УДК 004.02:519.254 doi: 10.15421/322103

С.М. Вовк
Дніпровський національний університет імені Олеся Гончара

УЗАГАЛЬНЕНА МАТЕМАТИЧНА МОДЕЛЬ ПРОЦЕСУ ОБРОБКИ ДАНИХ ДЛЯ СКЛАДНОГО ШУМОВОГО ОТОЧЕННЯ

Запропоновано узагальнену математичну модель процесу обробки даних для складного шумового оточення, сформованого шумом та аномальними значеннями. Модель подана узагальненим функціоналом, що підлягає мінімізації та враховує властивості розв'язку і відхилу розв'язку задачі обробки даних у припущенні лінійності системи їх формування. Зазначено особливості застосування та чисельної реалізації.

Ключові слова: математична модель, шумове оточення, аномальні значення.

S.M. Vovk
Oles Honchar Dnipro National University

GENERALIZED DATA PROCESSING MATHEMATICAL MODEL FOR COMPLICATED NOISE ENVIRONMENT

A generalized mathematical model of the data processing in a complex noise environment formed by noise and anomalous values is proposed. This model is represented by a generalized functional that needs to minimize.

The generalized functional allows one to take into account the various properties of both the solution and the solution residual of the data processing problem under the assumptions of the linearity of the data formation system, additivity of the measurement noise, and probabilistic nature of the anomalous values.

The basis of the generalized functional is the quasi-extent functional, which depends on three free parameters used to control its behavior to achieve its maximum efficiency. Applying this functional with a function that describes the solution residual of the data processing problem, one can obtain such a traditional criterion for the data processing as the minimum mean-square error criterion, minimum absolute deviation criterion, and others. Applying this functional with a function that describes a data processing problem solution, one can obtain the various types of solution regularization.

Among them are quadratic regularization, least-absolute-deviation regularization, and others. The conditions for obtaining the specific data processing models from the proposed generalized model are indicated, and their descriptions are given. The features of applying the quasi-extent functional to solve the problems of data approximation both with a linear and nonlinear parameter are discussed.

The range of the methods for numerical minimization is outlined. The expediency of applying a "greedy" strategy for selecting the best values from the set of parameter trial values generated by solving the corresponding equations or systems of equations is emphasized.

Keywords: mathematical model, noise environment, anomalous values.

С.М. Вовк
Днипровский национальный университет имени Олеся Гончара

ОБОБЩЕННАЯ МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ПРОЦЕССА ОБРАБОТКИ ДАННЫХ ДЛЯ СЛОЖНОЙ ШУМОВОЙ СРЕДЫ

Предложена обобщенная математическая модель процесса обработки данных для сложной шумовой среды, сформированной шумом и аномальными значениями. Модель представлена обобщенным функционалом, который подлежит минимизации и учитывает свойства решения и невязки решения задачи обработки данных в предположении линейности системы их формирования. Указаны особенности применения и численной реализации.

Ключевые слова: математическая модель, шумовая среда, аномальные значения.

Вступ. Розробка систем обробки вимірювальної інформації з об’єкта досліджень звичайно наштовхується на загальну проблему ефективної обробки отримуваних даних. Це продиктовано необхідністю мінімізації методичної похибки, зумовленої неадекватністю об’єкта вимірювання та його моделі, прийнятою при вимірюванні. На практиці розв’язання цієї проблеми є складним завданням як через розмаїття моделей шумів та ймовірних завад природного і штучного походження, так і через відсутність узагальненої моделі обробки спотворених даних, яка придатна до різних шумових оточень і до врахування властивостей отримуваних даних та/або властивостей розв’язку. Так, на даний час у повній мірі розв'язана проблема ефективної обробки даних, отримуваних у простому шумовому оточенні, сформованому шумом відомої статистичної природи. Але на даний час недостатньо досліджена проблема ефективної обробки даних, отримуваних у складному шумовому оточенні, сформованому шумом і аномальними значеннями. Це пояснюється тим, що статистична природа таких даних відома не повністю, що, з одного боку, обмежує застосування відомих статистичних підходів, але, з іншого боку, надає стимул до розробки нових критеріїв, математичних моделей, обчислювальних методів та алгоритмів обробки даних. Сьогодні основні напрямки досліджень у цій галузі становлять робастна обробка даних [7; 9; 12] і неквадратична регуляризація [4; 6; 8; 11].

