y_j - y_j^{pred})^2}{\sum (y_j - \bar{y})^2}. \quad (2.19)$$

Яка базується на послідовному вилученні точок із навчаючої вибірки і відповідної оцінки у для них. В формулі (2.19)  $y_j^{pred}$  – вектор що утримує передбачені величини залежної змінної.

В кінці цього підрозділу відзначимо кілька важливих моментів що стосуються ядерної (іноді кажуть Кернел) регресії:

1. Кернел-регресія – це оцінка середньозваженого значення  $y_i$ .
2. Ядро Гауса навколо певної «точки запиту»  $x_i$  дає оцінку  $y_i$  та внески за рахунок інших точок.
3. Оскільки вагові коефіцієнти плавно змінюються залежно від  $x$ , сама регресійна оцінка також плавна.
4. З приводу ядра. Звісно, що різні ядра можуть дати різні результати. Але реальні оцінки кажуть що, наприклад Гаусове ядро, дає результати близькі до Єпанечнікова.
5. Набагато більша різниця в описі походить від вибору різних значень параметру  $h$ .
6. Адекватний вибір  $h$  є основною проблемою розрахунку за ядерною регресією.
7. Позитивним моментом ядерної регресії є той факт, що лише один параметр  $h$  варіюється.
8. Мінімальний набір параметрів що характеризує адекватність ядерної регресії повинен включати величину  $Q^2$  (2.19) та кривизну (2.14)

## 2.2 Тестові розрахунки методом ядерної регресії

В цьому підрозділі ми описуємо кілька тестових прикладів використанні ядерної регресії.

Приклад №1. На рис. 2.1 наведено такий приклад (ми умовно позначаємо цю функцію як «квазі-сінус»).

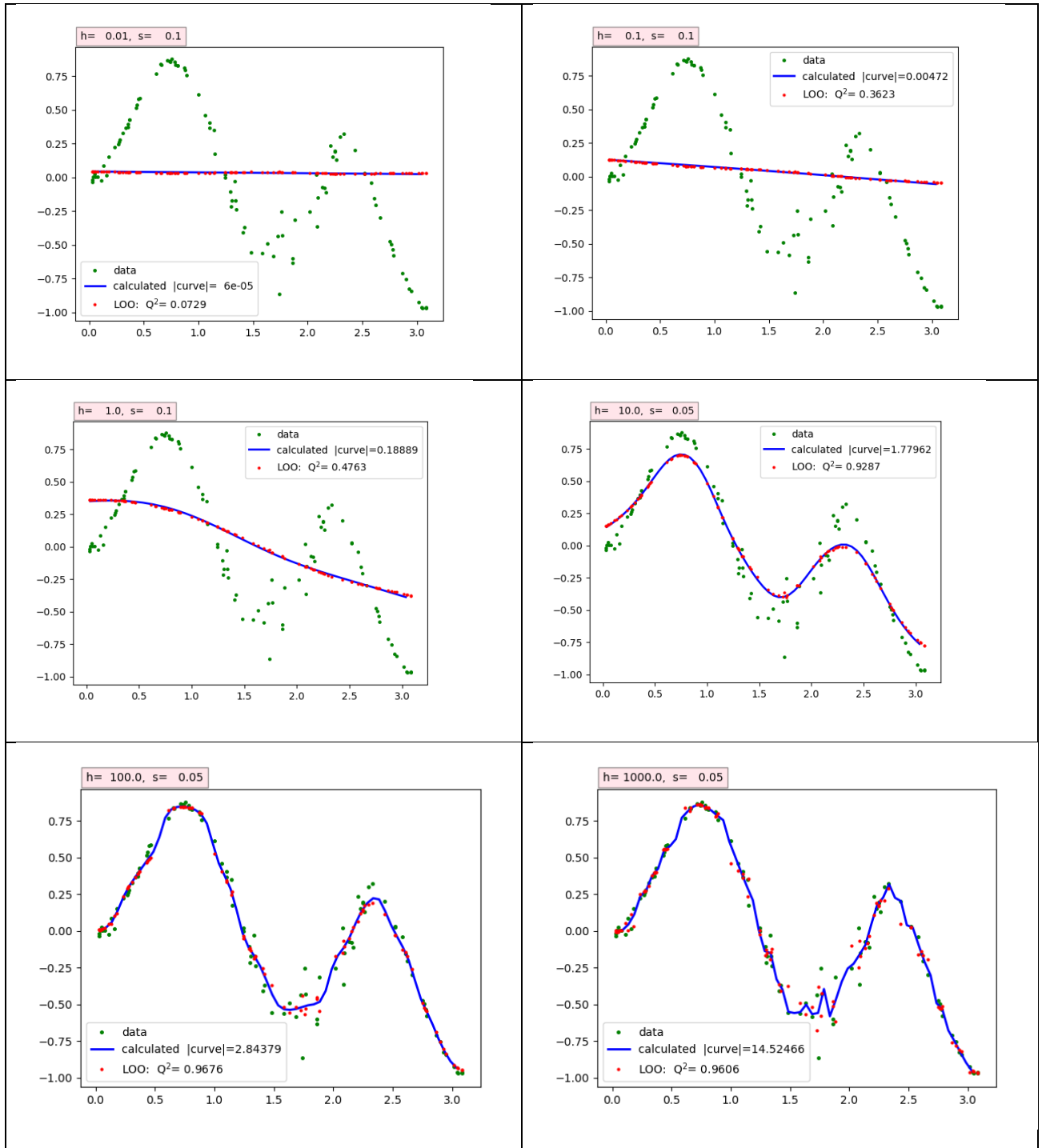


Рисунок 2.1. Функція «квазі-сінус» та її апроксимація методом ядерної регресії

$$\begin{aligned}
 x &= \text{random.uniform}(0, \pi) \\
 \sigma &= 0.01 + 0.05(1 - \sin(2.5x))^2 \\
 y &= \text{random.gauss}(\sin(2.5x) \sin(1.5x), \sigma)
 \end{aligned}
 \tag{2.20}$$

Функція Python *random.uniform*(0,  $\pi$ ) реалізує рівномірне розповсюджені на заданому інтервалі  $[0, \pi)$  випадкові числа. Функція *random.gauss*(mean,  $\sigma$ ) генерує випадкові числа за Гаусом навколо середнього значення mean із стандартним відхиленням  $\sigma$ . Загалом тут було реалізовано доволі великий розкид точок.

