

## Сжатие гиперспектральных данных методом главных компонент

А.С. Минкин<sup>1</sup>, О.В. Николаева<sup>2</sup>, А.А. Руссков<sup>2</sup>

<sup>1</sup> НИЦ «Курчатовский институт»,  
123182, Россия, г. Москва, пл. Академика Курчатова, д. 1,

<sup>2</sup> Институт прикладной математики РАН,  
123047, Россия, г. Москва, Миусская пл., д. 4

### Аннотация

Цель работы – построение алгоритма сжатия гиперспектральных данных, позволяющего достигнуть высокой степени сжатия при малой мере близости исходного и восстановленного сигналов.

Алгоритм опирается на метод главных компонент и метод исчерпывания. Методом исчерпывания последовательно находятся главные компоненты – сингулярные вектора матрицы исходного сигнала. Параллельно формируется матрица восстановленного сигнала. Процесс продолжается до достижения заданной меры близости исходного и восстановленного сигналов.

Представлено пошаговое описание алгоритма, приведены списки входных и выходных параметров.

Тестирование выполнено на данных эксперимента Aviris. Используются три снимка, отвечающие разной облачной ситуации (чистое небо, частичная облачность и сплошная облачность). Для каждого снимка тестирование выполнено отдельно для всего набора спектральных каналов и для набора, из которого исключены каналы, лежащие в полосах сильного поглощения излучения в водяном паре.

Представлена зависимость мер близости исходного и восстановленного сигналов от степени сжатия. Рассматриваются четыре меры близости: среднее квадратичное отклонение, отношение шума к сигналу, мера структурного сходства и среднее относительное отклонение.

Показано, что меры близости уменьшаются более чем на порядок при исключении из снимка спектральных каналов, лежащих в полосах сильного поглощения. Показано, что причиной этого являются погрешности измерения слабого в полосе поглощения сигнала, из-за чего снижается зависимость между спектрами в разных пространственных пикселях. Для оценки готовности снимка к сжатию представленным алгоритмом предложено использовать среднее по снимку косинусное расстояние между спектрами в разных пространственных пикселях.

**Ключевые слова:** гиперспектральные данные, сжатие данных, метод главных компонент, мера близости.

**Цитирование:** Минкин, А.С. Сжатие гиперспектральных данных методом главных компонент / А.С. Минкин, О.В. Николаева, А.А. Руссков // Компьютерная оптика. – 2021. – Т. 45, № 2. – С. 235-244. – DOI: 10.18287/2412-6179-CO-806.

**Citation:** Minkin AS, Nikolaeva OV, Russkov AA. Hyperspectral data compression based upon the principal component analysis. Computer Optics 2021; 45(2): 235-244. DOI: 10.18287/2412-6179-CO-806.

### Введение

Информационная производительность современных приборов дистанционного зондирования Земли из космоса составляет сотни мегабит/с, тогда как скорость передачи информации в используемых радиолиниях существенно (до 5 раз) меньше [1]. При этом требования к производительности продолжают повышаться, тогда как пропускная способность радиолинии остается ограниченной. Кроме того, постоянно нарастает объем данных дистанционного зондирования в наземных хранилищах. В связи с этим актуальной задачей является развитие алгоритмов сжатия данных зондирования.

При этом сжатие может выполняться без потерь и с потерями. Сжатие без потерь может быть выполнено кодированием данных (алгоритмы zip, rar и т.п.). Также сжатие без потерь достигается с помощью какого-либо ортогонального преобразования сигнала, дискретного косинусного (JPEG) или вейвлет (JPEG 2000), с последующим кодированием коэффициентов разложения сигнала по ортогональным базисным функциям.

Вместо ортогонального преобразования, может быть использовано представление значения каждого пикселя через значения в соседних пикселях с последующим кодированием разницы между истинным и предсказанным значением пикселя (LossLess JPEG).

