

## МЕТОД ЧИСЛЕННОГО ДИФФЕРЕНЦИРОВАНИЯ ЗАШУМЛЕННЫХ ДАННЫХ С ВЫБРОСАМИ

**Актуальность.** Применение традиционных методов численного дифференцирования к зашумленным данным с выбросами приводит к значительным ошибкам. Объектом данного исследования является процесс численного дифференцирования таких данных.

**Цель работы** – разработка метода численного дифференцирования зашумленных данных с выбросами, который позволяет получить гладкую аппроксимацию их первой производной и, соответственно, гладкую аппроксимацию самих данных.

**Метод.** Предложенный метод численного дифференцирования основан на решении задачи минимизации сглаживающего функционала, который построен на критерии минимума протяженности невязки решения и ограничении энергии первой производной решения. Критерий минимума протяженности задает основную часть функционала и обеспечивает его устойчивое поведение по отношению к аддитивному шуму и выбросам. Ограничение задает стабилизирующую часть функционала и обеспечивает гладкость решения задачи. Вклад указанных частей регулируется с помощью параметра регуляризации. Поскольку основная часть сглаживающего функционала не является выпуклой, то задача его минимизации является задачей нелинейного невыпуклого программирования. Для численного решения этой задачи используется метод сопряженных градиентов, в котором величина шага вдоль направления спуска определяется на множестве пробных шагов. Эти шаги минимизируют отдельные компоненты основной и стабилизирующей частей сглаживающего функционала, что позволяет переходить из одного локального минимума функционала в другой более глубокий локальный минимум.

**Результаты.** Моделирование задачи численного дифференцирования зашумленных данных с выбросами и обработка экспериментальных данных, которые представляли собой спектры фотолуминесценции при наличии в их составе узких линейчатых спектральных составляющих, подтвердили эффективность предложенного метода.

**Выводы.** Предложенный метод может быть использован для численного дифференцирования зашумленных данных с выбросами. При этом он позволяет получить гладкую аппроксимацию первой производной исходных данных и гладкую аппроксимацию самих исходных данных. Данный метод можно обобщить на случай негладкого решения путем построения стабилизирующей части функционала на основе ограничения полной вариации решения.

**Ключевые слова:** численное дифференцирование, гладкая аппроксимация, критерий минимума протяженности.

### НОМЕНКЛАТУРА

**A** – оператор (матрица) антидифференцирования;

**D** – оператор (матрица) дифференцирования;

**E** – функционал;

**N** – количество дискретных отсчетов;

$a_{jn}$  – элемент матрицы **A**;

$a_{kj}^T$  – элемент матрицы  $\mathbf{A}^T$ , которая получена транспонированием матрицы **A**;

**e** – вспомогательный вектор;

$e_j$  – элемент вектора **e**;

$d_{jn}$  – элемент матрицы **D**;

$\mathbf{grad}^{(t)}$  – вектор градиента на  $t$ -ой итерации;

$h^{(t)}$  – шаг вдоль направления спуска на  $t$ -ой итерации;

и;

$\mathbf{p}^{(t)}$  – вектор направления спуска на  $t$ -ой итерации;

$q$  – свободный параметр, который отвечает за степень сглаживания значений;

$t$  – номер итерации;

**u** – вектор искомого решения;

$\mathbf{u}^{(t)}$  – вектор решения на  $t$ -ой итерации;

$\mathbf{u}'$  – вектор первой производной искомого решения;

$u_n$  – элемент вектора **u**;

$x$  – независимая переменная;

$\hat{x}$  – точка нормировки функции  $\psi^{(\alpha, \beta, q)}(x)$  на единицу;

© Вовк С. М., 2017

DOI 10.15588/1607-3274-2017-3-5

**u** – вектор исходных зашумленных данных с выбросами;

$\alpha$  – свободный параметр, который отвечает за величину сглаживания значений;

$\beta$  – свободный параметр, который отвечает за форму закона распределения

$\gamma$  – параметр регуляризации;

$\Phi$  – одномерная целевая функция;

$\Psi$  – неквадратичный функционал, который реализует понятие «протяженность»;

$\Psi$  – стоимостная функция для построения функционала  $\Psi$ ;

$\|\dots\|^2$  – квадрат евклидовой нормы;

АЦП – аналого-цифровой преобразователь.

### ВВЕДЕНИЕ

Операция дифференцирования является мощным математическим инструментом анализа данных. Она составляет основу многих современных методов спектроскопии [1], астрофизики [2], обработки изображений [3] и мониторинга движения объектов [4]. Поскольку операция дифференцирования является неустойчивой и, следовательно, напрямую не применимой на практике из-за наличия шума и дискретности данных, то были разработаны различные методы численного дифференцирования [5]. При этом одни из них (полиномиальная аппроксимация методом наименьших квадратов, регуляризация Тихонова, сглаживание сплайнами) предполагают гладкое решение [5], а другие (метод регуляриза-

ции на основе ограничения полной вариации решения) предполагают даже разрывное решение [6]. Однако указанные методы плохо работают в случае, когда исходные данные искажены не только шумом, но и выбросами [7]. Это проявляется в виде ошибочных осцилляций решения из-за расплывания выбросов и является следствием того, что функционал невязки решения формируется на основе квадратичной метрики в условиях равноправности вкладов зашумленных элементов данных и выбросов. Для устранения этого недостатка такое равноправие должно быть нарушено.

Целью данной работы является разработка метода численного дифференцирования зашумленных данных с выбросами для получения гладкой аппроксимации их первой производной и, соответственно, гладкой аппроксимации самих данных. Предложенная постановка задачи заключается в минимизации сглаживающего функционала, который состоит из основной и стабилизирующей частей, а вклад этих частей регулируется с помощью параметра регуляризации. При этом основная часть функционала построена на базе критерия минимума протяженности [8]–[10], который применяется к невязке решения соответствующего интегрального уравнения Вольтерры и который обеспечивает уменьшение вклада выбросов. Стабилизирующая часть функционала построена на основе ограничения квадрата евклидовой нормы первой производной решения и обеспечивает получение гладкого решения задачи. Далее в работе приводится математическая постановка решаемой задачи как в общем, так и в развернутом виде, а также численный метод ее решения на основе метода сопряженных градиентов. В заключительной части работы приводятся результаты численного моделирования и обработки экспериментальных данных.

## 1 ПОСТАНОВКА ЗАДАЧИ

Пусть функция  $g(x)$  непрерывна в интервале  $[a, b]$  и имеет производную

$$u(x) = g'(x) = \frac{dg(x)}{dx}; \quad x \in [a, b], \quad (1)$$

причем  $g(a)$  задано. Тогда задачу определения производной (1) можно сформулировать как задачу решения линейного интегрального уравнения Вольтерры первого рода [5]

$$Au(x) = \int_a^x u(z) dz = \hat{g}(x); \quad x \in [a, b], \quad (2)$$

где  $\hat{g}(x) = g(x) - g(a)$  [6]. Если  $u(x)$  является непрерывной функцией, то  $g(x)$  является гладкой функцией. Далее будем считать, что  $u(x)$  также является гладкой функцией, полагая существование производной  $u'(x)$ .