При використанні значення  $h = 10$ , функція є гладкою із відносно невеликою величиною середньої кривизни  $|curve| = 1.78$  та непоганою величиною  $Q^2 = 0.9287$ . При зростанні  $h \geq 100$  функція ставала надто хвилястою  $|curve| \geq 2.84$ .

Приклад №2. Представлені на рисунку 2.2.2 залежності характеризують графік А як доволі гладкий. Величина середнього значення склала  $|curve| = 0.00493$ , тоді як у випадку В, коли  $h = 2.0$ ,  $|curve| = 0.36377$ , що є свідомством значного () зростання кривизни.

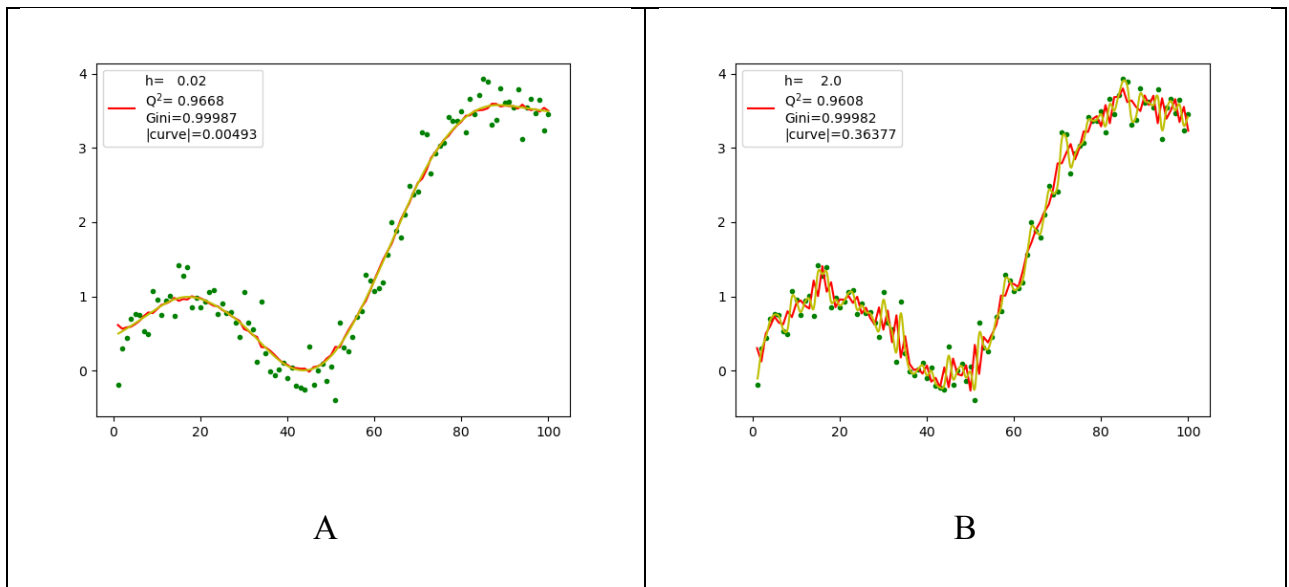


Рисунок 2.2. Приклад використання ядерної регресії з різними значеннями параметра  $h$



Приклад №3. В цьому прикладі розглядається так звана білінійна функція, яка має наступне вигляд.

$$y = -\log(\exp(a(x - x_0)) + \exp(b(x - x_0))) \quad (2.21)$$

Відповідні параметри обрано таким чином:  $a = 8$ ,  $b = -1.5$ ,  $x_0 = 0$ . Всього було генеровано  $N = 80$  точок на інтервалі  $x = [-1:1]$ . Певний статистичний розкид було генеровано за Гаусовим розподілом  $\sigma = 1$  (рис. 2.3)



Рисунок 2.3. Білінійна функція із розкидом. Апроксимація за методом найменших квадратів

На цьому ж рисунку представлено результат апроксимації методом найменших квадратів. Для її реалізації було використано процедуру Левенберга-Марквардта (див. літогляд). Програму наведено в додатку В. Можна бачити, що вказані параметри якісно близькі до початкових (до того як була внесена похибка за Гаусовим розподілом). Коефіцієнт детермінації отриманий за нелінійним методом найменших квадратів в процедурі LOO  $Q^2 = 0.8240$  непогано узгоджується із  $R^2$ . ( $R^2 - Q^2 = 0.013$ ).

Результати апроксимації методом ядерної регресії представлено на рис. 2.4. Можна бачити, що за параметра  $h = 20$  отримано адекватний опис даних  $Q^2 = 0.820$ , який може бути зіставлено із попереднім результатом отриманим МНК. Середня величина кривизни  $|curve| = 1.095$  є індикатором якості функції. Подальше збільшення  $h = 100$  веде до значного збільшення  $|curve| = 5.55$  що характеризує апроксимацію як таку, що близька до інтерполяції. Ще одна важлива обставина стає очевидною при аналізі трьох вказаних прикладів. А саме – з якоїсь величини  $h$  коефіцієнт детермінації  $Q^2$  починає зменшуватись (Табл. 2.1). Також дані таблиці характеризують індекси  $Inf(c)$  та  $Gini$  як малоінформативні.