В HGI (Hierarchical Grid Interpolation) методе для предсказания значения пространственного пикселя в каждом спектральном канале используется последовательная аппроксимация сигнала на иерархической последовательности постепенно разрежаемых сеток [2]. Степень сжатия данных в этих алгоритмах невелика (от 2 до 4 раз) [2, 3].

Усовершенствование алгоритма предсказания за счет учета корреляции между спектральными каналами повышает степень сжатия до 4–5 раз [3]. Развита алгоритм предсказания с использованием нескольких опорных спектральных каналов, что позволяет выполнять декомпрессию сигнала быстрее [4]. Усовершенствование алгоритма предсказания в области неоднородности сигнала позволяет повысить степень сжатия до 1,5 раз [5].

Развивается метод сжатия, опирающийся на аппроксимацию на последовательности разрежаемых сеток не самого сигнала, а коэффициентов его вейвлет-разложения [6]; здесь пока достигается сжатие только в 1,5–2 раза.

*Сжатие с потерями* может быть выполнено отбором наиболее информативных спектральных каналов. Определяются каналы, при отбрасывании которых вероятность распознавания объекта снижается не более чем на 5% [1]. При дополнительном применении архиватора zip к значениям в оставшихся каналах степень сжатия достигает 19 раз. Однако информация в отброшенных каналах полностью теряется.

Высокая степень сжатия может быть достигнута при применении метода главных компонент, если требовать только сохранения формы спектральной кривой [7].

Очень высокая степень сжатия (до 140 раз) достигается при использовании банка данных спектральных сигнатур. Здесь для спектра каждого пикселя подбирается сигнатура и параметры ее преобразования, при которых отклонение спектра от преобразованной сигнатуры меньше заданного параметра [8]. Тогда сжатый сигнал состоит из номеров сигнатур и параметров их преобразования. Однако такой метод требует большого набора сигнатур.

Степень сжатия до 60 раз достигается при отбрасывании несущественных коэффициентов разложения сигнала по дискретным косинусам [9]. Однако в отдельных пикселях для отдельных спектральных каналов исходный и восстановленный сигнал могут различаться в 2 раза. Меньшее отклонение исходного и восстановленного сигналов достигается при меньшей степени сжатия – от 6 до 30 раз [10]. Такие методы устойчиво обрабатывают в том числе и зашумленные сигналы.

Для сжатия с потерями также может быть использован HGI-метод [2, 4, 5]. Степень сжатия при внесении искажения в восстанавливаемый сигнал может быть повышена до 20 раз.

Для сжатия сигналов также используются нейронные сети [11], [12]; с их помощью может быть реали-

зован и метод главных компонент [13]. Однако в задачах сжатия сигналов нейронные сети обладают двумя важными недостатками:

1. Размерность сжатого сигнала задается априори.
2. Применение сети возможно лишь для сигналов, родственных сигналам из обучающей выборки.

В настоящей работе представлен алгоритм сжатия с потерями на основе метода главных компонент. Наша цель состоит в построении алгоритма, позволяющего добиться высокой степени сжатия при малой мере близости исходного и восстановленного сигналов, а не только сохранения формы спектральной кривой, как в [7]. Малая степень близости исходного и восстановленного сигналов необходима, если восстановленный сигнал будет использован для определения оптических и физических параметров атмосферы.

### 1. Меры близости сигналов

Важным является выбор используемой меры близости. Пусть исходный сигнал состоит из  $N$  пространственных пикселей, каждому из которых отвечает  $L$  спектральных каналов. Представим исходный сигнал набором векторов  $\mathbf{x}_n \{x_{n,\ell}, \ell = 1, \dots, L\}, n = 1, \dots, N$ . Восстановленный сигнал обозначим так  $\mathbf{x}'_n \{x'_{n,\ell}, \ell = 1, \dots, L\}, n = 1, \dots, N$ .

