УДК 519.213

© М.Н. Жуков, А.М. Тішаєва, І.В. Тішаєв, 2011

Київський національний університет імені Тараса Шевченка, м. Київ

ОЦІНКА ЩІЛЬНОСТІ РОЗПОДІЛУ СПЕКТРАЛЬНОЇ ЯСКРАВОСТІ НА ОСНОВІ КОМПОЗИЦІЙНОЇ МОДЕЛІ В ЗАДАЧАХ ДИСТАНЦІЙНИХ ЗОНДУВАНЬ ЗЕМЛІ

Запропоновано новий метод оцінки щільностей одно- та багатовимірних розподілів для використання в задачах дешифрування даних дистанційного зондування Землі на основі моделі суміші локальних розподілів. Показано високу ефективність, особливо в разі ускладнених неоднорідних розподілів, прийнятність для опису багатовимірних розподілів, з компонентами, корельованими у загально прийнятому розумінні. Наведено результати практичного застосування.

Ключові слова: щільність розподілу, композиційна модель, спектральні канали, дистанційні зондування.

Вступ. Оцінка функції щільності розподілу належить до поширених операцій в статистичній обробці геологічних даних. Значною мірою від якості апроксимації на початковому етапі аналізу експериментальних даних залежить результат розв'язання задачі. Істотні перспективи може дати зображення щільності багатовимірного розподілу у вигляді суперпозиції простіших розподілів.

Термін "суміш" стосовно задачі апроксимації розподілу випадкової величини за вибіркою експериментальних даних має достатньо підстав для використання в обробці геологічних даних. Суміш являє собою суму розподілів випадкових величин, з яких складається вибірка. Поширеним є підхід, відомий в іноземній літературі, як Gaussian mixture – суміш гауссівських розподілів [1, 2]. Ідея полягає у побудові загальної щільності розподілу вибірки у вигляді зваженої суми нормальних розподілів, якими з певним ступенем достовірності описуються складові частини вибірки:

$$f(\mathbf{x}) = \sum_i k_i \mathrm{N}\left(\mathbf{x} \middle| \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right). \qquad (1)$$

Тут

$$\mathrm{N}\left(\mathbf{x}\middle|\,\boldsymbol{\mu},\boldsymbol{\Sigma}\right)=\frac{1}{\left(2\pi\right)^{n/2}\left|\boldsymbol{\Sigma}\right|^{1/2}}\exp\left[-\tfrac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)^{T}\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)\right]- \qquad (2)$$

багатовимірний нормальний розподіл (вимірності n);

$$\sum_i k_i=1, \qquad (3)$$

де k_i – вагові коефіцієнти.

Моделювання. Розглянемо для прикладу одновимірну випадкову величину X із щільністю розподілу $f(x)$. Згенеруємо достатньо велику кількість s реалізацій випадкової величини X, щільність розподілу якої є сумішшю нормальних розподілів із параметрами $\mu_1 = 125$, $\sigma_1 = 0{,}5$ і $\mu_2 = 130$, $\sigma_2 = 1{,}5$ у частках, відповідно, $k_1 = k_2 = 0{,}5$ [3]. Уявімо, що ця сукупність реалізацій отримана внаслідок проведення деяких спостережень, тобто є набором емпіричних даних – вибіркою з усієї множини можливих значень X. Потім припустимо, як це зазвичай буває на практиці, що дані отримані з параметричної сім'ї нормальних розподілів (2), зокрема:

$$\hat{f}\left(x\right)=\frac{1}{\sqrt{2\pi\hat{\sigma}^2}}\exp\left[-\frac{1}{2}\frac{\left(x-\hat{\mu}\right)^2}{\hat{\sigma}^2}\right].$$

Визначивши оцінки [4] для цієї моделі, отримаємо: $\hat{\mu}=127{,}56$, $\hat{\sigma}=2{,}71$. Істинний розподіл і оцінена модель зображені на рис. 1.