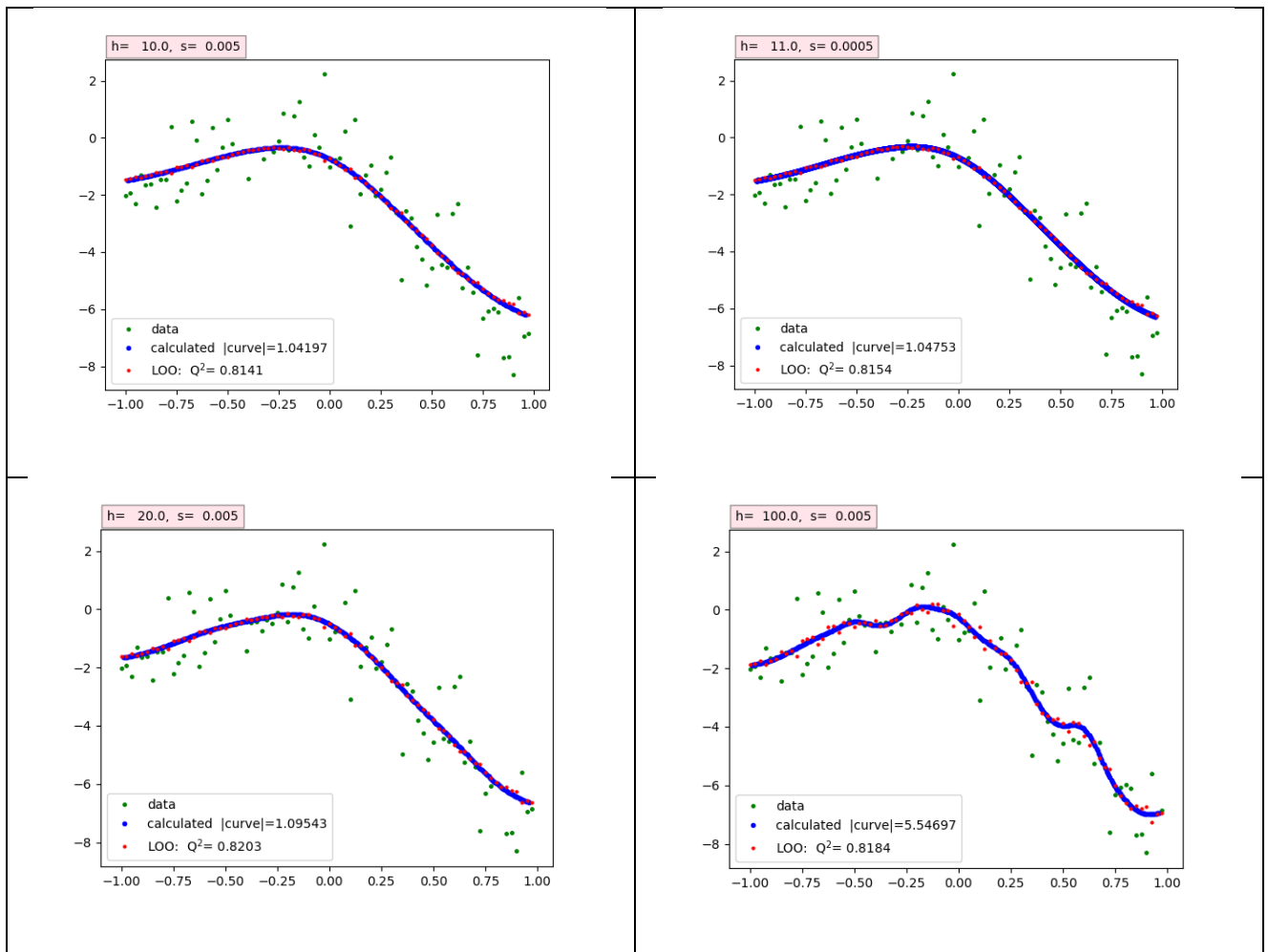


Рисунок 2.4. Білінійна функція із розкидом. Апроксимація за методом ядерної регресії

Таблиця 2.1. Параметри ядерної регресії для білінійної функції (2.21)

<i>h</i>	15	20	25	35	45	50	100
$Q^2$	0.8185	0.8203	0.8213	0.8222	0.8223	0.8222	0.8184
$ curve $	1.07	1.10	1.16	1.41	1.82	2.11	5.55
$Inf(c)$	6.49	6.31	6.27	6.46	6.63	6.66	6.36
<i>Gini</i>	0.9816	0.9779	0.9758	0.9774	0.9821	0.9840	0.9828

### 2.3 Апроксимація потенціальної енергії функцією Морзе

Одна з найпростіших (і найстаріших) функцій які можна використовувати для опису потенціальних кривих дисоціації двохатомних молекул є функція Морзе (2.23)

$$V(r)=D\left(1-e^{-\beta(r-r_e)}\right)^2, \quad (2.23)$$

де  $D$ ,  $\beta$  – параметри функції (енергія дисоціації, та параметр що визначає крутизну нахилу кривої відповідно);

$r_e$  – Оптимальна (мінімум) між'ядерна відстань;

$r$  – між'ядерна відстань.

Загалом така функція  $V(r)$  має вигляд :

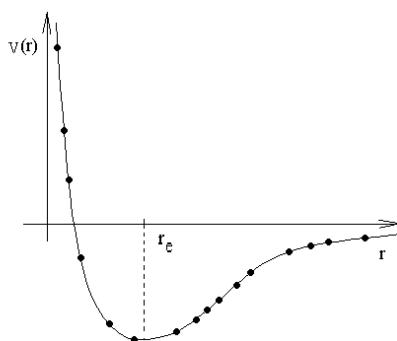


Рисунок 2.5. Енергія двохатомної молекули як функція між'ядерної відстані

Оптимальні параметри можуть бути знайдені МНК якщо відомі розрахункові дані залежності  $V_k=V(r_k)$  для певної кількості точок  $r_k$ . Отже параметри  $D$ ,  $\beta$  та  $r_e$  мають бути отримані мінімізацією виразу:

$$F(D, \beta, r_e) = \sum_{k=1}^N (V_k^{calc} - V_k)^2 \quad (2.24)$$

де  $V_k$  – задане з квантово-хімічного розрахунку значення функції за відстані  $r_k$ ;  $V_k^{calc}$  – розрахована згідно (2.3.1) величина енергії системи.

Відповідні (необхідні) похідні можуть бути легко знайдені:

$$\frac{\partial \Phi(D, \beta, r_e)}{\partial D} = -2 \sum_k (V_k^{calc} - V_k) \frac{\partial V_k}{\partial D}, \quad (2.25)$$

$$\frac{\partial \Phi(D, \beta, r_e)}{\partial \beta} = -2 \sum_k (V_k^{calc} - V_k) \frac{\partial V_k}{\partial \beta}, \quad (2.26)$$

$$\frac{\partial \Phi(D, \beta, r_e)}{\partial r_e} = -2 \sum_k (V_k^{calc} - V_k) \frac{\partial V_k}{\partial r_e}. \quad (2.27)$$

$$\text{де } \frac{\partial V_k}{\partial \mathbf{x}} = \left( \frac{\partial V(\mathbf{r})}{\partial \mathbf{x}} \right)_{\mathbf{r}=\mathbf{r}_k}$$

Простота функції Морзе (2.23) дає можливість легко розрахувати відповідні компоненти для (2.25-2.27):

$$\frac{\partial V(\mathbf{r})}{\partial D} = 1 - 2e^{-\beta(r-r_e)} + e^{-2\beta(r-r_e)}, \quad (2.28)$$

$$\frac{\partial V(\mathbf{r})}{\partial \beta} = 2D(r-r_e) \left( e^{-\beta(r-r_e)} - e^{-2\beta(r-r_e)} \right), \quad (2.29)$$

$$\frac{\partial V(\mathbf{r})}{\partial r_e} = 2D\beta \left( -e^{-\beta(r-r_e)} + e^{-2\beta(r-r_e)} \right). \quad (2.30)$$

