УДК 621.391:519.24

Дегтяр О.С., аспірант

Оцінювання концентрацій хлорофілу за допомогою розв'язання задачі регуляризації Тихонова двоїстим методом

Київський національний університет імені Тараса Шевченка, 03680, м. Київ, пр-т. Глушкова 4д,
e-mail: olga.degtiar@gmail.com

O. Degtiar, postgraduate (PhD)

Method of chlorophyll concentration estimation using dual solution for ridge regression optimization

Taras Shevchenko National University of Kyiv, 03680, Kyiv, Glushkova st., 4d,
e-mail: olga.degtiar@gmail.com

Розроблено метод оцінки концентрацій хлорофілу на основі спектру відображення, використовуючи ретроспективні виміри концентрацій та відповідні їм спектри. Невідомі параметри знаходяться шляхом розв'язання задачі регуляризації Тихонова двоїстим методом. Наведено порівняння результатів обчислювального експерименту для прямого та двоїстого методів.

Ключові слова: структурно-параметрична оптимізація, псевдообернення, регуляризація, двоїстий метод.

The method for chlorophyll concentration estimation based on spectrum reflection develop. Mathematical model is built using retrospective concentration measurements and corresponding chlorophyll spectrum reflections taken in different moments of time. Relation between spectrum reflection and concentration value suggested being linear, so the task is to find unknown parameters' vector for linear combination of spectrum components that best interpolates given concentrations. The problem is ill-posed, till the matrix is underdetermined. Pseudoinverse methods with Tikhonov regularization are used for solving the problem. Dual solution for ridge regression optimization is used for estimating unknown values. The main benefit is decreasing unknown parameters' vector dimension. Computative experiment results deviation measurement for dual and primal methods shows quite the same results, while computative complexity of dual solution is much less costly, which is a big advantage for the given data. Afterwards the model is used for concentration prediction.

Key Words: structural and parametric optimization, pseudoinversion, regularization, dual solution.

Статтю представив д. т. н., проф. Гаращенко Ф.Г.

Вступ

Велика кількість задач апроксимації та прогнозування фізичних процесів і явищ зводяться до представлення експериментальних даних у вигляді комбінацій функцій з деякого структурно заданого класу таким чином, щоб найкраще, в певному сенсі, описати отримані дані. При цьому найкращим представленням вважатимемо нульове середньоквадратичне відхилення між вихідним сигналом та його знайденим наближенням.

При прогнозуванні задача ускладнюється неможливістю дослідити структуру даних, що прогнозуються, і порівняти її з структурою вихідних даних, на основі яких будується математична модель.

Одним із способів вирішення цієї проблеми є розбиття вихідної вибірки на декілька підвибірок довільним чином з подальшою побудовою моделі лише для однієї з них та інтерполяцією результатів на інші підвибірки. Це дає можливість ввести деякі апріорні оцінки на невідомі параметри моделі і покращити результати прогнозованих значень.

В статті розглядається задача оцінювання концентрацій хлорофілу в рослинності на основі спектру відображення. Виконано ряд експериментів з метою запису відображених спектрів для озимої пшениці з різними значеннями концентрацій хлорофілу.

Постановка задачі

Нехай маємо відомі N $q-$вимірних сигналів спектрального розподілу для N різних концентрацій хлорофілу, де $c=(c_1,c_2,...,c_N)^T$ - вектор значень концентрацій,

$$X=\begin{pmatrix} x_1^1 & x_2^1 & \dots & x_q^1 \\ x_1^2 & x_2^2 & \dots & x_q^2 \\ \dots & \dots & \dots & \dots \\ x_1^N & x_2^N & \dots & x_q^N \end{pmatrix}=\begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_N \end{pmatrix}$$

- матриця відповідних значень спектрального розподілу.

Для зручності позначимо $x=(x_1,x_2,...,x_N)$ - вектор-рядок, елементами якого є q-вимірні вектор-стовпчики спектральних даних для відповідних концентрацій.