В работах [2, 4, 8, 11] используются абсолютные меры близости исходного и восстановленного сигналов

$$E = E_1, \quad E_1 = \max_{n,\ell} |x'_{n,\ell} - x_{n,\ell}|, \quad (1)$$

$$E = E_2, \quad E_2^2 = \frac{1}{NL} \sum_{n=1}^N \sum_{\ell=1}^L (x'_{n,\ell} - x_{n,\ell})^2. \quad (2)$$

Вместо (1) может применяться более «мягкая» оценка [8] меры близости  $E$ :

$$E = E_3, \quad E_3^2 = \frac{1}{L} \max_n \sum_{\ell=1}^L (x'_{n,\ell} - x_{n,\ell})^2. \quad (3)$$

Так как значения сигналов могут сильно различаться в разных спектральных каналах, лучше использовать относительную среднюю квадратичную ошибку [13]

$$E = E_4, \quad E_4^2 = \frac{E_2^2}{\left[ \frac{1}{NL} \sum_{n=1}^N \sum_{\ell=1}^L (x'_{n,\ell})^2 \right]} \quad (4)$$

или же величину

$$E = E_5 = \frac{1}{PSNR}, \quad (5)$$

где  $PSNR$  есть пиковое отношение сигнала к шуму [9, 10, 11]

$$PSNR = \frac{\max_{n,\ell} |x_{n,\ell}|}{E_2}. \quad (6)$$

Таким образом, мера  $E_5$  – отношение шума к сигналу. В качестве меры близости также используется мера структурного сходства

$$E = E_6, \quad E_6 = 1 - MSSIM, \quad (7)$$

где  $MSSIM$  – индекс структурного сходства [12, 15, 16]

$$MSSIM = \frac{1}{T} \sum_{n=1}^{N-p+1} \sum_{\ell=1}^{L-p+1} b_{n,\ell,p} c_{n,\ell,p} s_{n,\ell,p}, \quad (8)$$

$T = (N-p)(L-p)$ , величины  $b_{n,\ell,p}$ ,  $c_{n,\ell,p}$ ,  $s_{n,\ell,p}$  вычисляются для блоков, состоящих из  $p \times p$  пикселей; каждый блок включает пиксели с индексами

$$(n', \ell'), \quad n' = n, \dots, n+p-1, \quad \ell' = \ell, \dots, \ell+p-1.$$

Величины  $b_{n,\ell,p}$ ,  $c_{n,\ell,p}$ ,  $s_{n,\ell,p}$  определяются равенствами:

$$b_{n,\ell,p} = \frac{2\mu_{n,\ell,p} \mu'_{n,\ell,p}}{\mu_{n,\ell,p}^2 + \mu'_{n,\ell,p}{}^2 + C_1} -$$

функционал сравнения яркости в блоке,

$$c_{n,\ell,p} = \frac{2\sigma_{n,\ell,p} \sigma'_{n,\ell,p}}{\sigma_{n,\ell,p}^2 + \sigma'_{n,\ell,p}{}^2 + C_2} -$$

функционал сравнения контрастности в блоке,

$$s_{n,\ell,p} = \frac{r_{n,\ell,p}}{\sigma_{n,\ell,p} \sigma'_{n,\ell,p} + C_3} -$$

функционал сравнения структуры в блоке.

Здесь

$$\mu_{n,\ell,p} = \frac{1}{p^2} \sum_{n'=n}^{n+p-1} \sum_{\ell'=\ell}^{\ell+p-1} x_{n',\ell'} -$$

среднее значение исходного сигнала в блоке,

$$\mu'_{n,\ell,p} = \frac{1}{p^2} \sum_{n'=n}^{n+p-1} \sum_{\ell'=\ell}^{\ell+p-1} x'_{n',\ell'} -$$

среднее значение восстановленного сигнала в блоке,

$$\sigma_{n,\ell,p}^2 = \frac{1}{p^2 - 1} \sum_{n'=n}^{n+p-1} \sum_{\ell'=\ell}^{\ell+p-1} (x_{n',\ell'} - \mu_{n,\ell,p})^2 -$$