Подана стаття присвячена розробці узагальненої математичної моделі процесу обробки даних для систем обробки вимірювальної інформації. Підґрунтям цієї моделі є вимога з мінімізації протяжності функції, використовуваної для пошуку розв’язку, яка означена як критерій мінімуму протяжності [1,2]. Для практичних застосувань найбільш ефективною варто вважати імплементацію цього критерію в нестрогій формі через концепцію вартісних функцій шляхом побудови їх певної "універсальної" множини [5; 10]. За цих умов, застосування зазначеного критерію до відхилу розв’язку відповідає ідеї робастної обробки даних, а до розв’язку чи до певної функції розв’язку – ідеї неквадратичної регуляризації. У сукупності це дозволяє побудувати узагальнену модель обробки даних, яка вдосконалює розробку систем обробки вимірювальної інформації з належними характеристиками як щодо шумового

оточення, так і щодо властивостей отримуваних даних та/або результатів їх обробки.

Постановка задачі. Нехай дані вимірювань сформовані лінійною системою та спотворені адитивним шумом і аномальними значеннями під час їх реєстрації, тобто:

$$\mathbf{g} = \begin{cases} \mathbf{Au} + \xi & with \quad probability \quad (1-p) \\ \eta & with \quad probability \quad p \end{cases}, \tag{1}$$

де $\mathbf{g}$ - вектор значень отриманих даних; $\mathbf{A}$ - матриця, що відповідає прямому лінійному оператору, який описує процес формування даних; $\mathbf{u}$ - вектор, що описує шуканий розв'язок задачі обробки даних; ξ - вектор, що описує випадкову реалізацію шуму; η - вектор, що описує випадкову реалізацію аномальних значень, які виникають незалежно з імовірністю p. Постановка задачі полягає в побудові узагальненої моделі процесу обробки отриманих даних (1), яка має поєднувати моделі обробки даних як для задач традиційної та робастної апроксимації даних заданими параметричними моделями, так і для обернених задач із прямим лінійним оператором за умов наявності в даних випадкових помилок, спричинених шумом, та аномальних значень.

Метод розв'язання. Уведемо у розгляд функціонал:

$$E^{+}[f(x)] = \int_{-\infty}^{\infty} \rho_S[f(x)]dx, \tag{2}$$

де вартісна функція [5; 10]

$$\rho_S(x) = k_S \cdot [(1+|x|^q/\alpha^q)^{\beta/q} - 1] \tag{3}$$

залежить від трьох вільних параметрів $\alpha > 0$, $-\infty < \beta \le 1$, $0 < q < \infty$ та де $\beta < q$, $k_S = 1/[(1+|x_{\text{Н}}|^q/\alpha^q)^{\beta/q} - 1]$ і $\rho_S(x_{\text{н}}) = 1$. Граничні переходи з (3) за вільними параметрами α, β, q для $x_{\text{Н}} = 1$ дають наступні результати:

$$\lim_{\alpha\to\infty}\rho_S(x) = |x|^q; \qquad \lim_{\alpha\to 0}\rho_S(x) = \begin{cases} |x|^{\beta}; & 0 < \beta \le 1 \\ \varphi(x); & -\infty < \beta \le 0 \end{cases};$$

$$\lim_{\beta\to\pm 0}\rho_S(x) = k_1 \cdot \ln(1+|x|^q/\alpha^q); \quad \lim_{\beta\to-\infty}\rho_S(x) = \varphi(x); \tag{4}$$

$$\lim_{\beta\to 1}\rho_S(x) = k_2 \cdot [(1+|x|^q/\alpha^q)^{1/q} - 1]; \quad \lim_{\substack{\beta\to-2 \\ q\to 2}}\rho_S(x) = k_3 \cdot \frac{x^2}{x^2+\alpha^2},$$