Отже, ми маємо градієнт, що розраховується на кожному кроці оптимізаційної задачі

$$\mathbf{g} = \left( \frac{\partial V(\mathbf{r})}{\partial D}, \frac{\partial V(\mathbf{r})}{\partial \beta}, \frac{\partial V(\mathbf{r})}{\partial r_e} \right) \quad (2.31)$$

А сама ітераційна задача, в рамках методу Ньютонa, для шуканого вектору параметрів  $\mathbf{x} = (D, \beta, r_e)$ , має вигляд:

$$\mathbf{x}^{(\ell)} = \mathbf{x}^{(\ell-1)} - \alpha \mathbf{H}^{-1} \mathbf{g}^{(\ell-1)} \quad (2.32)$$

де  $\ell$  – ітераційний крок;

$\alpha$  – параметр збіжності ітераційної процедури  $\alpha \approx 0.5$ ;

$\mathbf{H}$  – матриця других похідних (матриця Гесса) з елементами:

$$H_{ij} = \frac{\partial^2 \Phi}{\partial x_i \partial x_j} = 2 \sum_k \frac{\partial V_k}{\partial x_i} \frac{\partial V_k}{\partial x_j} - 2 \sum_k (V_k^{calc} - V_k) \frac{\partial^2 V}{\partial x_i \partial x_j}, \quad (2.33)$$

де  $x_i, x_j = D, \beta, r_e$ .

$$\frac{\partial^2 V_k}{\partial D^2} = 0 \quad (2.34)$$

$$\frac{\partial^2 V_k}{\partial D \partial \beta} = 2(r - r_e) (e^{-\beta(r-r_e)} - e^{-2\beta(r-r_e)}) \quad (2.35)$$

$$\frac{\partial^2 V_k}{\partial D \partial r_e} = -2\beta (e^{-\beta(r-r_e)} - e^{-2\beta(r-r_e)}) \quad (2.36)$$

$$\frac{\partial^2 V_k}{\partial \beta^2} = -2D(r - r_e)^2 (-e^{-\beta(r-r_e)} + 2e^{-2\beta(r-r_e)}) \quad (2.36)$$

$$\frac{\partial^2 V_k}{\partial \beta \partial r_e} = 2D\beta(r - r_e) (e^{-\beta(r-r_e)} - 2e^{-2\beta(r-r_e)}) \quad (2.37)$$

$$\frac{\partial^2 V_k}{\partial r_e^2} = -2D\beta^2 (e^{-\beta(r-r_e)} - 2e^{-2\beta(r-r_e)}) \quad (2.38)$$

Ітераційна процедура зупиняється коли  $\|g\| < \epsilon_{ps}$ .

В якості узагальненої функції Морзе ми розглядаємо наступну:

$$V^{GM}(r) = \sum_{m \geq 2} \alpha_m (1 - e^{-\beta_m(r-r_e)})^m \quad (2.39)$$

Для описаного в цьому підрозділі методу розрахунку було створено відповідну програму на скриптовій мові Python3.

## 2.4 Регресійні моделі опису потенціальних кривих двохатомних молекул

### 2.4.1 Молекула ВН

Для молекули ВН було розраховано енергію системи для різних між'ядерних відстаней. Розрахунок був проведений в багато-конфігураційному методі самоузгодженого поля (Complete Active Space Self Consistent Field, CASSCF). Базиси розрахунку cc-pVDZ та cc-pVQZ. Апроксимацію табличних даних було проведено за функцією Морзе та методом ядерної регресії при різних значеннях параметру  $h$ . Згідно результатами попередніх підрозділів для аналізу точності регресії, ми розрахували для кожного значення параметру  $h$  відповідне значення  $Q^2$  за процедурою LOO.

В табл. 2.2 представлено дані що стосуються потенціальної енергії дисоціації молекули ВН. Можна бачити, що після  $h=300$  кривизна надто значно зростає.

Таблиця 2.2. Параметри, що характеризують апроксимацію потенційної кривої для молекули ВН

$h$	$Q^2$	$Gini$	$ curve $
2	0.7781	0.9960	0.0395
50	0.9735	0.9861	0.1435
150	0.9916	0.9849	0.1714
250	0.9952	0.9859	0.1993
<b>300</b>	<b>0.9961</b>	<b>0.9859</b>	<b>0.2228</b>
700	0.9984	0.8544	2.4601
1000	0.9989	0.8261	19.5186

При використанні МНК для функції Морзе була проведена ітераційна процедура, за допомогою якої були знайдені параметри  $D$ ,  $\beta$  та  $r_e$ , що характеризують криву:

$$D = 0.2033, \beta = 1.3320, r_e = -0.0124 \quad (2.40)$$

Для аналізу точності розрахунків, ми розраховували для двох методів значення  $Q^2$  – квадрату коефіцієнта кореляції за процедурою LOO,  $Q^2_2$  – квадрат коефіцієнта кореляції Пірсона та  $\sigma^2$  – дисперсію.

$$Q^2_2 = \frac{\left( \sum_j (y_j - \bar{y})(y_j(x) - \bar{y}(x)) \right)^2}{\sum_j (y_j - \bar{y})^2 \sum_j (y_j(x) - \bar{y}(x))^2}, \quad (2.41)$$

де  $y_j$  – теоретичне значення отримане з квантово-хімічного розрахунку;

$y(x)$  – результат отриманий з регресійного аналізу (МНК або ядерна регресія);

$\bar{y}$  – середнє теоретичне значення для величин  $y_j$ ;

$\bar{y}(x)$  – середнє значення величин  $y(x)$ .

Також напишемо вираз для дисперсії

$$\sigma^2 = \frac{\sum_j (y_j - y_j(x))^2}{N - p}, \quad (2.42)$$

де

$p$  – кількість параметрів регресійної моделі

Картинки, що стосуються апроксимації кривої для молекули ВН представлено на рис. 2.6.