несмещенная оценка дисперсии исходного сигнала в блоке,

$$\sigma'_{n,\ell,p}{}^2 = \frac{1}{p^2 - 1} \sum_{n'=n}^{n+p-1} \sum_{\ell'=\ell}^{\ell+p-1} (x'_{n',\ell'} - \mu'_{n,\ell,p})^2 -$$

несмещенная оценка дисперсии восстановленного сигнала,

$$r_{n,\ell,p} = \frac{1}{p^2 - 1} \sum_{n'=n}^{n+p-1} \sum_{\ell'=\ell}^{\ell+p-1} (x_{n',\ell'} - \mu_{n,\ell,p}) \cdot (x'_{n',\ell'} - \mu'_{n,\ell,p}) -$$

несмещенная оценка ковариации сигналов в блоке.

Положительные константы  $C_1$ ,  $C_2$ ,  $C_3$  введены для того, чтобы величина  $MSSIM$  не деградировала в области слабого сигнала (когда  $\mu_{n,\ell,p}^2 + \mu'_{n,\ell,p}{}^2 \sim 0$ ) или в

области постоянного сигнала (когда  $\sigma_{n,\ell,p}^2 + \sigma'_{n,\ell,p}{}^2 \sim 0$ ). В соотношении (8) усреднение выполняется по всем блокам. Индекс  $MSSIM$  был разработан для оценки взаимосвязи в снимок при сжатии визуальных искажений.

Положим  $C_1 = C_2 = C_3 = 0$ . Тогда формула (8) может быть упрощена

$$MSSIM = \frac{1}{(N-p)(L-p)} \sum_{n=1}^{N-p+1} \sum_{\ell=1}^{L-p+1} Y_{n,\ell,p}, \quad (9)$$

$$Y_{n,\ell,p} = \frac{2\mu_{n,\ell,p} \mu'_{n,\ell,p}}{\mu_{n,\ell,p}^2 + \mu'_{n,\ell,p}{}^2} \cdot \frac{2r_{n,\ell,p}}{\sigma_{n,\ell,p}^2 + \sigma'_{n,\ell,p}{}^2}. \quad (10)$$

Эта величина имеет и другое название – UNI (Universal Quality Index).

Мы будем использовать также среднее относительное отклонение

$$E = E_7, \quad E_7 = \frac{1}{\Lambda(\mathbf{x})} \sum_{n=1}^N \sum_{\ell=1, x_{n,\ell}>0}^L |1 - x'_{n,\ell}/x_{n,\ell}|, \quad (11)$$

где  $\Lambda(\mathbf{x})$  – число положительных элементов в векторах  $\mathbf{x}_n$ . Подчеркнем, что неположительные элементы в векторах  $\mathbf{x}_n$  могут появляться только в результате погрешности измерения, например, слабого сигнала, в сопряжении с грубой оценкой фона. В мере (11) оценивается среднее по спектральным и пространственным пикселям относительное отклонение сигналов. Эта мера чувствительна к ошибкам восстановления малых по абсолютной величине значений сигнала.

Оценки (2), (4) и (5), (6) также будем вычислять с использованием только положительных значений сигнала

$$E_2^2 = \frac{1}{\Lambda(\mathbf{x})} \sum_{n=1}^N \sum_{\ell=1, x_{n,\ell}>0}^L (x'_{n,\ell} - x_{n,\ell})^2, \quad (12)$$

$$E_4^2 = \frac{E_2^2}{\left[ \frac{1}{\Lambda(\mathbf{x})} \sum_{n=1}^N \sum_{\ell=1, x_{n,\ell}>0}^L (x'_{n,\ell})^2 \right]}, \quad (13)$$

$$E_5 = \frac{1}{PSNR}, \quad PSNR = \frac{\max_{n,\ell, x_{n,\ell}>0} |x_{n,\ell}|}{E_2}. \quad (14)$$