Ставиться задача визначення невідомих концентрацій $(c_{N+1},c_{N+2},...,c_{N+M})^T$ на підставі отриманих спектральних даних $(x_{N+1},x_{N+2},...,x_{N+M})^T$ [3], а саме визначення невідомих параметрів $\alpha=(\alpha_1,\alpha_2,...,\alpha_q)^T$ лінійної моделі

$$c=X\alpha. \qquad (1)$$

З огляду на те, що матриця X не є квадратною ($N<q$), розв'язок задачі шукатиметься у вигляді $\alpha=X^+c$, де X^+ - псевдообернена матриця [1]. Тоді, використавши метод регуляризації Тихонова, де λ - параметр регуляризації, остаточно отримаємо вираз для α:

$$\alpha=(X^TX+\lambda I_q)^{-1}X^Tc. \qquad (2)$$

Для знаходження q-вимірного вектору невідомих параметрів α доведеться розв'язувати систему з q лінійних рівнянь з q невідомими. Складність такої задачі - $O(q^3)$.

В результаті отримаємо функцію апроксимації у вигляді

$$f(x)=\langle x,\alpha\rangle=X(X^TX+\lambda I_q)^{-1}X^Tc. \qquad (3)$$

Вище розглянуто пряму задачу знаходження апроксимації за допомогою введення параметру регуляризації [4].

З іншого боку, (2) можна переписати у вигляді [2]

$$X^TX\alpha+\lambda I_q\alpha=X^Tc, \qquad (4)$$

$$\alpha=\lambda^{-1}X^T(c-X\alpha)=X^Tw, \qquad (5)$$

$$\alpha(w)=\sum_{i=1}^{N}w_ix_i,\ w=\lambda^{-1}(c-X\alpha). \qquad (6)$$

Підставивши (5) в (6) отримаємо

$$\lambda w=(c-XX^Tw) \qquad (7)$$

або, перегрупувавши,

$$(XX^T+\lambda I_N)w=c. \qquad (8)$$

А отже, остаточний вираз для вектору невідомих параметрів буде мати вигляд

$$w=(G+w\lambda I_N)^{-1}c, \qquad (9)$$

де $G=XX^T$ (матриця Грамма).

Знаходження w зводиться до розв'язання N лінійних рівнянь з N невідомими. Складність такої задачі - $O(N^3)$.Результуюча апроксимація матиме вигляд

$$g(x,\alpha(w))=\langle x,\alpha(w)\rangle=\left\langle x,\sum_{i=1}^{N}w_ix_i\right\rangle= \\ =\sum_{i=1}^{N}w_i\langle x,x_i\rangle=c^T(G+\lambda I_N)^{-1}k, \qquad (10)$$

де $k = \langle x_i, x_i^j \rangle, i = \overline{1,N}, j = \overline{1,q}.$

Отже, для задач, в яких спостереження представлені у вигляді прямокутних матриць, використання двоїстої до задачі регуляризації Тихонова суттєво зменшує кількість операцій та прискорює обчислення, що особливо важливо при роботі з даними, що надходять в режимі реального часу.

Алгоритм знаходження параметрів моделі

Одним з методів відшукання параметру регуляризації для вихідної задачі є розбиття вихідної N-вимірної вибірки на n-вимірну тренуючу та m-вимірну валідаційну ($m+n=N$).

Нехай $c_{train} = (c_{train\,1}, c_{train\,2}, ..., c_{train\,n})^T$,

$$X_{train} = \begin{pmatrix} x^1_{1\,train} & x^1_{2\,train} & \dots & x^1_{q\,train} \\ x^2_{1\,train} & x^2_{2\,train} & \dots & x^2_{q\,train} \\ \dots & \dots & \dots & \dots \\ x^n_{1\,train} & x^n_{2\,train} & \dots & x^n_{q\,train} \end{pmatrix},$$

$x_{train} = (x_{train\,1}, x_{train\,2}, ..., x_{train\,n})$,

$c_{valid} = (c_{valid\,1}, c_{valid\,2}, ..., c_{valid\,m})^T$,

$$X_{valid} = \begin{pmatrix} x^1_{1\,valid} & x^1_{2\,valid} & \dots & x^1_{q\,valid} \\ x^2_{1\,valid} & x^2_{2\,valid} & \dots & x^2_{q\,valid} \\ \dots & \dots & \dots & \dots \\ x^m_{1\,valid} & x^m_{2\,valid} & \dots & x^m_{q\,valid} \end{pmatrix},$$

$x_{valid} = (x_{valid\,1}, x_{valid\,2}, ..., x_{valid\,m})$- відомі значення концентрацій та відповідних їм спектральних розподілів.