Протестувавши на адекватність розраховану модель реальним даним, можна переконатися, що нормальна модель впевнено відторгаєть-

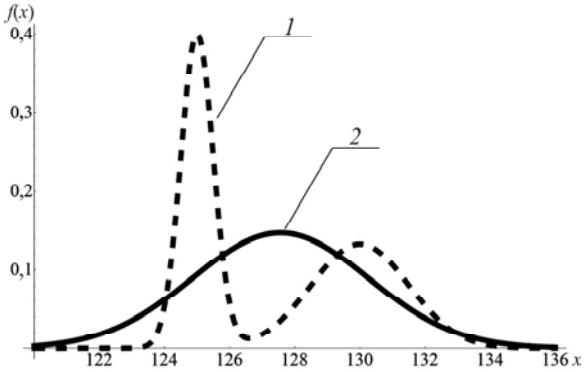


Рис. 1. Розподіл випадкової величини X: *1* – змодельований як гауссівська суміш; *2* – його апроксимація на основі параметричної моделі (2)

ся даними. В цій ситуації існує два способи: або шукати придатнішу модель, або відмовитися від параметричних сімей розподілів на користь непараметричних оцінок функції щільності.

Перший варіант заздалегідь є проблематичним, оскільки пересічно відсутнє теоретичне обґрунтування моделі, під яку підпадатимуть експериментальні дані. Тому не викликає сумніву актуальність дослідження сфери застосування непараметричних методів реконструкції функції щільності розподілу.

Пропонується використання непараметричного методу, умовно названого "композиційним" [5]. Його суть полягає у використанні сімейства елементарних нормальних розподілів X_i (1) для елементів вибірки. Надалі називатимемо такі розподіли локальними. В геологічний практиці, залежно від характеру дослідження, локальному розподілу відповідають розподіли: вмісту хімічного елемента (геохімічні методи пошуків корисних копалин, екогеохімія), параметрів фізичного поля (петрофізика, спостереження фізичних полів), кількісні показники ґрунтів (інженерна геологія) та підземних вод (гідрогеологія), яскравості елемента зображення земної поверхні (дистанційні зондування Землі).

Визначимо параметри цих розподілів так:

$$\mu_i = \mathrm{M}X_i, \quad \sigma_i = g(\mu_i), \tag{4}$$

де $\mathrm{M}X_i$ – математичне сподівання величини X_i; $g(\mu_i)$ – функція, що описує залежність локального середнього квадратичного відхилення (СКВ) від математичного сподівання. За результатами експериментів, така модель є набагато гнучкішою навіть за використання лінійного наближення до функції $g(\mu)$. Композиційний розподіл, як наближення істинної функції щільності розподілу випадкової величини, визначиться як зважена сума елементарних нормальних розподілів:

$$\hat{f}(x) = \frac{1}{s}\sum_{i=1}^{s} \mathrm{N}\left(x \middle| \mu_i, \sigma_i\right), \tag{5}$$

де s – кількість елементів вибірки.

Такий підхід обходить традиційні труднощі параметричного оцінювання і є працездатним в умовах браку інформації та складного вигляду розподілів спостережених даних. У разі застосування аналогічної багатовимірної моделі з'являється така важлива перевага, як автоматичне врахування статистичних зв'язків, причому навіть у нелінійній формі, тоді як за використання класичного підходу, наприклад, багатовимірного нор-

мального розподілу, виникають труднощі навіть з лінійною формою через необхідність обертання кореляційної матриці та, відповідно, наджорсткі вимоги щодо репрезентативності вибірок.

У наведеному вище прикладі змодельована гауссівська суміш двох рівноймовірних розподілів $f(x)$ визначена у вигляді

$$f(x) = \frac{1}{2}\left[\mathrm{N}\left(x \middle| 125, 0.5\right) + \mathrm{N}\left(x \middle| 130, 1.5\right)\right]. \qquad (6)$$

Побудована для того самого прикладу оцінка функції щільності розподілу на основі запропонованого композиційного розподілу (вирази (1) і (4)) матиме такий аналітичний вигляд:

$$\hat{f}(x) = \frac{1}{s}\sum_{i=1}^{s} \mathrm{N}\left(x \middle| X_i, g(X_i)\right). \qquad (7)$$

Виходячи з цілком правдоподібного припущення, принаймні для геологічних застосувань, лінійності функції $g(\mu)$ та умови $\sigma = 0$ при $\mu = 0$, одержимо $g(\mu) = a\mu$.