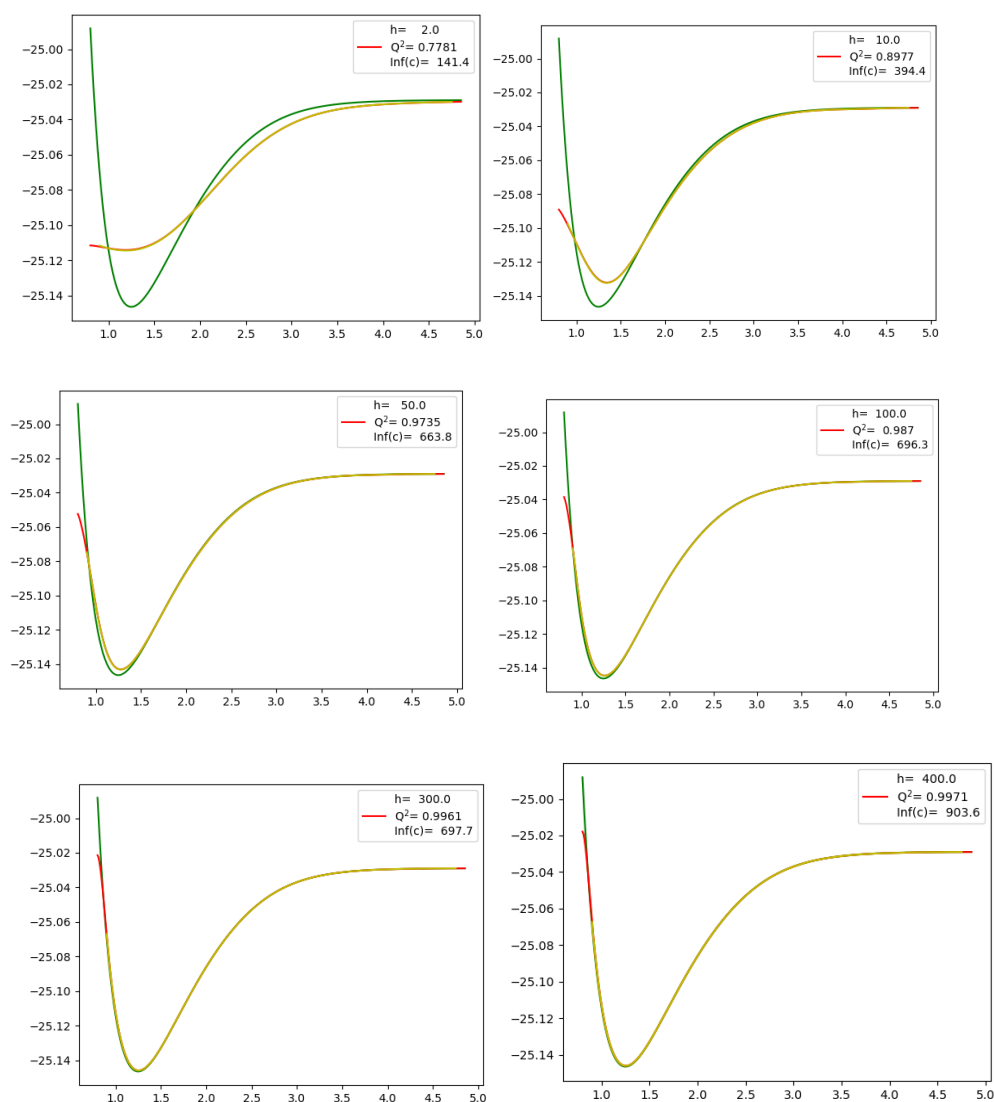


Рисунок 2.6. Залежність характеру апроксимації потенціальної поверхні ВН від параметру  $h$

Можна бачити, що апроксимації, що відповідають  $h=100-300$  є задовільними і можуть бути використані при розв'язанні радіального рівняння Шредингера (1.7).

Його розв'язок представлено в Табл. 2.3. Тут наведено дані отримані для трьох типів апроксимації.  $V(\text{Morse})$  (2.23) відповідає МНК апроксимації функції Морзе (три параметри –  $D$ ,  $\beta$ ,  $r_e$ ).  $V_6(\text{GM})$  – узагальнена функція Морзе (2.39) – сім параметрів ( $\alpha_i$ ,  $\beta_i$ ,  $i=2-4$ ,  $r_e$ ). І, нарешті, ядерна регресія – один параметр. В дужках позначено відхилення від експериментальних величин.

Можна бачити, що результати наших нових розрахунків значно краще ніж звичайний розрахунок за потенціалом Морзе і близький, до результатів із узагальненим потенціалом Морзе.

Таблиця 2.3. Розрахунок коливальних станів ВН. В дужках представлено відхилення розрахованих величин від експериментальних. Усе в  $(\text{cm}^{-1})$

v	CASSCF(2,2)/cc-pVDZ			$E_v$ Експеримент
	V(Morse)	$V_6(\text{GM})$	Ядерна Регресія, $h=300$	
0	1192.2 (21.2)	1185.3 (14.2)	1184(13)	1171.1
1	3504.3 (63.9)	3470.4 (30.0)	3476 (35)	3440.4
2	5726.8 (112.6)	5634.6 (20.4)	5654 (39)	5614.2
3	7868.1 (173.4)	7711.9 (17.2)	7719 (24)	7694.7
4	9939.4 (255.4)	9772.8 (88.8)	9723 (39)	9684.0

#### 2.4.2 Молекула FH

Для молекули HF (CASSCF(2,2)/ cc-pVDZ) ми провели два набора розрахунків – методом ядерної регресії та нелінійним МНК (Морзе). Особливістю вибірки було наявність пропущених даних на проміжній між'ядерній відстані (рис. 2.7). Тут можна бачити значну проблему методу ядерної регресії жодний вибір  $h$  нажаль не дав адекватного опису системи. Можна бачити, що після  $h = 300$  коефіцієнт кореляції



незначно зменшується (Табл. 2.4). Отже для задовільного опису даних із пропусками необхідно ліквідувати пропуски і додати недостатні точки.