З лінійної моделі $c_{train} = X_{train}\alpha$ з (10) знаходиться апроксимація концентрацій для тренувальної вибірки

$$g(x_{train}, \alpha(w_{train})) = \langle x_{train}, \alpha(w_{train}) \rangle = \\ = c^T_{train}(G_{train} + \lambda I_n)^{-1} k_{train},$$

де $G_{train} = X_{train} X^T_{train}$, $k_{train} = \langle x_i, x_i^j \rangle$, $i = \overline{1,N}$, $j = \overline{1,q}$, $\lambda \in (0, \lambda_1]$, λ_1 обирається емпірично.

Для валідаційної вибірки використовується знайдений раніше q-вимірний вектор w_{train}. Тоді апроксимація концентрацій для валідаційної вибірки знаходиться з (10) і має вигляд $g(x_{valid}, \alpha(w_{train})) = \langle x_{valid}, \alpha(w_{train}) \rangle$.

Нижче наведено графіки середньоквадратичного відхилення фактичних значень концентрацій від відновлених за спектрами значень $\|c_{train} - g(x_{train}, \alpha(w_{train}))\|^2$ та $\|c_{valid} - g(x_{valid}, \alpha(w_{train}))\|^2$ для тренуючої та валідаційної вибірки для різних значень параметру регуляризації $\lambda \in (0, \lambda_1]$.

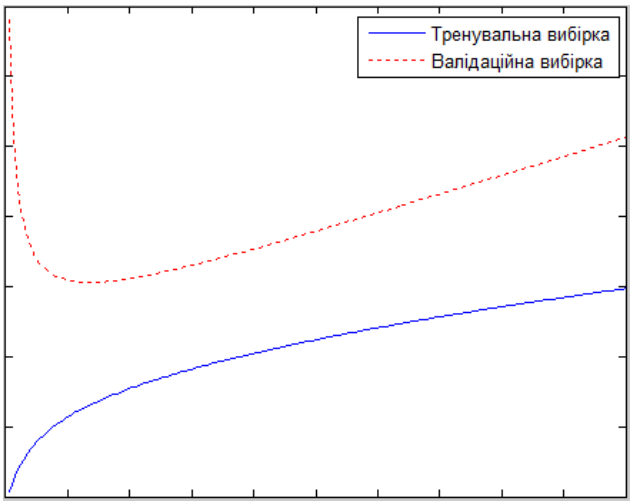


Рис. 1 Графік середньоквадратичного відхилення змодельованого розв'язку від реальних значень концентрацій

Для знаходження остаточного розв'язку вихідної задачі необхідно обрати значення параметру регуляризації λ_{opt}. Для задачі, що розглядається, пропонується обрати компромісне значення параметру для тренуючої та валідаційної вибірок, що є мінімумом відхилення валідаційної вибірки, який береться на розв'язках, знайдених для тренувальної вибірки

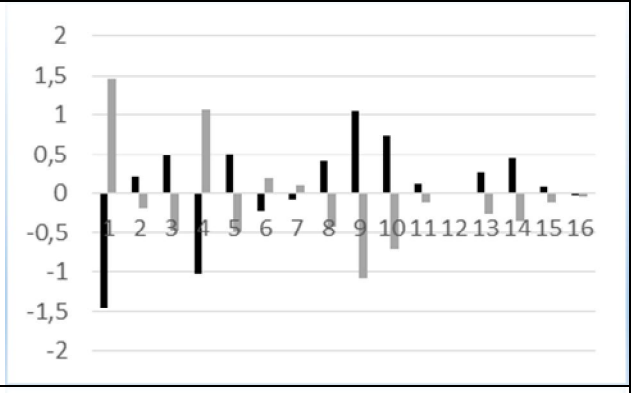
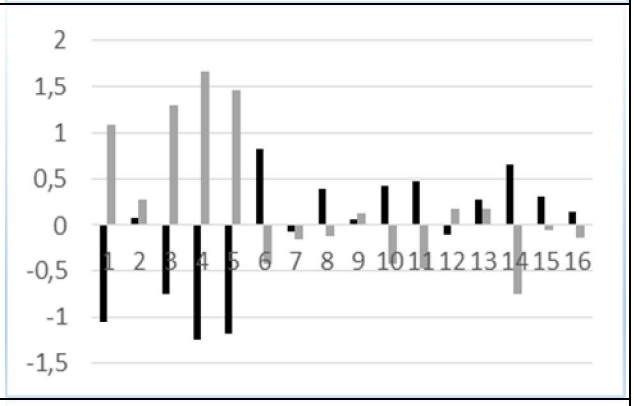
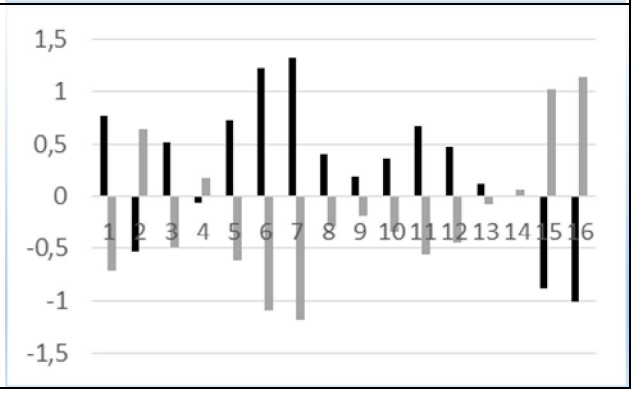
$$\lambda_{opt} = \arg\min_{\lambda \in (0, \lambda_1]} \|c_{valid} - g(x_{valid}, \alpha(w_{train}))\|^2. \quad (4)$$