де $k_1 = 1/\ln(1+1/\alpha^q)$, $k_2 = 1/[(1+1/\alpha^q)^{1/q} - 1]$, $k_3 = 1+\alpha^2$ та де $\varphi(0) = 0$; $\varphi(x) = 1 \quad if \quad x \ne 0$ – "0-1" функція, що має сенс характеристичної функції множини. В [3] зазначено, що інтеграл Рімана від характеристичної функції множини може бути названий протяжністю цієї множини. З цієї причини в подальшому розгляді функціонал (2) згадується як функціонал квазіпротяжності множини ненульових значень функції $f(x)$, або скорочено як функціо-

нал квазіпротяжності функції $f(x)$. З (4), зокрема, випливає, що граничними випадками функціонала (2) є норми просторів L_p; $1 \le p < \infty$, квазінорми ненормованих просторів L_p; $0 < p < 1$, 1-норма (один-норма) простору L_0, а також інші функції. Через широкий спектр породжуваних функцій, множину вартісних функцій, яка утворюється на основі (3), було запропоновано називати "супермножиною вартісних функцій" [5; 10].

Вартісну функцію (3) можна модифікувати таким чином, щоб усунути залежність її параметра α, відповідального за згладжування значень аргументу, від параметрів β і q. Така модифікація полягає у перерахунку значення α відносно значення $\alpha_{\log}$, заданого для $\beta \to \pm 0$, з метою вирівнювання значень других похідних вартісних функцій в нулі, що для $x_{\text{Н}} = 1$ здійснюється шляхом розв'язання нелінійного рівняння:

$$\frac{\beta}{\alpha^q[(1+1/\alpha^q)^{\beta/q} - 1]} = \frac{q}{\alpha_{\log}^q \ln[1+1/\alpha_{\log}^q]}. \qquad (5)$$

З (5) випливає, що узагальнена вартісна функція модифікованої множини для випадку $\beta \to -\infty$ прямує до узагальненої функції Мєшалкіна: $\rho_M(x) = k_4 \cdot [1 - \exp(-|x|^q / \alpha_M^q)]$; де $k_4 = 1/[1 - \exp(-|x_{\text{Н}}|^q / \alpha_M^q)]$, $0 < q < \infty$, $\rho_M(x_{\text{н}}) = 1$ та де $\alpha_M > 0$ розраховується з $\alpha_{\log}$ шляхом розв'язання нелінійного рівняння, аналогічного (5). У свою чергу, для узагальненої функції Мєшалкіна виконуються такі граничні переходи (для $x_{\text{Н}} = 1$): якщо $0 < q < \infty$, то $\lim\limits_{\alpha \to 0} \rho_M(x) = \varphi(x)$ та $\lim\limits_{\alpha \to \infty} \rho_M(x) = |x|^q$; якщо $\alpha > 0$, то $\lim\limits_{q \to 0} \rho_M(x) = \varphi(x)$; якщо $0 < \alpha < x_{\text{Н}} < \infty$, то $\lim\limits_{q \to \infty} \rho_M(x)\big|_{|x|<\alpha} = 0$, $\lim\limits_{q \to \infty} \rho_M(x)\big|_{|x|=\alpha} = (e-1)/e \approx 0{,}63$ та $\lim\limits_{q \to \infty} \rho_M(x)\big|_{|x|>\alpha} = 1$. Останнє означає, що за цих умов мінімізація функціонала (2) для функції відхилу розв'язку відповідає критерію обробки даних, за яким необхідно знаходити максимум гістограми значень даних з шириною колодязів 2α. Утім для великих значень $q > 50$ зазначена множина і модифікована множина вартісних функцій майже не відрізняються, і для таких випадків побудова модифікованої множини не є доцільною. В цілому можна зробити висновок про те, що функціонал (2) дозволяє реалізовувати велику кількість критеріїв обробки даних, які можна вживати як до відхилу розв'язку (якщо функція $f(x)$ описує відхил розв'язку), так і до розв'язку задач (якщо функція $f(x)$ описує розв'язок) обробки даних. Варто зазначити, що для практичних чисельних розрахунків важливим є дискретний випадок представлення та обробки даних, для якого функціонал (2) приймає вигляд: $E^+[f_n] = \sum_{n=1}^{N} \rho_S[f_n]$, де N - кількість дискретних відліків f_n функції $f(x)$ вздовж осі аргументу x.