Пусть гладкая функция  $f(x)$  аппроксимирует функцию  $g(x)$ . Тогда, если  $u(x)$  найдено, то  $f(x)$  можно получить с помощью формулы:  $f(x) = g(a) + \int_a^x u(z) dz$ .

Пусть значения функции  $g(x)$  в  $N$  дискретных точках  $a \leq x_1 \leq \dots \leq x_N \leq b$  искажены значениями  $\xi_i$  адди-

тивного шума с выбросами, образуя числовой вектор исходных данных  $\mathbf{y}$  с элементами

$$y_i = g(x_i) + \xi_i; \quad i = 1, \dots, N. \quad (3)$$

Тогда предлагаемая постановка задачи численного дифференцирования имеет следующий общий вид:

$$\arg \min_{\mathbf{u}} \left\{ \Psi[\mathbf{A}\mathbf{u} - \mathbf{y}] + \gamma \|\mathbf{u}'\|^2 \right\}, \quad (4)$$

где  $\mathbf{u}' = \mathbf{D}\mathbf{u}$ , а неквадратичный функционал  $\Psi$  реализует концепцию «протяженность» на основе стоимостной функции [10]

$$\psi^{(\alpha, \beta, q)}(x) = k^{(\alpha, \beta, q)} [(1 + |x|^q / \alpha^q)^{\beta/q} - 1]; \quad |x| < \infty, \quad (5)$$

где  $\alpha > 0$ ;  $-\infty < \beta \leq 1$ ;  $0 < q < \infty$ ;  $\beta < q$ ;

$$k^{(\alpha, \beta, q)} = 1 / [(1 + |\hat{x}|^q / \alpha^q)^{\beta/q} - 1]; \quad \psi^{(\alpha, \beta, q)}(\hat{x}) = 1;$$

$\hat{x} \neq 0$ . Для дискретного случая  $\Psi[\mathbf{e}] = \sum_{j=1}^N \psi^{(\alpha, \beta, q)}(e_j)$ .

Подставляя это выражение вместе с (5) в (4) и задавая линейные операторы  $\mathbf{A}$  и  $\mathbf{D}$  с помощью соответствующих матриц, получим постановку задачи (4) в развернутом виде:

$$\arg \min_{u_1, \dots, u_n} \left\{ k^{(\alpha, \beta, q)} \sum_{j=1}^N [(1 + \sum_{n=1}^N a_{jn} u_n - y_j |^q / \alpha^q)^{\beta/q} - 1] + \gamma \sum_{j=1}^N \sum_{n=1}^N d_{jn} u_n^2 \right\}. \quad (6)$$

Подчеркнем, что для  $0 < \beta \leq 1$  имеем  $k^{(\alpha, \beta, q)} > 0$ , а для  $-\infty < \beta < 0$  имеем  $k^{(\alpha, \beta, q)} < 0$ . Однако для  $\beta = 0$  запись задачи (6) и коэффициента  $k^{(\alpha, \beta, q)}$  преобразуется в запись на основе логарифмов [10]. Поэтому для практического использования задачу (6) следует переписать в виде двух задач оптимизации. Так, отбрасывая несущественные коэффициенты, для случая  $\beta \neq 0$  из (6) имеем задачу

$$\arg \min_{u_1, \dots, u_N} \left\{ \text{sgn}(\beta) \sum_{j=1}^N [(1 + \sum_{n=1}^N a_{jn} u_n - y_j |^q / \alpha^q)^{\beta/q}] + \gamma_1 \sum_{j=1}^N \sum_{n=1}^N d_{jn} u_n^2 \right\}, \quad (7)$$

где  $\gamma_1 = \gamma |(1 + |\hat{x}|^q / \alpha^q)^{\beta/q} - 1|$  и где  $\text{sgn}(\beta) = 1$  для  $0 < \beta \leq 1$  и  $\text{sgn}(\beta) = -1$  для  $-\infty < \beta < 0$ . Для случая  $\beta = 0$  из (6) путем предельного перехода по  $\beta \rightarrow 0$  получаем задачу

$$\arg \min_{u_1, \dots, u_N} \left\{ \sum_{j=1}^N \ln [1 + \sum_{n=1}^N a_{jn} u_n - y_j |^q / \alpha^q] + \gamma_2 \sum_{j=1}^N \sum_{n=1}^N d_{jn} u_n^2 \right\}, \quad (8)$$

где  $\gamma_2 = \gamma \ln(1 + |\hat{x}|^q / \alpha^q)$ . Таким образом, (7) и (8) есть частные случаи (6).

## 2 ОБЗОР ЛИТЕРАТУРЫ

Основу численного дифференцирования составляют методы аппроксимации [5]. Так, если в окрестности точки  $x$  функция  $g(x)$  приближается некоторой другой функцией  $f(x)$ , для которой производная  $f'(x)$  в точке  $x$  легко вычисляется, то полагают  $g'(x) \approx f'(x)$ . Наиболее просто такое приближение реализуется с помощью полиномиальной интерполяции. Однако при малой величине шага соответствующие формулы численного дифференцирования становятся плохо обусловленными и при наличии шума в исходных данных приводят к некорректным результатам [5]. Хотя выбор оптимального значения шага решает проблему устойчивости этих вычислений [4], однако при этом остаются следующие недостатки: 1) значения исходной функции могут быть не доступны для всех желаемых значений шага; 2) реализуется кусочно-полиномиальная интерполяция, которая не обеспечивает гладкость решения в точках стыковки полиномов.