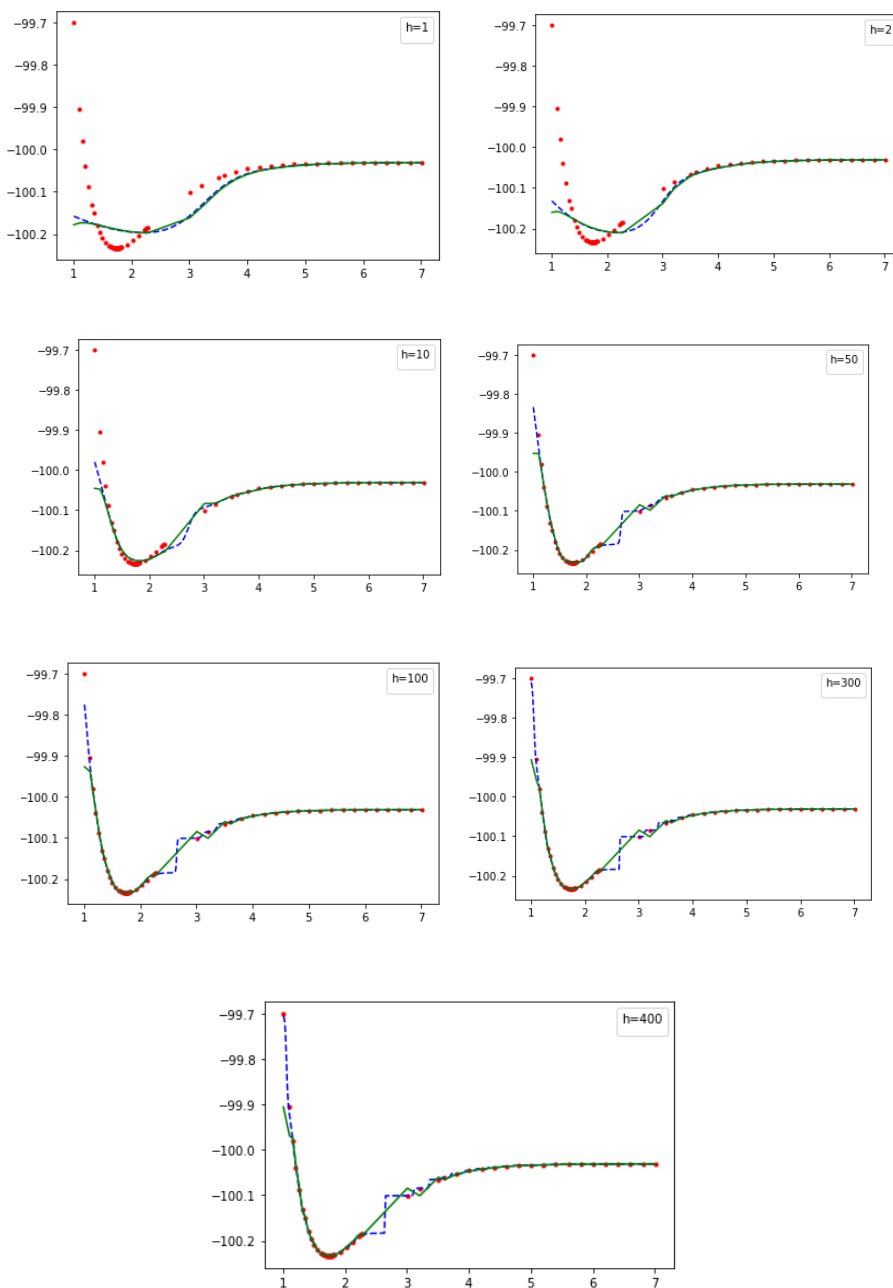


Рисунок 2.7. Апроксимація молекули HF методом ядерної регресії в залежності від параметру  $h$

Разом із тим розрахунки МНК Морзе виявились адекватними навіть у такому разі (Рис. 2.8, Табл. 2.4). Таким чином наведені графіки та таблиця що характеризують

вигляд апроксимуючої кривої для молекули HF при використанні двох зазначених методів, дозволяють зробити висновок, що метод МНК дає кращі результати при наявності пропусків.

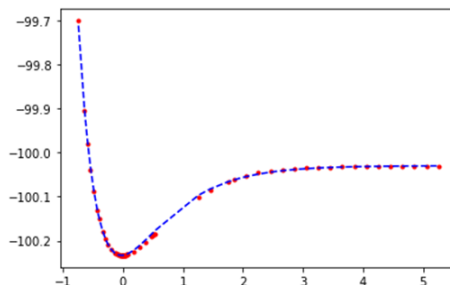


Рисунок 2.8. Апроксимація молекули HF МНК-Морзе

Таблиця 2.4. Параметри, що характеризують апроксимацію потенційної кривої молекули HF методом ядерної регресії та нелінійний МНК

Метод найменших квадратів		$Q^2$	$Q^2_2$	$\sigma^2$
		0.9992	0.9993	$1 \cdot 10^{-5}$
Метод ядерної регресії	$h$	$Q^2$	$Q^2_2$	$\sigma^2$
	1	0.4015	0.4081	$7.2 \cdot 10^{-3}$
	2	0.4831	0.4925	$6.2 \cdot 10^{-3}$
	10	0.7812	0.8022	$2.6 \cdot 10^{-3}$
	50	0.9036	0.9138	$1.2 \cdot 10^{-3}$
	100	0.9234	0.9324	$9.2 \cdot 10^{-4}$
	300	0.9333	0.9428	$8.0 \cdot 10^{-4}$
	400	0.9325	0.9422	$8.1 \cdot 10^{-4}$

## ВИСНОВКИ

1. На скриптовій мові Python розроблено комплекс програм для апроксимації нелінійних залежностей зокрема потенціальних кривих. Комплекс включає програми для апроксимації білінійної функції, функції Морзе та її узагальнення. Створено програму для реалізації методу ядерної регресії.
2. На ряді тестових прикладів проаналізовано можливості ядерної регресії та методу найменших квадратів для побудови нелінійних залежностей. Показано, що коефіцієнти детермінації отримані за процедурою LOO у купі з характеристиками кривизни нелінійної функції здатні бути критеріями що визначають параметр ядерної регресії  $h$ .
3. Методами ядерної регресії та нелінійним методом найменших квадратів проведено тестові розрахунки потенційних кривих ряду двохатомних молекул. Встановлено, що метод ядерної регресії здатен якісно представляти табличні дані щодо потенціальної енергії і може бути використаний для розв'язку радіального рівняння Шредингера.