З урахуванням (1)-(5), узагальнену модель процесу обробки даних доцільно побудувати у вигляді:

$$\min_{\mathbf{u}}\{\gamma_1 E_1^+[\mathbf{Au}-\mathbf{g}]+\gamma_2 E_2^+[\mathbf{Tu}]\}, \qquad (6)$$

де $E_1^+[...]$ і $E_2^+[...]$ – функціонали квазіпротяжності відхилу розв'язку та розв'язку зі своїми значеннями вільних параметрів, відповідно; $\mathbf{T}$ - лінійний оператор (матриця), що задає апріорні відомості про розв'язок задачі обробки; γ_1 та γ_2 - вагові коефіцієнти. З урахуванням (4), узагальнена модель (6) може трансформуватися у багато різних моделей обробки даних, що надає гнучкість в її використанні. Зокрема, вона може трансформуватися у традиційну модель апроксимації даних без використання апріорних відомостей щодо функції розв'язку (коли $\gamma_2=0$, $\mathbf{A}=\mathbf{I}$, де $\mathbf{I}$ - тотожний оператор, а розв'язок $\mathbf{u}$ задано моделлю з невеликою кількістю невідомих параметрів), модель розв'язання обернених задач із прямим лінійним оператором, розв'язок яких має малу протяжність (коли $\mathbf{T}=\mathbf{I}$), модель чисельного диференціювання зашумлених даних з викидами (коли $\mathbf{A}$ - оператор антидиференціювання, $\mathbf{T}=\mathbf{D}$, де $\mathbf{D}$ – оператор диференціювання та $\alpha_2\to\infty$ й $q_2=2$), модель екстраполяції просторового спектра точкових джерел без урахування відхилу розв'язку (коли $\gamma_1=0$ і $\mathbf{T}=\mathbf{I}$, а розв'язок задано сумою Фур'є-перетворень відомої та екстрапольованої частин спектра), традиційну модель згладжування даних (коли $\mathbf{A}=\mathbf{I}$, $\mathbf{T}=\mathbf{D}^r$, де $\mathbf{D}^r$ – оператор диференціювання r–го порядку, $\alpha_1\to\infty$, $q_1=2$, $\alpha_2\to\infty$, $q_2=2$) та інші. Серед зазначених моделей практично важливими є математичні моделі процесів апроксимації та розв'язання обернених задач із прямим лінійним оператором.

Математична модель процесу апроксимації даних, яка побудована за критерієм мінімуму протяжності відхилу даних $\mathbf{g}$ від апроксимуючої функції $\mathbf{u}(\theta)$, де θ - вектор шуканих невідомих параметрів, випливає з (6) у припущенні $\gamma_2=0$ і $\mathbf{A}=\mathbf{I}$ та має таку стислу форму запису:

$$\min_{\theta} E^+[\mathbf{g}-\mathbf{u}(\theta)]. \qquad (7)$$

Для дискретного випадку розгорнутий запис (7) на основі (3) для $\beta\neq 0$ є:

$$\begin{aligned}\min_{\theta}\{k_S\cdot\sum_{n=1}^{N}[(1+|g_n-u_n(\theta)|^q/\alpha^q)^{\beta/q}-1]\}=\\ =\min_{\theta}\{k\cdot\sum_{n=1}^{N}[(1+|g_n-u_n(\theta)|^q/\alpha^q)^{\beta/q}]\}\end{aligned}, \qquad (8)$$

де $k=1$ для $0<\beta<1$ та $k=-1$ для $-\infty<\beta<0$, і для $\beta\to\pm 0$ є:

$$\min_{\theta}\sum_{n=1}^{N}\ln(1+|g_n-u_n(\theta)|^q/\alpha^q). \qquad (9)$$

На основі (7) у стислій формі запису можна подати окремі математичні моделі процесу апроксимації даних, які отримуються шляхом заміни $\mathbf{u}(\theta)$ на відповідні вирази моделей даних. Крім цього, використовуючи (4), з (8)-(9)

отримуються такі традиційні моделі, як $\min_{\theta} \|\mathbf{g}-\mathbf{u}(\theta)\|_2^2$, $\min_{\theta} \|\mathbf{g}-\mathbf{u}(\theta)\|_1$ та інші.