Современными методами численного дифференцирования зашумленных данных являются [5]: 1) полиномиальная аппроксимация на базе метода наименьших квадратов; 2) регуляризация Тихонова; 3) сглаживание сплайнами; 4) свертка со сглаживающим ядром Фридриха (Friedrichs mollifier); 5) вариационный метод Ноулеса и Уоллеса (Knowles and Wallace); 6) регуляризация на основе ограничения полной вариации решения (total variation regularization). Сравнение первых пяти методов показало преимущество метода регуляризации Тихонова, который заключается в минимизации выпуклого функционала [5]

$$E(\mathbf{u}) = \|\mathbf{A}\mathbf{u} - \mathbf{y}\|^2 + \gamma \|\mathbf{D}_k \mathbf{u}\|^2, \quad (9)$$

с дифференциальным оператором  $\mathbf{D}_k$  порядка  $k$  [5]. Важным является то, что задача (9) имеет аналитическое решение [5]

$$\mathbf{u} = (\mathbf{A}^T \mathbf{A} + \gamma \mathbf{D}_k^T \mathbf{D}_k)^{-1} \mathbf{A}^T \mathbf{y}. \quad (10)$$

Метод регуляризации на основе ограничения полной вариации решения предназначен для случаев, когда искомая производная имеет разрывы. Данный метод заключается в минимизации функционала [5]

$$E(\mathbf{u}) = \frac{1}{2} \|\mathbf{A}\mathbf{u} - \mathbf{y}\|^2 + \gamma \sum_{j=1}^{n-1} \sqrt{(u_{j+1} - u_j)^2 + \varepsilon}, \quad (11)$$

в котором малая положительная величина  $\varepsilon$  необходима для дифференцируемости (11) при  $\mathbf{u}' = 0$  [6]. Задача (11) также сформулирована в [2], где указано, что путем изменения  $\varepsilon$  стабилизирующая часть функционала (11) меняется от  $L_1$ -нормы производной решения  $\mathbf{u}'$ , что заставляет решение  $\mathbf{u}$ , быть кусочно-постоянным, до  $L_2$ -нормы  $\mathbf{u}'$ , что заставляет  $\mathbf{u}$  быть гладким. С геометрической точки зрения, стабилизирующая часть функционала (11) задает длину кривой, которая описывает график функции  $u(x)/\sqrt{\varepsilon}$ . Минимизация длины этой кривой приводит к ее спрямлению и уменьшает осцилляции в решении.

Однако методы, которые основаны на решении задачи (2) в рамках квадратичных метрик, дают плохие результаты для случая, когда исходные данные искажены выбросами. Это обусловлено расплыванием выбросов из-за их сглаживания, причем чем больше количество выбросов и чем больше их амплитуды, тем хуже результат численного дифференцирования. Для подавления выбросов обычно применяют метод медианной фильтрации [4]. Однако более эффективными методами обработки зашумленных данных с выбросами являются методы «мирадной» [11] и «меридианной» [12] фильтрации. Эти методы разработаны для фильтрации шумов импульсного типа с «тяжелыми» хвостами их закона распределения, а именно шума с законом распределения Коши и шума с «меридианным» законом распределения. В [10] дано обобщение этих методов и предложен общий подход к построению методов фильтрации, который основан на требовании минимизировать временную протяженность (длительность) получаемого решения. Эффективность этого подхода достигается путем настройки его трех свободных параметров, которые связаны с масштабом (дисперсией) шума, величиной «тяжести» хвостов и формой закона распределения. При этом возможна настройка как на обычный шум с законом распределения Гаусса или Лапласа, так и на шум импульсного типа с законом обобщенного распределения Коши, а также на комбинацию этих шумов. Данный подход кладется в основу предлагаемого метода численного дифференцирования зашумленных данных с выбросами.

## 3 МАТЕРИАЛЫ И МЕТОДЫ

Далее рассмотрим только задачу (8) по следующим причинам: 1) простота записи из-за отсутствия свободного параметра  $\beta$ ; 2) рассмотрение задачи (7) выполняется аналогично.

В общем случае задача (8) имеет только численное решение. Чтобы получить его, используем метод сопряженных градиентов, который задан с помощью следующей вычислительной схемы:

$$\begin{aligned} \mathbf{u}^{(t+1)} &= \mathbf{u}^{(t)} + h^{(t)} \mathbf{p}^{(t)}, \quad t \geq 0; \\ \mathbf{p}^{(0)} &= -\mathbf{grad}^{(0)}; \quad t = 0; \\ \mathbf{p}^{(t)} &= -\mathbf{grad}^{(t)} + b^{(t-1)} \mathbf{p}^{(t-1)}; \quad t \geq 1; \\ b^{(t-1)} &= \|\mathbf{grad}^{(t)}\|^2 / \|\mathbf{grad}^{(t-1)}\|^2; \\ h^{(t)} &= \arg \min_{h \geq 0} \Phi(\mathbf{u}^{(t)} + h \mathbf{p}^{(t)}). \end{aligned} \quad (12)$$

В (12) одномерная целевая функция  $\Phi(\mathbf{u}^{(t)} + h \mathbf{p}^{(t)})$  зависит от  $h$  и получается подстановкой элементов вектора  $\mathbf{u}^{(t)} + h \mathbf{p}^{(t)}$  в выражение, записанное в фигурных скобках задачи (8). Далее рассмотрим вопросы построения вектора градиента и минимизации одномерной целевой функции для случая  $q = 2$ , который обеспечивает оптимальную (в смысле критерия максимального правдоподобия) обработку данных, искаженных шумом Коши [10].

Из необходимого условия минимума (8) для случая  $q = 2$  получаем систему из  $N$  нелинейных уравнений

$$\sum_{j=1}^N a_{kj}^T \frac{v_j}{v_j^2 + \alpha^2} + \gamma_2 \sum_{n=1}^N \hat{d}_{kn} u_n = 0; \quad k = 1, \dots, N, \quad (13)$$

где  $v_j = (\sum_{n=1}^N a_{jn} u_n - y_j)$  задает элемент вектора невязки

$$\mathbf{v} = \mathbf{A}\mathbf{u} - \mathbf{y}, \quad \hat{d}_{kj} = \sum_{n=1}^N d_{kn}^T d_{nj}$$

$\mathbf{D}^T \mathbf{D}$ . Левая часть (13) представляет собой  $k$ -ый элемент вектора градиента.

Решение одномерной задачи:

$$h^{(t)} = \arg \min_h \Phi(\mathbf{u}^{(t)} + h\mathbf{p}^{(t)})$$

заключается в выборе величины шага  $h$  вдоль направления спуска  $\mathbf{p}^{(t)}$ . Так как  $\Phi$  не является унимодальной функцией, то для решения этой задачи используем метод оптимального пассивного поиска на множестве «пробных шагов». Это множество включает шаги с такими условными названиями: «шаги, которые обнуляют компоненты невязки решения», «шаги, которые обнуляют компоненты стабилизирующей части функционала» и «шаг по методу Ньютона».

Шаги, которые обнуляют компоненты невязки решения, определяются формулой

$$h_j^{[1]} = - \left( \sum_{n=1}^N a_{jn} u_n^{(t)} - y_j \right) / \left( \sum_{n=1}^N a_{jn} p_n^{(t)} \right); \quad j = 1, \dots, N, \quad (14)$$

т.е. на каждой итерации существует  $N$  таких шагов.