Аналогично оценка (7), (9) вычисляется с использованием равенств

$$MSSIM = \frac{1}{Z} \sum_{n=1}^{N-p+1} \sum_{\ell=1, \mu_{n,\ell,p}>0, \sigma_{n,\ell,p}>0}^{L-p+1} Y_{n,\ell,p}, \quad (15)$$

$$Z = \sum_{n=1}^{N-p+1} \sum_{\ell=1, \mu_{n,\ell,p}>0, \sigma_{n,\ell,p}>0}^{L-p+1} 1, \quad (16)$$

$$\mu_{n,\ell,p} = \frac{1}{p^2} \sum_{n'=n}^{n+p-1} \sum_{\ell'=\ell, x_{n',\ell'}>0}^{\ell+p-1} x_{n',\ell'}, \quad (16)$$

$$\mu'_{n,\ell,p} = \frac{1}{p^2} \sum_{n'=n}^{n+p-1} \sum_{\ell'=\ell, x'_{n',\ell'}>0}^{\ell+p-1} x'_{n',\ell'}, \quad (17)$$

$$\sigma_{n,\ell,p}^2 = \frac{1}{p^2 - 1} \sum_{n'=n}^{n+p-1} \sum_{\ell'=\ell, x_{n',\ell'}>0}^{\ell+p-1} (x_{n',\ell'} - \mu_{n,\ell,p})^2, \quad (18)$$

$$\sigma_{n,\ell,p}^2 = \frac{1}{p_{n,\ell}^2 - 1} \sum_{n'=n}^{n+p-1} \sum_{\ell'=\ell, x_{n',\ell'} > 0}^{\ell+p-1} (x'_{n',\ell'} - \mu'_{n,\ell})^2, \quad (19)$$

$$r_{n,\ell,p} = \frac{1}{p_{n,\ell}^2 - 1} \sum_{n'=n}^{n+p-1} \sum_{\ell'=\ell, x_{n',\ell'} > 0}^{\ell+p-1} (x_{n',\ell'} - \mu_{n,\ell,p}) \times (x'_{n',\ell'} - \mu'_{n,\ell,p}). \quad (20)$$

Здесь  $p_{n,\ell}$  – число положительных элементов  $x_{n,\ell}$  с индексами из интервала  $\{n \leq n' \leq n+p-1, \ell \leq \ell' \leq \ell+p-1\}$ . В формуле (15) суммирование выполняется с отбрасыванием всех неположительных элементов исходного сигнала.

## 2. Алгоритм сжатия методом главных компонент

Составим из векторов  $\mathbf{x}_n \in R^L, n = 1, \dots, N$ , матрицу  $\mathbf{X} \in R^{L \times N}$ . У этой матрицы существуют левые  $\mathbf{q}_m \in R^L$  и правые  $\mathbf{v}_m \in R^N$  сингулярные вектора, удовлетворяющие условиям

$$\mathbf{X} \mathbf{v}_m = S_m \mathbf{q}_m, \quad \mathbf{X}^T \mathbf{q}_m = S_m \mathbf{v}_m, \quad (21)$$

$$\mathbf{q}_m^T \mathbf{q}_n = \delta_{m,n}, \quad \mathbf{v}_m^T \mathbf{v}_n = \delta_{m,n}.$$

Здесь  $S_m$  – сингулярное число матрицы  $\mathbf{X}$ ,  $\delta_{m,n}$  – символ Кронекера. Тогда для матрицы  $\mathbf{X}$  справедливо представление

$$\mathbf{X} = \sum_{m=1}^N S_m \mathbf{q}_m \mathbf{v}_m^T. \quad (22)$$

Зададим натуральное число  $M$  и выберем наибольшие по модулю сингулярные числа  $S_m, m = 1, \dots, M$ . Матрицу восстановленных векторов  $\mathbf{X}'(M)$  определим равенством

$$\mathbf{X}'(M) = \sum_{m=1}^M \mathbf{X}'_m, \quad (23)$$

где

$$\mathbf{X}'_m = S_m \mathbf{q}_m \mathbf{v}_m^T \in R^{L \times N} - \quad (24)$$