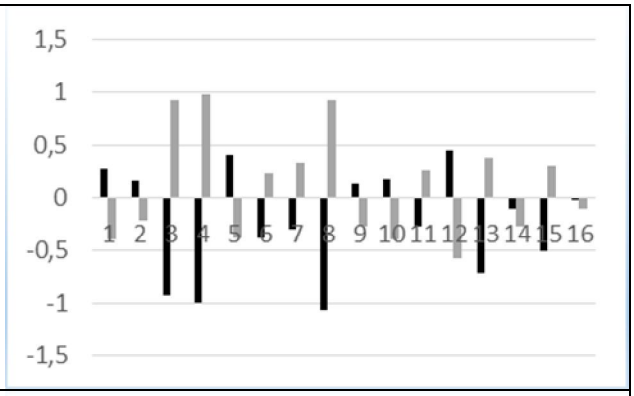
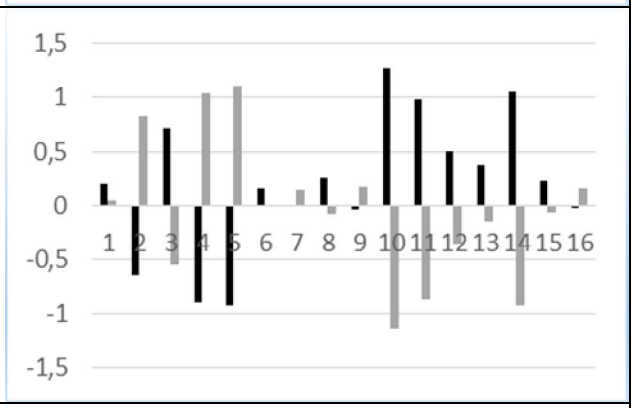
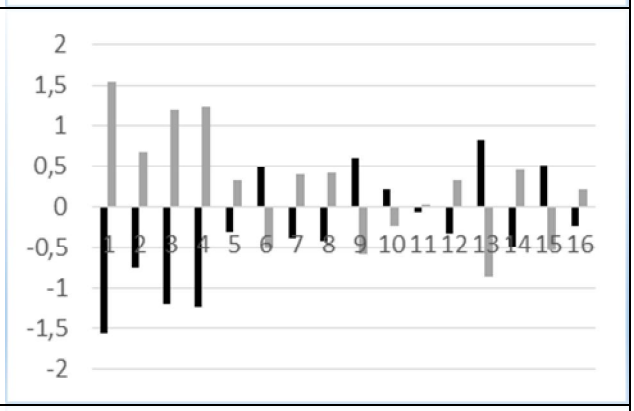
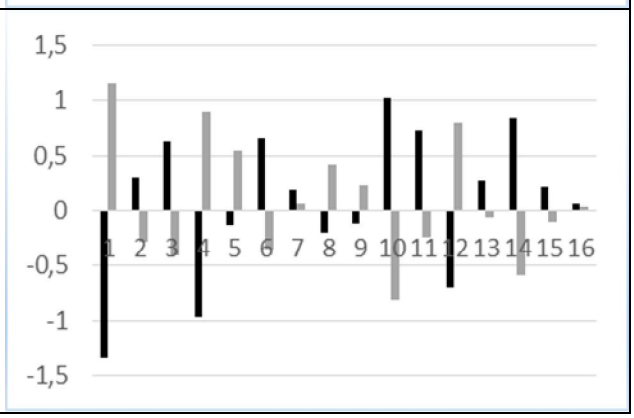
Результати обчислювального експерименту