Шаги, которые обнуляют компоненты стабилизирующей части функционала, получаются из условия равенства нулю этих компонент. Рассматривая стабилизирующую часть функционала в (8) как скалярное произведение:  $\|\mathbf{u}'\|^2 = (\mathbf{D}\mathbf{u}, \mathbf{D}\mathbf{u}) = (\mathbf{D}^T \mathbf{D}\mathbf{u}, \mathbf{u})$ , из условия равенства нулю каждой из компонент скалярного произведения  $(\mathbf{D}^T \mathbf{D}(\mathbf{u}^{(t)} + h\mathbf{p}^{(t)}), (\mathbf{u}^{(t)} + h\mathbf{p}^{(t)}))$  получаем два набора из  $N$  шагов:

$$h_n^{[2]} = -u_n^{(t)} / p_n^{(t)}; \quad n = 1, \dots, N, \quad (15)$$

$$h_n^{[3]} = - \sum_{j=1}^N \hat{d}_{nj} u_j^{(t)} / \sum_{j=1}^N \hat{d}_{nj} p_j^{(t)}; \quad n = 1, \dots, N. \quad (16)$$

Шаг, который отвечает методу Ньютона, используется при условии, что решение находится в окрестности некоторого локального минимума. Этот шаг задается формулой

$$h_n^{[4]} = -(\mathbf{g}^{(t)}, \mathbf{p}^{(t)}) / ([\mathbf{A}^T \mathbf{Q} \mathbf{A} + \gamma_2 \mathbf{D}^T \mathbf{D}] \mathbf{p}^{(t)}, \mathbf{p}^{(t)}), \quad (17)$$

где  $\mathbf{Q} = \text{diag} \left( \dots, \frac{(\alpha^2 - [v_j^{(t)}]^2)}{(\alpha^2 + [v_j^{(t)}]^2)^2}, \dots \right)$  есть  $N \times N$  диагональная матрица, зависящая от  $\alpha^2$  и от вектора невязки

$\mathbf{v}^{(t)}$  на  $t$ -ой итерации. Всего существует один такой шаг на  $t$ -ой итерации.

Пробные шаги не являются оптимальными, но они являются квазиоптимальными, так как приводят к уменьшению значения функции  $\Phi$ . Алгоритм поиска на множестве пробных шагов следующий. Сначала по формулам (14)–(17) генерируются значения пробных шагов.

Затем в качестве  $h^{(t)}$  выбирается тот шаг, для которого

$\Phi$  принимает наименьшее значение. Но если шаг  $h^{[4]}$ , который отвечает методу Ньютона, оказывается наилучшим  $N$  раз подряд, то выполняется обновление направления

спуска по формуле:  $\mathbf{p}^{(t)} = -\mathbf{grad}^{(t)}$ . Если число выполненных итераций равно  $N$ , то также выполняется обновление направление спуска. Если ни один из пробных шагов не приводит к уменьшению значения  $\Phi$ , то  $h^{(t)}$  полагается равным нулю и выполняется обновление направление спуска. Если  $h^{(t)}$  равно нулю дважды подряд, то итерационный процесс метода сопряженных градиентов завершается.

Матрицы  $\mathbf{A}$  и  $\mathbf{D}$  имеют следующее строение:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 \dots \\ 1 & 1 & 0 & 0 \dots \\ 1 & 1 & 1 & 0 \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}, \quad \mathbf{D} = \begin{bmatrix} -1 & 1 & 0 & 0 \dots \\ 0 & -1 & 1 & 0 \dots \\ 0 & 0 & -1 & 1 \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}.$$

Матрица  $\mathbf{A}$  имеет размер  $N \times N$ , а матрица  $\mathbf{D}$  имеет размер  $N \times (N-1)$ . Матрицы  $\mathbf{A}^T \mathbf{A}$  и  $\mathbf{D}^T \mathbf{D}$  имеют размер  $N \times N$ . Когда решение  $\mathbf{u}$  получено, то возможно

построить вектор  $\mathbf{f}$  с элементами  $f_i = g(a) + \sum_{k=1}^i u_k$ , ко-

торый аппроксимирует вектор исходных данных  $\mathbf{y}$ .

Решение задачи численного дифференцирования чувствительно к выбору значения  $g(a)$ . Если  $g(a)$  априорно не задано, то часто полагают  $g(a) = y_1$  [6]. Однако такой выбор даст плохой результат для случая, когда первый отсчет  $y_1$  содержит выброс. Для устранения этого недостатка можно решить задачу в «обратном направлении» следования дискретных отсчетов  $y_i$ , предполагая отсутствие выброса в последнем отсчете  $y_N$ . Тогда в полученном решении  $\mathbf{u}$  потребуется обратить направление следования элементов и умножить их на  $-1$ . Но если выбросы есть и в первом, и в последнем отсчетах, то следует сделать два подобных прохода. При этом сначала задача решается в «обратном направлении». В результате выброс в  $y_N$  будет размыт, но выброс в  $y_1$  будет подавлен. Далее получаем вектор  $\mathbf{f}$  (его элементы будут следовать в «обратном направлении»), полагаем  $y_1 = f_N$  и решаем задачу в «прямом направлении». Такой алгоритм условно назовем двухпроходным алгоритмом.

#### 4 ЭКСПЕРИМЕНТЫ

Для демонстрации возможностей предложенного метода были выполнены моделирование задачи численного дифференцирования зашумленных данных с выбросами и обработка экспериментальных данных со спектрами фотолюминесценции (ФЛ). Важная цель моделирования состояла в том, чтобы имитировать шумовую обстановку, характерную для измерений спектра ФЛ, и настроить параметры метода на нее. При этом параметры  $\beta = 0$  и  $q = 2$  были фиксированы, а настройке подлежали только параметры  $\alpha$  и  $\gamma_2$ .

Спектр ФЛ моделировался в диапазоне длин волн 380–720 нм с шагом  $\Delta\lambda = 0,5$  нм, что отвечало  $N = 681$  дискретным отсчетам  $x_i = i \cdot \Delta\lambda$ ;  $i = 1, \dots, N$ , причем  $a = x_1$  и  $b = x_N$ . Модель спектра задавалась суммой пяти гауссовых функций с параметрами:  $A_1 = 3000$ ,  $m_1 = 410\Delta\lambda$ ,  $\sigma_1 = 40\Delta\lambda$ ;  $A_2 = 700$ ,  $m_2 = 310\Delta\lambda$ ,  $\sigma_2 = 40\Delta\lambda$ ;  $A_3 = 250$ ,  $m_3 = 210\Delta\lambda$ ,  $\sigma_3 = 40\Delta\lambda$ ;  $A_4 = 350$ ,  $m_4 = 500\Delta\lambda$ ,  $\sigma_4 = 40\Delta\lambda$ ;  $A_5 = 100$ ,  $m_5 = 580\Delta\lambda$ ,  $\sigma_5 = 40\Delta\lambda$ , где  $A_i$ ,  $m_i$  и  $\sigma_i$  есть амплитуда, местоположение и полуширина  $i$ -ой гауссовой функции. Амплитуды задавались в условных единицах в пределах диапазона значений 16-разрядного АЦП, а привязка местоположения гауссовых функций к шкале длин волн выполнялась по формуле:  $m_i^{[\lambda]} = 380 + m_i - \Delta\lambda$ ;  $i = 1, \dots, N$ . Так, если  $m_1 = 410\Delta\lambda$ , то  $m_1^{[\lambda]} = 584,5$  нм. К сумме гауссовых функций добавлялся постоянный фон амплитуды  $A_0 = 50$  и шум.