## ПЕРЕЛІК ПОСИЛАНЬ

1. V. A. Epanechnikov, Nonparametric estimation of a multidimensional probability density, *Teor. Veroyatnost. i Primenen.*, 1969, Volume 14, Issue 1, 153–158.
2. E. Parzen, On estimation of a probability density function and more, *Ann. Math. Statist.*, 35 (1962), pp. 1065-1076.
3. M. Rosenblatt, Remarks on some nonparametric estimates of a density function, *Ann Math. Statist.*, 27 (1956), pp. 1065-1076.
4. G. M. Maniya, Remarks on nonparametric estimates of a bivariate probability density, *Soobshch. Akad. Nauk Gruz. SSR*, 27, 4 (1961), pp. 385-400.
5. E. A. Nadaraya, Estimation of a bivariate probability density, *Soobshch. Akad. Nauk Gruz. SSR*, 36, 2 (1964), pp. 267-268.
6. R. Bellman, I. Glicksberg and O. Gross, *Some Aspects of the Mathematical Theory of Control Processes*, Rand Corporation, R-313, 1958.
7. Altman, N. S. (1992). "An introduction to kernel and nearest neighbor nonparametric regression". *The American Statistician*. 46 (3): 175–185.
8. Cleveland, W. S.; Devlin, S. J. (1988). "Locally weighted regression: An approach to regression analysis by local fitting". *Journal of the American Statistical Association*. 83 (403): 596–610.
9. W. Zucchini, *Applied smoothing techniques, Part 1 Kernel Density Estimation.*, 2003.
10. Magen H., Kestner N. R., Meath W. J., et al. *Theory of intermolecular forces*. Pergamon Press, 1971; Meath W.J. et al. *Intermolecular forces*. B. Pullman 1978, p. 69. Mulliken R.S., Ermer W.C. *Diatomic Molecules, Results of Ab-Initio Calculations*, Academic Press, N.Y., 1977., *Polyatomic Molecules*, 1981. Patil S. H. *J. Phys. B*, 20, 3075, 1987.
11. N. D. Sokolov *Trends and problems in modern quantum chemistry. The successes of chemistry v. LVII*, N2, p.177-203, 1988.

12. B. O. Roos The Multiconfigurational Self-Consistent Field (MCSF) Theory. Lecture Notes in Quantum Chemistry, N7, Springer-Verlog. 1992.
13. A. M. Boichenko, at. all. Broadband continuums in inert gases and mixtures of inert gas with halogen molecules. J. Rus. Q. Electron., 1993, v.20 N1, p.7-30, 1993.
14. K. P. Lawley Dynamic of the exited states. J. Wiley Sons., N.Y.,1982.
15. Murrell J. N. Molecular potential energy functions. J. Wiley and Sons, N.Y., 1984.
16. D. A. Micha, Adv. Chem. Phys. 30, 7, 1975.
17. Clary D. C. The Theory of Chemical Reaction Dynamics. D. Reidel Publ. Co., Boston, 1986.
18. Hsu C. C., Pozdneev S. Int. Conf. Atomic and Molecular Processes in Fusion Plasma, Book of Abstracts. Nagoya, Japan, 1996.
19. T. K. Rebane, N. N. Penkina The scale transformation in quantum theory of atoms and molecules. - LGUpress, Leningrad, 1985.
20. Berens P. H., Wilson K. R., J. Comput. Chem., 4, N3, 313, 1983.
21. White D. N., J. Mol. Graphics, 3, N4, 136, 1985.
22. Pozdneev S. The Third International Conference on Computational Physics, Book of Abstracts, Chung-Li, Taiwan, 1995.
23. Grosdidier G. Comput. Phys. Comm., 52, 207, 1989.
24. M. G. Veselov Methods of the calculation of electronic structures of atoms and molecules. LGTJ-press, Leningrad, 1975.
25. V. A. Fock The main principles of the quantum mechanics. M., Nauka, 1976.
26. P. Gombas The Problem of many particles in quantum mechanics. M., 1953.
27. Кереселидзе Т. М., Фирсов О. Б., ЖЭТФ, 65, вып.1, 98, 1973.
28. Воронин А. И., Ошеров В. И., Динамика молекулярных реакций., М., Наука, 1991.

29. Дмитриева И. К., Зеничев В. А., Плиндов Г. Н., Препринт ИТМ АН БССР N1, 1986.
30. Sharp T. E., Atomic data, 2, 119, 1971.
31. Fayyazuddin M. R., Phys Lett., A 205, 383, 1995.
32. Коптев Г. С., Вестник МГУ, сер. 2, 35, 135, 1994.
33. Ackerman J., Helfrick K. Z., Phys. D, 18, 365, 1991.
34. Kyu Soo Jhung, In Ho Kim, Ki Hwan Oh., Phys. Rev., 42A, 6497, 1990.
35. Abrahamson A. A., Phys. Rev., 130, N2, 693, 1963.
36. Baylis W. E., J. Chem. Phys., 51, N6, 2665, 1969.
37. Хардле В. Прикладная непараметрическая регрессия. М.: Мир, 1993. - 349 с.
38. Акулич И.Л. Математическое программирование в примерах и задачах: Учеб. пособие для студентов эконом. спец. вузов. — М.: Высш. шк., 1986.
39. Амосов А. А., Дубинский Ю. А., Копченова Н. П. Вычислительные методы для инженеров : Учеб. пособие. — М. : Высшая школа, 1994. — 544 с.
40. Бахвалов Н. С., Жидков Н. П., Кобельков Г. Г. Численные методы. — 8-е изд. — М.: Лаборатория Базовых Знаний, 2000.

## ДОДАТОК А

Програма на скриптовій мові Python для апроксимації білінійної функції методом  
найменших квадратів

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
#   bln_nls.py
#----->
'''
    Nonlinear Least Squares for Bilinear function
'''
from math import exp,sqrt
from bln_nls_func import input_data, nls, correl,rnv,bilf
import matplotlib.pyplot as mp

n,x,y=input_data('bilinear3.dat')
#n,x,y=input_data('parabola.dat')

ksi=0.5

ai=-1 # initial guess
bi= 1
x0i=0.0

a,b,x0,yc=nls(n,x,y,ai,bi,x0i,ksi,50000,1.0e-8)
r,sig=correl(y,yc,n)
print ("%s %6.6f" %('R2=',r*r) )

mp.title('Bilinear Regression')
mp.scatter(x,y,marker='.',s=100,c='g')
mp.plot(x,yc,'-r',label='R$^2$='+rnv(r*100,4,7)+'\n'+
'$\sigma$='+rnv(sig,4,7)+'\n'+
'a='+rnv(a,4,7)+'\n'+
'b='+rnv(b,4,7)+'\n'+
'x$_0$='+rnv(x0,4,7) )
mp.legend(loc='best')
mp.savefig('Fig1.png')
mp.show()