Нижче наведено результати обчислювального експерименту для 7 різних комбінацій даних 40 352-вимірних ($q = 352$) спектральних розподілів та відповідних їм концентрацій: вибірки розбивалися на тренувальну ($n = 10$), валідаційну ($m = 14$) та тестову ($M = 16$) довільним чином.

Зокрема для кожної комбінації в таблиці наводяться знайдені за наведеним вище алгоритмом середньоквадратичні відхилення фактичних значень концентрацій від відновлених за спектрами значень для прямої ($\|c_{test} - f(x_{test})\|^2$)та двоїстої $\|c_{test} - g(x_{test}, \alpha(w_{train}))\|^2$ задач та графіки цих відхилень.

Результати обчислювального експерименту для прямої та двоїстої задачі

№	Пряма задача $\|c_{test} - f(x_{test})\|^2$	Двоїста задача $\|c_{test} - g(x_{test}, \alpha(w_{train}))\|^2$	■ пряма задача ■ двоїста задача
1	0,3702	0,3681	
2	0,4067	0,5718	
3	0,4899	0,4558	

4	0,2886	0,2595	
5	0,4317	0,4015	
6	0,5239	0,5147	
7	0,4179	0,3023	

За результатами експерименту можна зробити висновок, що прогнозовані значення концентрацій для тестової вибірки за допомогою двоїстої задачі здебільшого є більш точними ніж аналогічні результати з використанням прямої задачі (за винятком однієї вибірки). При цьому двоїста задача є суттєво кращою за часовою обчислювальною

складністю ($O(n^3)$ та $O(q^3)$ відповідно, де $n=10$, $q=352$).

Висновки

В роботі запропоновано алгоритм знаходження апроксимації відомих концентрацій хлорофілу на основі даних спектральних розподілів та оцінка невідомих концентрацій за допомогою двоїстої задачі до задачі регуляризації Тихонова.

Проведено обчислювальний експеримент для оцінки розв'язку з використанням двоїстої та прямої задачі, що показав ефективність використання двоїстої задачі для некоректно визначених, а саме недовизначених задач. Розглянутий алгоритм суттєво кращий за швидкодією, що вкрай важливо для задач, в яких дані надходять в режимі реального часу.

Список використаних джерел

1. *Albert A.* Regression, And The Moor-Penrose Pseudoinverse / A. Albert.– New York: Academic Press, 1972. – 224 c.
2. *Shave-Taylor J.* Kernel Methods for Pattern Analysis / J. Shave-Taylor, N. Cristiani.– New York: Cambridge University Press, 2004. – 478 p.
3. *Гаращенко Ф.Г.* Адаптивные модели аппроксимации сигналов в структурно-параметрических классах функций / Ф.Г. Гаращенко, О.Ф. Швец, О.С. Дегтяр // Проблемы управления и информатики. – 2011. – №2. – С. 69-77.
4. *Дегтяр О.С.* Алгоритм розв'язання задачі оцінювання концентрації хлорофілу за допомогою введення параметру регуляризації / О.С. Дегтяр // Вісник Київського національного університету імені Тараса Шевченка: Серія фізико-математичні науки. – 2013. – №4. – С. 104-107.

References

1. ALBERT, A. (1972) *Regression, And The Moor-Penrose Pseudoinverse*. New York: Academic Press.
2. SHAVE-TAYLOR, J. and CRISTIANI, N. (2004) *Kernel Methods for Pattern Analysis.* New York: Cambridge University Press.
3. GARASHCHENKO, F.G., SHVETS, O.F. and DEGTIAR, O.S. (2011) Adaptive models of approximation of signals in structural and parametric classes of functions *Problemy informatiki i upravleniya*. 2. p. 69-77.
4. DEGTIAR, O.S. (2013) Algorithm for solution of chlorophyll concentration estimation problem using regularization parameter *Bulletin of Taras Shevchenko National University of Kyiv: Series Physics & Mathematics.* 4. p. 104-107.

Надійшла до редколегії 22.05.14