При измерениях спектров ФЛ с помощью фотоэлектронного умножителя шум можно считать аддитивным, а его модель задавать суммой шума, который зависит от сигнала, и теплового шума. В результате оценивания значений параметров этой модели на основе экспериментальных данных для нескольких спектров ФЛ, шум был описан соотношением:  $\xi_i = |g(x_i)|^{1/2} \eta_i + 15\zeta_i$ , где  $\eta_i$  и  $\zeta_i$  есть значения гауссовых случайных величин с нулевым математическим ожиданием и единичной дисперсией, а число 15 указывает на величину стандартного отклонения теплового шума (что составляет 4 двоичных разряда из допустимых 16 разрядов). Однако далее шумовая обстановка дополнительно ухудшалась за счет добавления положительных выбросов с вероятностью их появления 0,03 и равномерным распределением их амплитуды в интервале  $[0, 2000]$ .

Непосредственно для обработки экспериментальных данных использовались те зависимости спектров ФЛ, которые помимо широких «гладких» составляющих спектра содержали очень узкие и достаточно большие по амплитуде составляющие линейчатого спектра, которые интерпретировались как выбросы значений. Задача обработки состояла в получении гладких зависимостей как для исходного спектра ФЛ, так и для его первой производной.

#### 5 РЕЗУЛЬТАТЫ

Результаты моделирования задачи численного дифференцирования зашумленных данных с выбросами при имитации шумовой обстановки, характерной для измерений спектра ФЛ, представлены на рис. 1.

На рис. 1а приведена исходная последовательность зашумленных данных с выбросами, а на рис. 1б – результат ее численного дифференцирования обычным методом вычисления правой разностной производной (например, это выполняет функция `diff()` среды Matlab). Видно, что результат такого дифференцирования является неудовлетворительным.

На рис. 1в и рис. 1г приведен результат применения метода регуляризации Тихонова (9) для  $k = 1$  с оптимальным значением параметра регуляризации, где пунктирная линия отвечает истинной зависимости, а сплошная линия отвечает полученной зависимости. Видно, что выбросы исказили искомое решение (рис. 1г), вызывая неверное поведение хвостов полученной зависимости. При этом относительная ошибка решения и относительная ошибка аппроксимации составили 11,2% и 4,1%, соответственно. Набор статистики по 100 реализациям шума дал усредненные значения этих показателей величиной в 12,7% и 4,2%, соответственно.

На рис. 1д и рис. 1е приведены результаты применения предложенного метода после использования двухпроходного алгоритма (242 итерации на первом проходе и 188 итераций на втором проходе) для  $\alpha^2 = 10000$  и  $\gamma_2 = 10000$ , где пунктирная линия отвечает истинной зависимости, а сплошная линия отвечает полученной зависимости. Видно, что на обоих графиках сплошная линия практически полностью накрыла пунктирную линию. При этом относительная ошибка решения и относительная ошибка аппроксимации составили 4,9% и 1,2%, соответственно. Набор статистики по 100 реализациям шума дал усредненные значения этих показателей величиной в 5,7% и 1,4%, соответственно.

На рис. 2 приведено три экспериментальных зависимости спектров ФЛ кристаллов ZnS:Mn, а также результаты их сглаживания и численного дифференцирования предложенным методом для  $\alpha^2 = 10000$  и  $\gamma_2 = 10000$ . При этом использовался двухпроходный алгоритм, причем для обработки первой экспериментальной зависимости потребовалась 251 итерация на первом проходе и 293 итерации на втором проходе, для обработки второй экспериментальной зависимости – соответственно 133 и 127 итераций, а для третьей экспериментальной зависимости – 254 и 227 итераций, соответственно.

Видно, что выбросы на экспериментальных кривых расположены не случайно, а их ширина у основания составляет примерно 2,5 нм. По этой причине их следует рассматривать как детерминированные линейчатые составляющие спектра ФЛ, получая последние путем вычитания сглаженной зависимости из исходной зависимости данных. Также следует отметить, что применение метода регуляризации Тихонова превращало такие выбросы на экспериментальных кривых в достаточно широкие «холмы», что приводило к появлению нефизических осцилляций в результатах численного дифференцирования.