Математична модель процесу розв'язання обернених задач із прямим лінійним оператором, розв'язок яких має малу протяжність, випливає з (2) у вигляді задачі мінімізації функціонала, який є унормованою (через параметр регуляризації $\gamma = \gamma_2 / \gamma_1$) сумою функціоналів квазіпротяжності відхилу розв'язку та квазіпротяжності розв'язку. Її стисла форма запису є:

$$\min_{\mathbf{u}}\{E_1^+[\mathbf{Au}-\mathbf{g}]+\gamma E_2^+[\mathbf{u}]\}, \qquad (10)$$

а розгорнута форма запису отримується за допомогою (2) та (3). Видно, що для моделі (10) виникають дві групи вільних параметрів, де одна група відповідає функціоналу квазіпротяжності E_1^+, а друга група – функціоналу E_2^+. Крім цього, додатковим параметром налаштування такого процесу обробки даних є параметр регуляризації γ. Враховуючи граничні переходи (4), з (10) отримуємо такі традиційні моделі: $\min_{\mathbf{u}}\{\|\mathbf{Au}-\mathbf{g}\|_2^2+\gamma\|\mathbf{u}\|_2^2\}$, $\min_{\mathbf{u}}\{\|\mathbf{Au}-\mathbf{g}\|_2^2+\gamma\|\mathbf{u}\|_1\}$ та інші. Модель (10) в певному сенсі узагальнює (7) і є основою для розробки методів розв'язання систем лінійних алгебраїчних рівнянь з розрідженим вектором розв'язку та з погрішностями і грубими помилками у векторі вільних членів, а також методів екстраполяції просторового спектра точкових джерел.

Аналіз одержаних результатів. Основними властивостями функціонала квазіпротяжності (2) є його невід'ємність та обмеженість для будь-якої обмеженої функції на скінченному інтервалі спостереження, а також його можливість трансформуватися у інші, зокрема, традиційні функціонали обробки даних шляхом зміни значень його вільних параметрів. Останнє визначає окремі властивості функціонала квазіпротяжності для заданих значень його вільних параметрів. Такими, наприклад, є наступні. Якщо $\alpha = 0;\ 0 < \beta < 1$, то функціоналом квазіпротяжності є квазінорма $\|...\|_{0<\beta<1}$, яка за умови відсутності шуму в даних утворює мінімуми цільової функції у формі гострих «шипів» донизу, а за умови наявності шуму – у формі шипів донизу зі згладженим вістрям. Аналітично було отримано, що для лінійного параметра моделі даних і випадку адитивного спотворення шумом його справжнього значення згладжування шипів відбувається у відповідності із законом розподілу шуму. За цих умов математичне очікування функціонала квазіпротяжності за множиною реалізацій шуму є його згорткою із законом розподілу шуму, яка веде до розмиття мінімумів зі зменшенням їх глибини та їх розширенням. Чисельне моделювання підтвердило ефект згладжування мінімумів як для цього випадку, так і для випадку адитивного шуму, що накладався на дані. Аналогічний ефект згладжування отримано також і для шумів імпульсного типу із законом розподілу Коші та із меридіанним законом розподілу. Для нелінійного параметра, яким була частота синусоїди, функціонал квазіпротяжності через обмеженість значень синусоїди породжував цільову функцію з великою кіль-

кістю неглибоких мінімумів, які складали певний фоновий рівень її значень, та з відносно вузьким і глибоким глобальним мінімумом, глибина якого суттєво зменшувалася тільки за умови суттєвого зростання шуму (коли відношення амплітуди синусоїди до параметра масштабу шуму убувало до одиниці). Утім для нелінійного параметра, яким був коефіцієнт затухання в експоненціальній моделі даних, функціонал квазіпротяжності породжував цільову функцію з необмеженим зростанням її значень у разі необмеженого зростання відхилу даних від їх моделі. Загалом, залежно від моделі даних та від значень своїх вільних параметрів функціонал квазіпротяжності мав відповідну поведінку в околицях локальних мінімумів та "на нескінченності". Крім того, функціонал (2) уможливлював отримання декількох значень невідомого параметра.