За вибором з ймовірностями $p_1 = p_2 = 0{,}5$ однієї із складових X_i суміші побудовані графіки щільності змодельованого вище розподілу (6) і його оцінок (7) з використанням $g(X_i) = aX_i$ з подальшою генерацією нормального розподілу цієї складової з $\mu_i = X_i$, $\sigma_i = a\mu_i$, при $i = 1, \ldots, s$ (рис. 2). Після одержання s таких локальних розподілів обчислена щільність композиційного розподілу за (7).

Як видно з рис. 2, оцінки функції щільності композиційного розподілу правдоподібніші, ніж оцінка, обчислена на основі параметричної моделі.

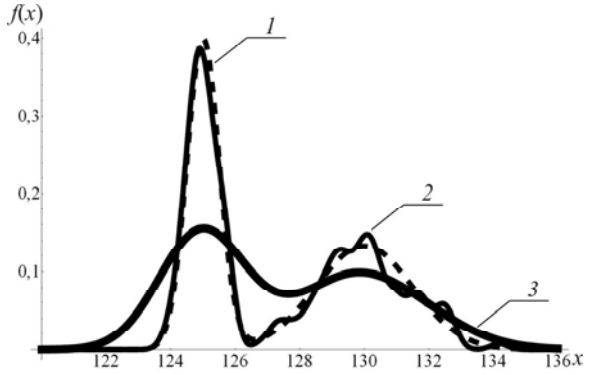


Рис. 2. Істинний розподіл випадкової величини X (*1*) та змодельовані розподіли на основі композиційної моделі при $g(\mu) = a\mu$, $a = 0{,}002$ (*2*) та $a = 0{,}01$ (*3*)

Якщо вважати, що кожен елемент вибірки отриманий як найімовірніше значення локального нормального розподілу, то стає цілком очевидним вибір самого елементу вибірки як оцінки математичного сподівання цього розподілу. Для обґрунтованого вибору відповідної оцінки СКВ пропонується попередньо встановлювати статистичну залежність між СКВ і математичним сподіванням за результатами спеціально влаштованого експерименту.

Практична реалізація. Проілюструємо запропоновані рішення на прикладі даних дистанційних зондувань. Дані космічної зйомки переважно подаються у вигляді матриці цілих невід'ємних чисел, кожен елемент якої (піксел) характеризується значенням інтенсивності відбитої або випроміненої енергії [6]. Кожен піксел (у подальшому – об'єкт) характеризується вектором ознак $\mathbf{x}_{ij} = \{x_1, x_2, \ldots, x_n\}_{ij}$ (рис. 3). Компонентами вектора $\mathbf{x}$ є значення інтенсивності відбитої або випроміненої енергії, зареєстрованої в n спектральних каналах, в яких проводиться зйомка; i, j – умовні координати пікселу в межах знімка (фактично, це номери рядка та стовпчика матриці значень, на перетині яких знаходиться піксел).

Діапазон можливих значень компонент вектора x визначається радіометричною роздільною здатністю системи реєстрації і знаходиться в межах $[0,2^b - 1]$, де b – розрядність АЦП каналу реєстрації. Тому достатньо розглянути діапазон $|\mathbf{x}| \in \left[0, (2^b - 1)\sqrt{n}\right]$.

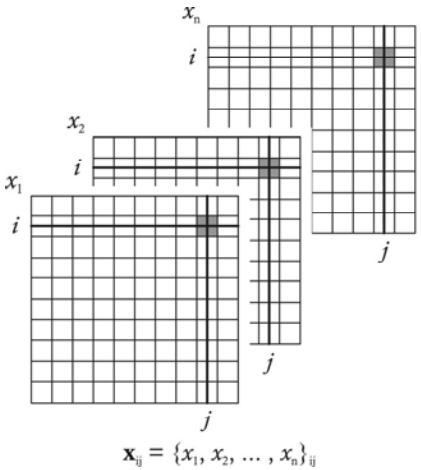
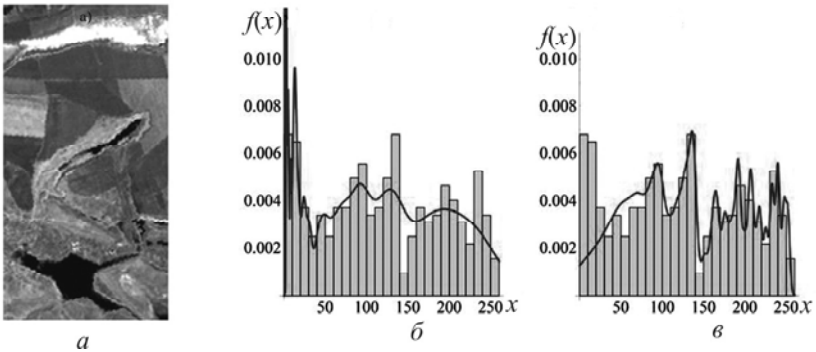