#LOO

ai=a # initial guess
bi=b
```

```

x0i=x0

x1=[None]*(n-1)
y1=[None]*(n-1)
yp=[None]*n

for i in range(n):
    print ('LOO',i)
    k=-1
    for j in range(n):
        if i != j:
            k=k+1
            x1[k]=x[j]
            y1[k]=y[j]

    a,b,x0,yc=nls(n-1,x1,y1,ai,bi,x0i,ksi,50000,1.0e-8)
    yp[i]=bilf(a,b,x[i],x0)

r,sig=correl(y,yp,n)
print(r*r,sig)

#end

#!/usr/bin/env python3
# -*- coding: utf-8 -*-
#   bln_nls_func.py
#
from math import exp,sqrt,log

def rnv(x,m,mj):
    return str(round(x,m)).rjust(mj)

def bilf(a,b,x,x0):
    return -log( exp(a*(x-x0))+exp(b*(x-x0)))

def input_data(file1):
    fin=open(file1,'r')
    line1=fin.read(-1).split()
    fin.close()

    n=len(line1)
    n=int(n/2)

    x=[None]*n

```



```

y=[None]*n

j=0
for i in range(n):
    x[i]=float(line1[j])
    y[i]=float(line1[j+1])
    j+=2

return n,x,y
# Pearson correlation coeff
def correl(x,z,n):
    sx=sz=0.000e0
    for i in range(n):
        sx+=x[i]
        sz+=z[i]
    sx=sx/n
    sz=sz/n

    ss1=ss2=ss3=ss4=0.0000e0
    for i in range(n):
        x1=x[i]-sx
        z1=z[i]-sz
        ss1+=x1*z1
        ss2+=x1**2
        ss3+=z1**2
        ss4+=(x[i]-z[i])**2

    r=ss1/sqrt(ss2*ss3)
    sig=sqrt(ss4/n)
    return r,sig

# Gradients
def gradients(n,x,y,a,b,x0):
    grd=[None]*3

    g0=g1=g2=0
    for i in range(n):
        delt=x[i]-x0
        aaa=exp(a*delt)
        bbb=exp(b*delt)

        den=aaa+bbb
    #    print (i, den)

```

```

ch0=aaa*delt
ch1=bbb*delt
ch2=-a*aaa-b*bbb

t1=y[i]+log(aaa+bbb)

g0+=t1*ch0/den
g1+=t1*ch1/den
g2+=t1*ch2/den

grd[0]=g0
grd[1]=g1
grd[2]=g2

return grd

#----> diagonal Hessian
def hess_diag(n,x,y,a,b,x0):
    hd=[None]*3

    hd0=hd1=hd2=0
    for i in range(n):
        delt=x[i]-x0
        aaa=exp(a*delt)
        bbb=exp(b*delt)

        hd0+=(aaa*aaa*delt*delt + (y[i]+log(aaa+bbb))*(aaa*delt*delt*(aaa+bbb)-
aaa*aaa*delt*delt))/(aaa+bbb)**2
        hd1+=(bbb*bbb*delt*delt + (y[i]+log(aaa+bbb))*(bbb*delt*delt*(aaa+bbb)-
bbb*bbb*delt*delt))/(aaa+bbb)**2
        hd2+=((-a*aaa-b*bbb)**2+(y[i]+log(aaa+bbb))*((a*a*aaa+b*b*bbb)*(aaa+bbb)-(-
a*aaa-b*bbb)**2))/(aaa+bbb)**2

    hd[0]=hd0*2
    hd[1]=hd1*2
    hd[2]=hd2*2

    return hd

#----> Hessian
def hess(n,x,y,a,b,x0):

    fa=[None]*6
    hd=[None]*3

```

```

s00=s11=s01=s02=s12=s22=0.00000000e0
for i in range(n):
    delt=x[i]-x0
    aaa=exp(a*delt)
    bbb=exp(b*delt)

    # a-a
    s00+=(aaa*aaa*delt*delt + (y[i]+log(aaa+bbb))*(aaa*delt*delt*(aaa+bbb)-
aaa*aaa*delt*delt))/(aaa+bbb)**2
    # b-b
    s11+=(bbb*bbb*delt*delt + (y[i]+log(aaa+bbb))*(bbb*delt*delt*(aaa+bbb)-
bbb*bbb*delt*delt))/(aaa+bbb)**2
    # x0-x0
    s22+=((-a*aaa-b*bbb)**2+(y[i]+log(aaa+bbb))*((a*a*aaa+b*b*bbb)*(aaa+bbb)-(-
a*aaa-b*bbb)**2))/(aaa+bbb)**2
    # a-b
    s01+=delt*delt*aaa*bbb*(1-(y[i]+log(aaa+bbb)))/(aaa+bbb)**2
    # a-x0
    s02+=(-a*delt*aaa**2-b*delt*aaa*bbb+(y[i]+log(aaa+bbb))*(-
a*delt*aaa*bbb+b*delt*aaa*bbb))/(aaa+bbb)**2
    # b-x0
    s12+=(-a*delt*aaa*bbb-b*delt*bbb**2+(y[i]+log(aaa+bbb))*(-
b*delt*aaa*bbb+a*delt*aaa*bbb))/(aaa+bbb)**2

    hd[0]=s00*2
    hd[1]=s11*2
    hd[2]=s22*2

    dd=s00*s11*s22-s00*s12*s12-s01*s01*s22+2*s01*s02*s12-s02*s02*s11

    fa[0]=s11*s22-s12*s12
    fa[1]=-s01*s22+s02*s12
    fa[2]=s01*s12-s02*s11
    fa[3]=s00*s22-s02*s02
    fa[4]=-s00*s12+s01*s02
    fa[5]=s00*s11-s01*s01

    return hd, fa, dd

#-----> NLS
def nls(n,x,y,ai,bi,x0i,ksi,itmax,eps):

```

```

yc=[None]*n

a = ai
b = bi
x0=x0i

iter=0
accur=1
while iter <= itmax and accur > eps:
    grd=gradients(n,x,y,a,b,x0)
    accur=abs(grd[0])+abs(grd[1])+abs(grd[2])

    hd=hess_diag(n,x,y,a,b,x0)

    a -= ksi*grd[0]/hd[0]
    b -= ksi*grd[1]/hd[1]
    x0-= ksi*grd[2]/hd[2]

    iter+=1

for i in range(n):    #  calculated y[i]
    yc[i]=-log( exp(a*(x[i]-x0))+exp(b*(x[i]-x0)) )

text1= ('iter=',iter,'accuracy=', accur)
print ("%s %3.0f %s %3.1e" %text1)
print ('a=',a,' b=',b,' x0=', x0)
return a,b,x0,yc

```