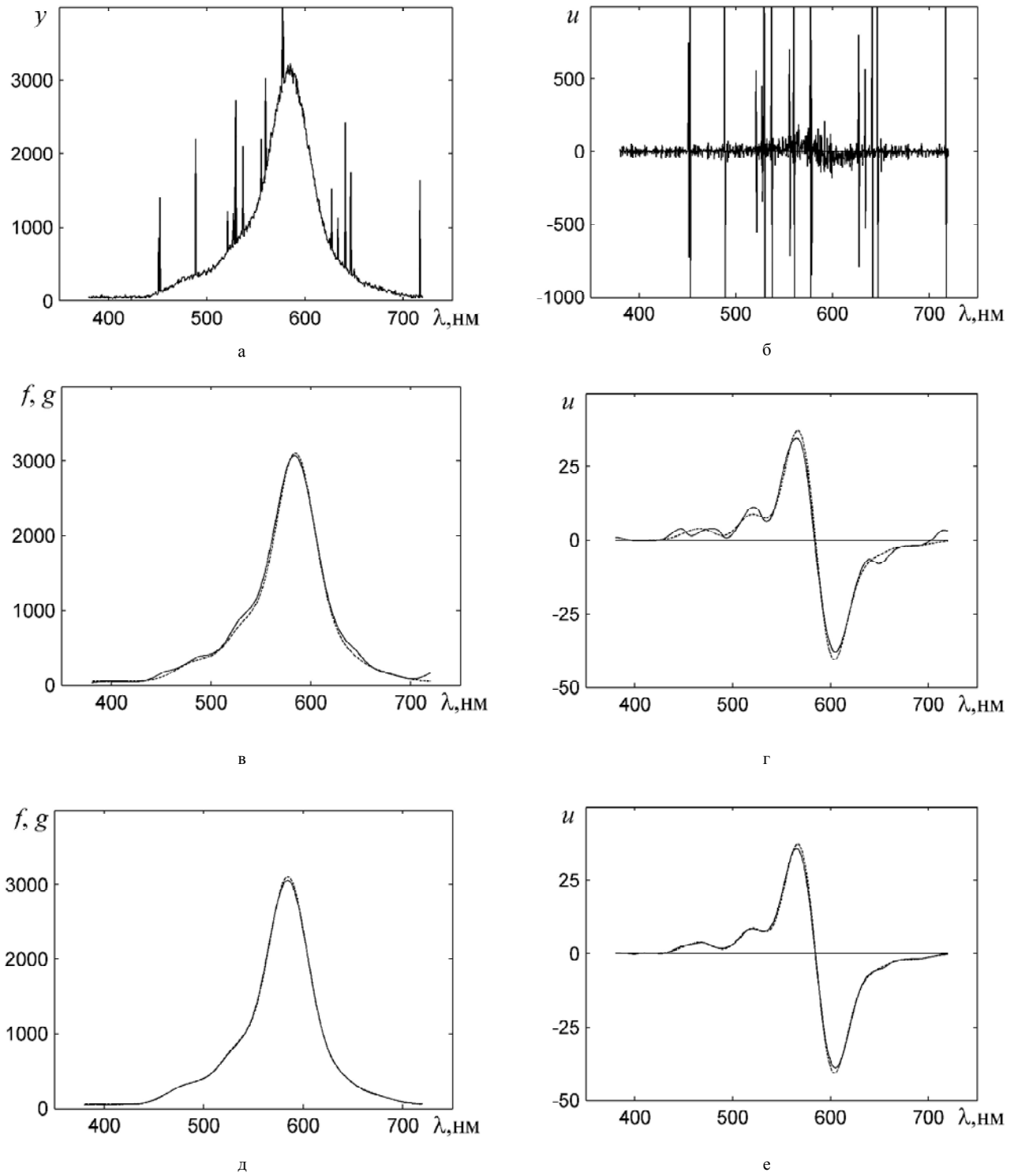


Рисунок 1 – Результаты моделирования:

а, б – данные и их численное дифференцирование обычным методом; в, г – сглаживание данных и численное дифференцирование методом Тихонова; д, е – сглаживание и численное дифференцирование предложенным методом

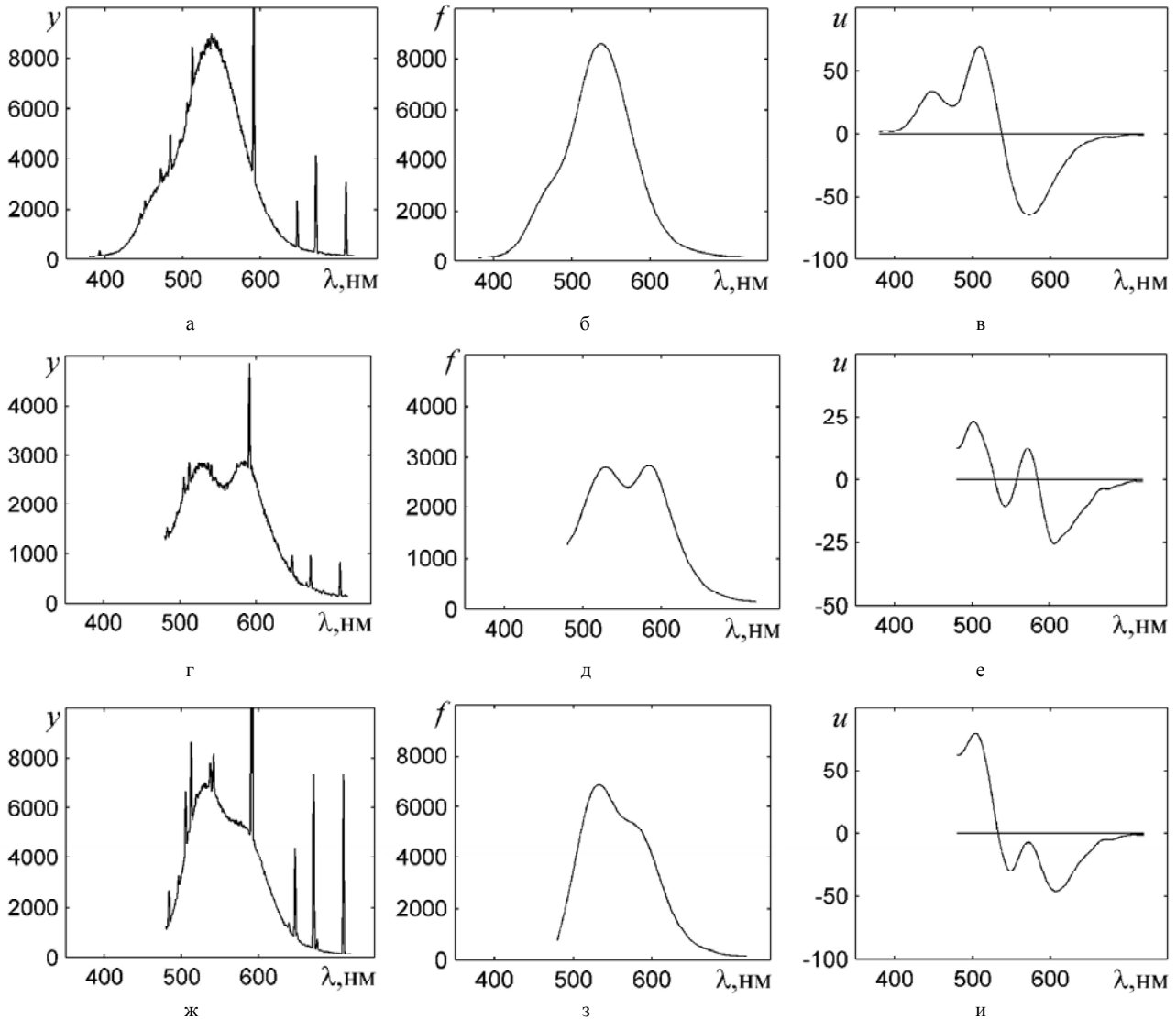


Рисунок 2 – Обработка экспериментальных данных:  
а, г, ж – исходные данные; б, д, з – сглаженные данные; в, е, и – первая производная

## 6 ОБСУЖДЕНИЕ

Результаты моделирования и обработки экспериментальных данных демонстрируют эффективность предложенного метода. Это во многом определяется хорошим качеством настройки его свободных параметров  $\alpha$ ,  $\beta$ ,  $q$  и параметра регуляризации  $\gamma_2$ . Поскольку для заданной шумовой обстановки была типична некоторая шумовая полоса с выбросами значений, то для грубой настройки на эту шумовую обстановку в качестве прототипа был выбран шум импульсного типа с законом распределения Коши [11]. Для этого шума  $\beta = 0$ ,  $q = 2$ , а  $\alpha$  должно настраиваться на полуширину шумовой полосы (т.е. на масштаб шума Коши) [10]. Вторым настраиваемым параметром был параметр регуляризации  $\gamma_2$ , который имеет смысл отношения величины протяженности невязки решения к величине энергии производной решения. На практике настройка параметров  $\alpha$  и  $\gamma_2$  выполнялась на этапе численного моделирования типовых за-

дач данного класса путем минимизации абсолютной ошибки решения. Затем полученные значения этих параметров применялись к другим задачам этого класса, в том числе, к задачам с экспериментальными данными. Хотя более точная настройка  $\beta$  и  $q$  позволяет улучшить результаты [10], однако здесь она не выполнялась.