Рис. 3. Формування набору даних на основі результатів дистанційної мультиспектральної зйомки

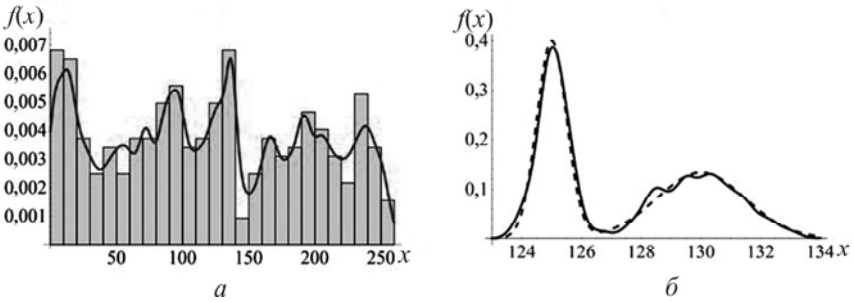
Припустимо, що між СКВ і математичним сподіванням існує зв'язок у вигляді $\sigma = g(\mu) = a\mu$. У випадку прямої залежності в області малих значень компонент вектора **x** можливі нерегулярні високоамплітудні викиди значень функції щільності розподілу, внаслідок того що $\sigma \sim \mu \to 0$ і одночасно $\int_{-\infty}^{+\infty} f(x) = 1$. Таким чином, в області малих значень випадкової величини функція щільності розподілу може містити велику кількість голкоподібних викидів. Натомість в області великих значень, внаслідок тих самих властивостей: $\sigma \sim \mu \to \left(2^b - 1\right)\sqrt{n}$ і водночас, $\int_{-\infty}^{+\infty} f(x) = 1$, може відбуватись надмірне згладжування функції щільності розподілу випадкової величини (рис. 4, *б*). У випадку оберненої залежності ситуація змінюється на протилежну (рис. 4, *в*).

Для встановлення регресії $\sigma = g(\mu)$ можна виходити з таких міркувань. Певна річ, чим більшою є кількість повторів $C(X)$ у вибірці даних певного значення випадкової величини X, тим більшою є щільність імовірності для цього значення. Відповідно, тим "вужчою" має бути побудована на цьому значенні функція щільності локального нормального розподілу $N(X, \sigma)$, а отже, тим меншим має бути значення середнього квадратичного відхилу σ. Таким чином, взявши за основу обернений зв'язок $\sigma_i \sim 1/C(X_i)$, можна побудувати критерій для визначення СКВ для кожного значення випадкової величини індивідуально. Зокрема для даних, зображених на рис. 4, використана евристична залежність вигляду $\sigma_i = 10/\sqrt{C(X_i)}$ (рис. 5).



а *б* *в*

Рис. 4. Фрагмент 8-бітного цифрового зображення: *а* – космічний знімок частини території Керченського півострова; *б*, *в* – оцінки функції щільності розподілу яскравості зображення



а *б*

Рис. 5. Результат апроксимації функції щільності розподілу емпіричних даних із використанням залежності вигляду $\sigma_i \sim 1/C(X_i)$: *а* – апроксимація даних дистанційних зондувань, наведених на рис. 4; *б* – апроксимація (суцільна лінія) даних модельного прикладу, наведеного на рис. 1

Дослідивши апроксимацію, показану на рис. 5, можна констатувати в цілому задовільну відповідність побудованих залежностей вхідним даним. Універсальніший підхід передбачає знаходження параметрів функціональної залежності $\sigma = Y(C(x), a_1, \ldots, a_t)$ методом найменших квадратів після проведення спеціального експерименту.