Аналіз ефективності методів чисельної мінімізації функціонала квазіпротяжності показав, що коло цих методів обмежено методами оптимізації нульового порядку, які придатні для знаходження декількох локальних мінімумів, а також методами градієнтного спуску та спряжених градієнтів. Чисельну реалізацію цих методів доцільно ґрунтувати на «жадібній» стратегії використання множини пробних значень шуканих параметрів, які отримуються в результаті розв'язання відповідних рівнянь чи систем рівнянь, або множини пробних значень кроку уздовж напрямку спуску.

Висновки. Запропоновано та обґрунтовано узагальнену модель процесу обробки даних, виконуваної через розв'язання апроксимаційних задач та обернених задач із прямим лінійним оператором, яка враховує розмаїття моделей шумів і ймовірних завад та властивості отримуваних даних або результатів їх обробки. Аналітичні виведення та чисельне моделювання розкривають широкі потенційні можливості використовуваного в цій моделі функціонала квазіпротяжності для розв'язання задач із лінійними та нелінійними параметрами за відсутності і за наявності шуму різної статистичної природи та аномальних значень.

Запропонована узагальнена модель дозволить підвищити ефективність процесу обробки даних для складних шумових оточень, сформованих шумом і аномальними значеннями, та удосконалить розробку систем обробки вимірювальної інформації з належними характеристиками щодо поточних шумових оточень й властивостей даних та/або результатів їх обробки.

Бібліографічні посилання

1. Вовк С.М. Постановка задач обработки данных на основе критерия минимума протяженности. *Радіоелектроніка, інформатика, управління*. 2019. № 1. С. 157–166.
2. Вовк С.М. Применение функционала квазипротяженности в задачах аппроксимации искаженных данных. *Системні технології*. 2020. Вип. 5 (130). С. 79 – 87.
3. Титчмарш Е. Теория функций. М.: Наука, 1980. 464 с.
4. Bardsley J.M., Howard M. L1-regularized inverse problems for image deblurring via bound- and equality-constrained optimization. *Research in Shape Analysis. Association for Women in Mathematics Series*. 2018. Vol. 12. P.1-16.

5. Borulko V.F., Vovk S.M. Minimum-duration filtering. *Radio Electronics, Computer Science, Control*. 2016. № 1. P. 7–14.
6. Guo W., Lou Y., Qin J., Yan M. A Novel Regularization Based on the Error Function for Sparse Recovery. *Journal of Scientific Computing*. 2021. Vol. 87, 31.
7. Huang X., Zhang L., Chen Z., Zhao R. Robust detection and motion parameter estimation for weak maneuvering target in the alpha-stable noise environment. *Digital Signal Processing*. 2021. Vol. 108, 102885.
8. Nikolova M. Relationship between the optimal solutions of least squares regularized with L0-norm and constrained by k-sparsity. *Applied and Computational Harmonic Analysis*. 2016. Vol. 41, N. 1. P. 237–265.
9. Rousseeuw P. J., Hubert M. Anomaly detection by robust statistics. *WIREs Data Mining Knowledge Discovery*. 2018. Vol. 8, N. 2. P. 1-14.
10. Vovk S.M. General approach to building the methods of filtering based on the minimum duration principle. *Radioelectronics and Communications Systems*. 2016. V. 59, N. 7. P. 281–292.
11. Wang M., Wang Q., Chanussot J., Hong D. L0-L1 Hybrid Total Variation Regularization and Its Applications on Hyperspectral Image Mixed Noise Removal and Compressed Sensing. *IEEE Transactions on Geoscience and Remote Sensing*. 2021. V. 59, N. 9. P. 7695-7710.
12. Wen F., Liu P., Liu Y., Qiu R.C., Yu W. Robust sparse recovery in impulsive noise via Lp-L1 optimization. *IEEE Transactions on Signal Processing*. 2017. Vol. 65, N. 1. P. 105-118.

Надійшла до редколегії 15.06.2021.