При решении задачи одномерной оптимизации наилучшими шагами более часто оказывались пробные шаги, которые обнуляли компоненты стабилизирующей части функционала, а менее часто – шаги, которые обнуляли невязку решения. Шаги, которые отвечали методу Ньютона, при вычислениях практически не использовались, что указывает на их низкую эффективность для решения подобных задач.

## ВЫВОДЫ

Предложенный метод может быть использован для численного дифференцирования зашумленных данных с выбросами. При этом он позволяет получить гладкую

аппроксимацию первой производной от исходных данных и, соответственно, гладкую аппроксимацию самих исходных данных.

Результаты моделирования и обработки экспериментальных данных, которые представляли собой спектры фотолюминесценции при наличии в их составе узких линейчатых спектральных составляющих, подтвердили эффективность предложенного метода.

Метод можно обобщить на случай негладкого решения путем построения стабилизирующей части функционала на основе ограничения полной вариации решения.

### БЛАГОДАРНОСТИ

Работа выполнена в рамках госбюджетной научно-исследовательской темы Днепропетровского национального университета «Методы и информационные технологии цифровой обработки многоканальных данных» (номер государственной регистрации 0116U001297). Автор выражает благодарность проф. Коваленко А. В. и инж. Плахтию Е. Г. за предоставленные экспериментальные данные и полезное обсуждение результатов работы.

### СПИСОК ЛИТЕРАТУРЫ

1. Derivative spectroscopy and the continuous relaxation spectrum / [A. R. Davies, R. S. Anderssen, F. R. de Hoog, N. J. Goulding] / *Journal of Non-Newtonian Fluid Mechanics*. – 2016. – Vol. 233, July. – P. 107–118. DOI: 10.1016/j.jnnfm.2016.01.018.
2. Autonomous Gaussian decomposition / [R. R. Lindner, C. Vera-Ciro, C. E. Murray and other] // *The Astronomical Journal*. – 2015. – Vol. 149, No. 4. – P. 138–149. DOI: 10.1088/0004-6256/149/4/138.
3. Maitra I. K. Adaptive Edge Detection Method towards Features Extraction from Diverse Medical Imaging Technologies / I. K. Maitra, S. K. Bandhyopadhyay // *Intelligent Multidimensional Data Clustering and Analysis*. – 2016. – Chapter 7. – P. 159–192. DOI: 10.4018/978-1-5225-1776-4
4. Selected algorithms for measurement data processing in impulse-radar-based system for monitoring of human movements /

Вовк С. М.

Канд. фіз.-мат. наук, доцент, доцент кафедри автоматизованих систем обробки інформації Дніпропетровського національного університету, Дніпро, Україна

### МЕТОД ЧИСЕЛЬНОГО ДИФЕРЕНЦІЮВАННЯ ЗАШУМЛЕНИХ ДАНИХ З ВИКИДАМИ

**Актуальність.** Застосування традиційних методів чисельного диференціювання до зашумлених даних з викидами призводить до значних похибок. Об'єктом цього дослідження є процес чисельного диференціювання таких даних.

**Мета роботи** – розробка методу чисельного диференціювання зашумлених даних з викидами, який дозволяє отримати гладку аппроксимацию їх першої похідної та, відповідно, гладку аппроксимацию самих даних.

**Метод.** Запропонований метод чисельного диференціювання заснований на рішенні задачі мінімізації згладжувального функціонала, який побудований на критерії мінімуму протяжності відхилення та обмеженні енергії першої похідної рішення. Критерій мінімуму протяжності формує основну частину функціонала й забезпечує його стійку поведінку у відношенні до адитивного шуму й викидів. Обмеження формує стабілізуювальну частину функціонала й забезпечує гладкість рішення задачі. Внесок зазначених частин регулюється за допомогою параметра регуляризації. Оскільки основна частина згладжувального функціонала не є опуклою, то задача його мінімізації є задачею нелінійного неопуклого програмування. Для чисельного рішення цієї задачі використовується метод спряжених градієнтів, в якому величина кроку вздовж напрямку спуска визначається на множині випробувальних кроків. Ці кроки мінімізують окремі компоненти основної та стабілізуювальної частин функціонала, що дозволяє переходити з одного локального мінімуму функціонала в інший більш глибокий локальний мінімум.

**Результати.** Моделювання задачі чисельного диференціювання зашумлених даних з викидами та обробка експериментальних даних, які являли собою спектри фотолюмінисценції з присутністю в їх складі вузьких лінійчатих спектральних складових, засвідчили ефективність запропонованого методу.

**Висновки.** Запропонований метод може бути використаний для чисельного диференціювання зашумлених зашумлених даних з викидами. При цьому він дозволяє отримати гладку аппроксимацию першої похідної початкових даних, а також гладку аппроксимацию самих початкових даних. Поданий метод можна узагальнити на випадок негладкого рішення шляхом побудови стабілізуювальної частини функціонала на основі обмеження повної варіації рішення.