компонента сигнала, отвечающая паре векторов  $\mathbf{v}_m$  и  $\mathbf{q}_m$ . Сингулярные вектора  $\mathbf{q}_m$  задают ортонормированный базис в пространстве  $R^L$ , которому принадлежат вектора значений измеренного сигнала  $\mathbf{x}_n$ . Сингулярные вектора  $\mathbf{v}_m$  содержат коэффициенты разложения векторов  $\mathbf{x}_n$  по базису  $\mathbf{q}_m$ . Формулами (23), (24) восстановленный сигнал  $\mathbf{X}'(M)$  определяется как разложение исходного сигнала по первым  $M$  базисным векторам  $\mathbf{q}_m$ . Оставшаяся часть сигнала содержится в матрице  $\bar{\mathbf{X}}(M+1)$ , определяемой выражением

$$\bar{\mathbf{X}}(M+1) = \mathbf{X} - \mathbf{X}'(M) = \sum_{m=M+1}^N \mathbf{X}'_m.$$

Отсюда следует рекуррентное соотношение

$$\bar{\mathbf{X}}(M+1) = \bar{\mathbf{X}}(M) - \mathbf{X}'_M, \quad M > 0, \quad \bar{\mathbf{X}}(1) = \mathbf{X}. \quad (25)$$

Из выражения (23) следует равенство

$$\mathbf{X}'(M) = \mathbf{X}'(M-1) + \mathbf{X}'_M. \quad (26)$$

Вместо матрицы  $\mathbf{X} \in R^{L \times N}$ , на Землю будем передавать вектора  $\mathbf{q}_m \in R^L, \mathbf{v}_m \in R^N$  и собственные числа  $S_m, m = 1, \dots, M$ . Коэффициент сжатия равен

$$C = (LN) / [(L+N)M]. \quad (27)$$

Заметим, что из выражений (21) следуют равенства

$$\mathbf{X} \mathbf{X}^T \mathbf{q}_m = S_m \mathbf{X} \mathbf{v}_m = S_m^2 \mathbf{q}_m, \quad (28)$$

$$\mathbf{X}^T \mathbf{X} \mathbf{v}_m = S_m \mathbf{X}^T \mathbf{q}_m = S_m^2 \mathbf{v}_m. \quad (29)$$

Поэтому сингулярные вектора  $\mathbf{q}_m$  и  $\mathbf{v}_m$  являются собственными векторами матриц  $\mathbf{X} \mathbf{X}^T \in R^{L \times L}$  и  $\mathbf{X}^T \mathbf{X} \in R^{N \times N}$  соответственно; оба вектора отвечают собственному числу  $S_m = S_m^2$ .

Отметим, что обычно в методе главных компонент предварительно выполняется центрирование векторов, то есть их элементы преобразуются следующим образом

$$x_{n,\ell} := x_{n,\ell} - \bar{x}_\ell, \quad \bar{x}_\ell = (1/N) \sum_{n=1}^N x_{n,\ell}.$$

При этом вектор  $\bar{\mathbf{x}} \{\bar{x}_\ell\}$  принимается в качестве вектора  $\mathbf{q}_1$  с сингулярным числом  $S_1 = 1$ , а матрица  $\mathbf{X} \mathbf{X}^T$  становится матрицей ковариации. Для нахождения остальных векторов  $\mathbf{q}_m$  в этом случае решается уравнение (28), а для вычисления векторов  $\mathbf{v}_m$  используется второе уравнение (21).

Вычисление матрицы  $\mathbf{X} \mathbf{X}^T$  требует много времени; поэтому некоторые авторы предлагают использовать приближенные матрицы ковариации [7].

Мы не вычисляем матрицу  $\mathbf{X} \mathbf{X}^T$ , а решаем непосредственно уравнения (21) с помощью метода исчерпывания. Этот метод позволяет найти старшие сингулярные число и вектора матрицы.