Переваги запропонованого методу стають набагато значнішими для оцінки щільності багатовимірного розподілу. Як відомо, за спроби використання моделі багатовимірного нормального розподілу (1) у вигляді

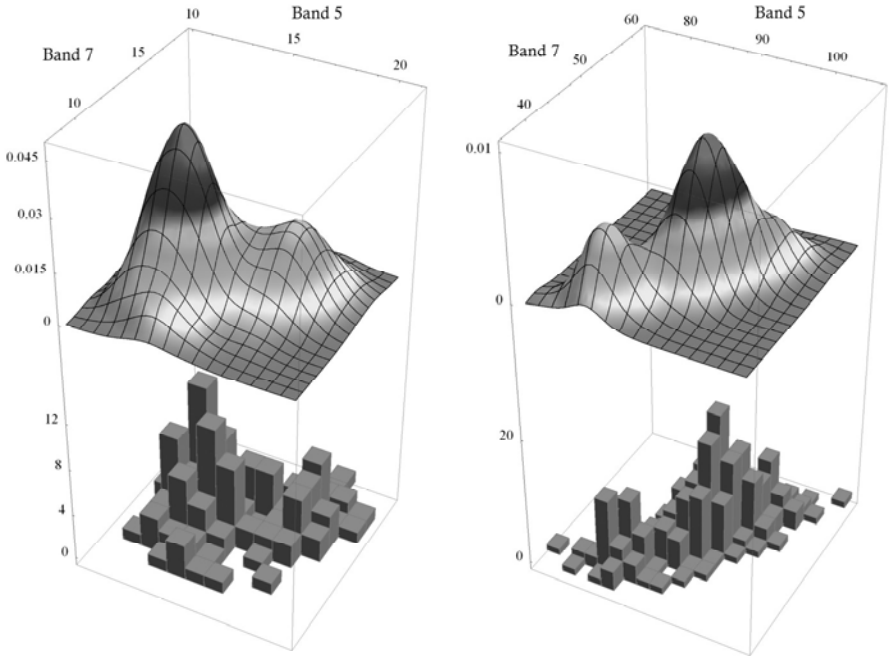
$$\hat{f}(\mathbf{x}) = \frac{1}{s}\sum_{i=1}^{s} \mathrm{N}\left(\mathbf{x} \middle| \mathbf{x}_i, \mathbf{\Sigma}_i\right) \qquad (8)$$

виникнуть серйозні обчислювальні проблеми ще на етапі побудови багатовимірних розподілів. Зокрема, отримання стійких оцінок коваріаційної матриці $\mathbf{\Sigma}$ потребує виконання умови $s >> n$ (s – кількість елементів вибірки; n – вимірність), що, як правило, для навчальних вибірок не виконується. Композиційна модель обходить ці труднощі: кожна компонента x_j ($j = \overline{1,n}$) вектора $\mathbf{x}_i$ ($i = \overline{1,s}$) визначається як випадкова величина деякого нормального розподілу $\mathrm{N}(x, \sigma)$; фактично це означає, що усі компоненти x_j ($j = \overline{1,n}$) вектора $\mathbf{x}_i$ у розумінні локального розподілу є незалежними. Такий підхід дає змогу використовувати композиційну модель для оцінки функцій розподілу багатовимірних випадкових величин без залучення процедур обертання кореляційних матриць. Отже, вираз (8) може бути замінений на вираз

$$\hat{f}(\mathbf{x}) = \frac{1}{s}\sum_{i=1}^{s}\prod_{j=1}^{n} \mathrm{N}\left(\mathbf{x}_i \middle| x_{ij}, \sigma_{ij}\right). \qquad (9)$$

Визначення оцінки (9) може бути ефективно реалізоване як за достатньо високої вимірності n, так і за умови малої чисельності s вибірки. В цьому – найцінніша перевага запропонованої багатовимірної моделі.