**Ключові слова:** чисельне диференціювання, гладка аппроксимация, критерій мінімуму протяжності.

- [A. Mikkina, J. Wagner, P. Mazurek, R. Z. Morawski] // *Journal of Physics*. – 2016. – Vol. Conference Series 772. – P. 1–6. DOI:10.1088/1742-6596/772/1/012057.
5. Knowles I. Methods for numerical differentiation of noisy data / I. Knowles, R. J. Renka // *Electronic Journal of Differential Equations*. – 2014. – Vol. Conference 21. – P. 235–246.
  6. Chartrand R. Numerical differentiation of noisy, nonsmooth data / R. Chartrand // *ISRN Applied Mathematics*. – 2011. – Vol. 2011. – P. 1–11. DOI: 10.5402/2011/164564.
  7. Liao H. Outlier Impact and Accommodation Methods: Multiple Comparisons of Type I Error Rates / H. Liao, Y. Li, G. Brooks // *Journal of Modern Applied Statistical Methods*. – 2016. – Vol. 15, May. – No. 1. – P. 452–471.
  8. Borulko V. F. Principle of minimum extent in spatial spectrum extrapolation problems of complex-valued sources / V. F. Borulko, S. M. Vovk // *Telecommunications and Radio Engineering*. – 2013. – Vol. 72, No. 7. – P. 581–592. DOI: 10.1615/TelecomRadEng.v72.i7.30.
  9. Вовк С. М. Двойственный метод минимума пространственной протяженности для робастного оценивания параметров дипольных источников излучения / С. М. Вовк, В. Ф. Борулько // *Радіоелектроніка, інформатика, управління*. – 2014. – № 2. – С. 8–17. DOI: 10.15588/1607-3274-2014-2-1.
  10. Vovk S. M. General approach to building the methods of filtering based on the minimum duration principle / S. M. Vovk // *Radioelectronics and Communications Systems*. – 2016. – Vol. 59. – № 7. – P. 281–292. DOI: 10.3103/S0735272716070013.
  11. Gonzalez J. G. Optimality of the myriad filter in practical impulsive-noise environments / J. G. Gonzalez, G. R. Arce // *IEEE Trans. on Signal Processing*. – 2001. – Vol. 49, No. 2. – P. 438–441. DOI: 10.1109/78.902126.
  12. Aysal T. C. Meridian filtering for robust signal processing / T. C. Aysal, K. E. Barner // *IEEE Trans. on Signal Processing*. – 2007. – Vol. 55, No. 8. – P. 3949–3962. DOI: 10.1109/TSP.2007.894383.

Статья поступила в редакцию 01.02.2017.  
После доработки 22.04.2017.



Vovk S. M.

PhD, Associate professor, Associate professor of Department of Automated Systems of Information Processing, Dnipropetrovs'k National University, Dnipro, Ukraine

#### METHOD FOR NUMERICAL DIFFERENTIATION OF NOISY DATA WITH OUTLIERS

**Context.** Using of conventional methods of numerical differentiation to the noisy data with outliers leads to significant errors. The object of this study is the process of numerical differentiation of such data.

**Objective.** The goal of this work is the development of a method of numerical differentiation of the noisy data with outliers to obtain a smooth approximation of the first derivative of original data as well as a smooth approximation of the original data themselves.

**Method.** The proposed method of numerical differentiation is based on solving the problem of minimizing the smoothing functional, which is built on the criteria of a minimum of extent of the solution residual and of an energy constraint of the first derivative of solution. The minimum-extent criterion defines the main part of functional and ensures its stable behavior with respect to the additive noise and outliers. The energy constraint defines the stabilizing part of the functional and provides a smooth solution of the problem. The contribution of these parts is controlled by a regularization parameter. Since the main part of smoothing functional is not convex, then the minimization problem is the non-convex nonlinear programming problem. For the numerical solution of this problem the conjugate gradient method is used. In this method the step size along the descent direction is defined on the set of test steps. These steps minimize the individual components of the main and stabilizing parts of the smoothing functional that allows to move from the one local minimum of the functional to another deeper local minimum.

**Results.** Simulation of the problem of numerical differentiation of noisy data with outliers and processing of the experimental data, which are photoluminescence spectra with narrow line components in their compositions, confirmed the performance of the proposed method.

**Conclusions.** The proposed method can be used for numerical differentiation of noisy data with outliers. It provides a smooth approximation of the first derivative of the original data, as well as a smooth approximation of the original data themselves. This method can be generalized to the case of non-smooth solutions by constructing a stabilizing part of the functional based on the criterion of minimum total variation.

**Keywords:** numerical differentiation, smooth approximation, minimum-extent criteria.

#### REFERENCES

1. Davies A. R., Anderssen R. S., de Hoog F. R., Goulding N. J. Derivative spectroscopy and the continuous relaxation spectrum, *Journal of Non-Newtonian Fluid Mechanics*, 2016, Vol. 233, pp. 107–118. DOI: 10.1016/j.jnnfm.2016.01.018.
2. Lindner R. R., Vera-Ciro C., Murray C. E., Stanimirovic S., Babler B. L., Heiles C., Henebelle P., Goss W. M., Dickey J. Autonomous Gaussian decomposition, *The Astronomical Journal*, 2015, Vol. 149, No. 4, pp. 138–149. DOI: 10.1088/0004-6256/149/4/138.
3. Maitra I. K., Bandhyopadhyay S. K. Adaptive Edge Detection Method towards Features Extraction from Diverse Medical Imaging Technologies, *Intelligent Multidimensional Data Clustering and Analysis*, 2016, Chapter 7, pp. 159–192. DOI: 10.4018/978-1-5225-1776-4
4. Mikkina A., Wagner J., Mazurek P., Morawski R. Z. Selected algorithms for measurement data processing in impulse-radar-based system for monitoring of human movements, *Journal of Physics*, 2016, Vol. Conference Series 772, pp. 1–6. DOI:10.1088/1742-6596/772/1/012057.
5. Knowles I., Renka R. J. Methods for numerical differentiation of noisy data, *Electronic Journal of Differential Equations*, 2014, Vol. Conference 21, pp. 235–246.
6. Chartrand R. Numerical differentiation of noisy, nonsmooth data, *ISRN Applied Mathematics*, 2011, Vol. 2011, pp. 1–11. DOI: 10.5402/2011/164564.
7. Liao H., Li Y., Brooks G. Outlier Impact and Accommodation Methods: Multiple Comparisons of Type I Error Rates, *Journal of Modern Applied Statistical Methods*, 2016, Vol. 15, No. 1, pp. 452–471.
8. Borulko V. F., Vovk S. M. Principle of minimum extent in spatial spectrum extrapolation problems of complex-valued sources, *Telecommunications and Radio Engineering*, 2013, Vol. 72, No. 7, pp. 581–592. DOI: 10.1615/TelecomRadEng.v72.i7.30.
9. Vovk S. M., Borulko V. F. Dvoystvennyy metod minimuma prostranstvennoy protyazhennosti dlya robastnogo otsenivaniya parametrov dipolnyih istochnikov izlucheniya, *Radio Electronics, Computer Science, Control*, 2014, No. 2, pp. 8–17. DOI: 10.15588/1607-3274-2014-2-1.
10. Vovk S. M. General approach to building the methods of filtering based on the minimum duration principle, *Radioelectronics and Communications Systems*, 2016, Vol. 59, No. 7, pp. 281–292. DOI: 10.3103/S0735272716070013.
11. Gonzalez J. G., Arce G. R. Optimality of the myriad filter in practical impulsive-noise environments, *IEEE Trans. on Signal Processing*, 2001, Vol. 49, No. 2, pp. 438–441. DOI: 10.1109/78.902126.
12. Aysal T. C., Barner K. E. Meridian filtering for robust signal processing, *IEEE Trans. on Signal Processing*, 2007, Vol. 55, No. 8, pp. 3949–3962. DOI: 10.1109/TSP.2007.894383.