Продемонструємо суть викладеного методу на прикладі розв'язання задачі дешифрування даних дистанційного зондування Землі. Для практичної реалізації використаємо мультиспектральний космічний знімок, зроблений 21 серпня 2000 р. спектрорадіометром "ETM+" із супутника "Landsat-7": band 1 (0,45–0,52 мкм), band 2 (0,52–0,60 мкм), band 3 (0,63–0,69 мкм), band 4 (0,76–0,9 мкм), band 5 (1,55–1,75 мкм), band 7 (2,08–2,35 мкм), band 61 (10,4–12,5 мкм), band 62 (10,.4–12,5 мкм). Принцип формування вектора **x** описаний вище і зображений на рис. 3. На космознімку виділена ділянка, в межах якої незалежними методами проведене тематичне районування на різні класи. На рис. 6 зображені двовимірні гістограми і функції щільності розподілів вектора **x** = {band 5, band 7} для двох класів, отримані згідно з (9).



а *б*

Рис. 6. Приклади оцінки двовимірної щільності розподілу на основі моделі композиційного розподілу

Зазначений підхід є перспективним для оцінювання щільностей розподілу, умовних за класом, у задачах класифікації на основі критерію Байєса. Питання, пов'язані із застосуванням даних дистанційних зондувань у класифікаційних задачах, зокрема, в задачі локалізації зон підтоплення за даними мультиспектральної космічної зйомки, розглянуті у статті [7].

Висновки. Запропонована модель композиційного розподілу дає можливість знаходити ефективні оцінки складних багатомодальних розподілів, у тому числі багатовимірних. Така ситуація є типовою для даних дистанційних зондувань Землі, коли вимірність простору ознак (кількість спектральних каналів, у яких ведеться зйомка) може сягати кількох сотень. Викладений вище непараметричний підхід довів свою спроможність, і можна сподіватися на результати в задачах тематичної класифікації певної ділянки досліджень на основі супутникової зйомки.

1. *Gaussian* mixture density modeling, decomposition, and applications / Xinhua Zhuang, Yan Huang, Palaniappan K.; Yunxin Zhao // Image Processing. – 1996. – **5**, № 9. – P. 1293–1302.
2. *The topography* of multivariate normal mixtures / Surajit Ray and Bruce G. Lindsay // The Ann. Statistics. – 2005. – **33**, № 5. – P. 2042–2065.
3. *Расин Дж.* Непараметрическая эконометрика: вводный курс / Джеффри Расин // Квантиль. – 2008. – № 4. – С. 7–56.
4. *Жуков Н.Н.* Вероятностно-статистические методы анализа геолого-геофизической информации. – К.: Вища шк.., 1975. – 304 с.
5. *Жуков М.Н.* Метод багатовимірної статистичної фільтрації різновидової інформації для вирішення задач картування та прогнозу: Дис…д-ра геол. наук. – К., 1997. – 337 с.
6. *Schowengerdt R.A.* Remote sensing: models and methods for image processing. – 3rd ed. – Elsevier Inc., 2007. – 515 p.
7. *Оцінка* стану підтоплення за даними дистанційного зондування методом багатовимірної статистичної класифікації на основі моделі композиційних розподілів / М. Жуков, А. Тішаєва // Вісн. Київ. нац. ун-ту ім. Т. Шевченка. Серія Геологія. – 2010. – Вип. 50. – С. 34–36.

Оценка плотности распределения спектральной яркости на основе композиционной модели в задачах дистанционных зондирований Земли Н.Н. Жуков, А.Н. Тишаева, И.В. Тишаев

РЕЗЮМЕ. Предлагается новый метод оценки плотностей одно- и многомерных распределений для использования в задачах дешифрирования данных дистанционного зондирования Земли на основе модели смеси локальных распределений. Продемонстрированы высокая эффективность, особенно в случае осложненных

неоднородных распределений, приемлемость для описания многомерных распределений, с компонентами, коррелированными в общепринятом понимании. Приведены результаты практического использования.
Ключевые слова: плотность распределения, композиционная модель, спектральные каналы, дистанционные зондирования.

Estimation of spectral radiance density distribution based on the compositional model, remote sensing case study M.N. Zhukov, A.M. Tishaieva, I.V. Tishaiev

SUMMARY. In the paper is suggested new approach for density estimation of univariate and multivariate distributions, which are used in interpretation of remote sensing data. The approach is based on mixture of elemental local distributions. It is shown effectiveness of the method, especially in case of complex inhomogeneous distributions. Acceptability for describing multivariate distributions with correlated components is considered. Practical applications examples are given.
Keywords: probability density function, compositional model, spectral bands, remote sensing.