Для нахождения сингулярных векторов  $\mathbf{q}_M, \mathbf{v}_M$  метод исчерпывания применяется к матрице  $\bar{\mathbf{X}}(M)$ . Выполняется следующий итерационный процесс

$$\mathbf{v}_{M,k+1/2}^T = \mathbf{q}_{M,k}^T \bar{\mathbf{X}}(M), \quad (30)$$

$$\mathbf{v}_{M,k+1}^T = \mathbf{v}_{M,k+1/2}^T / \|\mathbf{v}_{M,k+1/2}^T\|, \quad (31)$$

$$\mathbf{q}_{M,k+1/2} = \bar{\mathbf{X}}(M) \mathbf{v}_{M,k+1}, \quad (32)$$

$$\mathbf{q}_{M,k+1}^T = \mathbf{q}_{M,k+1/2}^T / \|\mathbf{q}_{M,k+1/2}^T\|, \quad (33)$$

$$S_{M,k+1} = \mathbf{q}_{M,k+1}^T \bar{\mathbf{X}}(M) \mathbf{v}_{M,k+1} = \mathbf{q}_{M,k+1}^T \mathbf{q}_{M,k+1/2}. \quad (34)$$

Здесь  $k \geq 0$  – номер итерации. При выполнении условия

$$|1 - S_{M,k} / S_{M,k+1}| < \delta, \quad (35)$$

где  $\delta$  – априори заданный малый параметр, итерации прекращаются.

После получения пары сингулярных векторов  $\mathbf{q}_M, \mathbf{v}_M$  находится новая матрица восстановленных векторов  $\mathbf{X}'(M)$ , см. формулы (24), (26), и матрица оставшейся части сигнала  $\bar{\mathbf{X}}(M+1)$  по формуле (25). Да-

лее метод исчерпывания может быть применен уже к матрице  $\bar{\mathbf{X}}(M+1)$  для нахождения следующей пары сингулярных векторов.

Зададим условие выбора числа учитываемых сингулярных чисел  $M$ . Будем требовать, чтобы мера близости  $E$  исходного и восстановленного сигнала не превышала заданного порога  $\Delta$

$$E(\mathbf{X}, \mathbf{X}'(M)) < \Delta. \quad (36)$$

Наконец, все полученные сингулярные числа и элементы сингулярных векторов перед пересылкой на Землю требуется перевести из величин двойной точности в целые  $B$ -битовые числа и применить к ним процедуру квантования.

Таким образом, *входными параметрами* алгоритма сжатия являются следующие величины:

1. Число спектральных каналов  $L$  и число пространственных пикселей  $N$ .
2. Матрица  $\mathbf{X} \in R^{L \times N}$  значений измеренного сигнала.
3. Параметр  $\delta$ , определяющий прекращение числа итераций в методе исчерпывания.
4. Пороговое значение  $\Delta$ .
5. Количество бит  $B$  в представлении сингулярных чисел и векторов.

Алгоритм сжатия может быть записан в виде последовательности шагов.

Задание начального значения количества сингулярных чисел  $M=1$ .

1. Вычисление текущего сингулярного числа  $S_M$  и векторов  $\mathbf{q}_M$  и  $\mathbf{v}_M$  методом исчерпывания (30)–(35).
2. Получение отвечающей полученным векторам компоненты сигнала  $\mathbf{X}'_M$  по формуле (24).
3. Получение матрицы оставшейся части сигнала  $\bar{\mathbf{X}}(M)$  по формуле (25).
4. Получение новой матрицы восстановленных векторов  $\mathbf{X}'(M)$  по формуле (26).
5. Вычисление меры близости  $E(\mathbf{X}, \mathbf{X}'(M))$  исходного и восстановленного сигналов по формуле (13), (12) или (14), (12) или (7), (10), (15–20) или (11).
6. Проверка выполнения условия (36).
7. Если условие не выполнено, то число  $M$  увеличивается на единицу и совершается возврат на шаг 2.
8. Перевод векторов  $\mathbf{q}_M, \mathbf{v}_M$  в  $B$ -битовые целые и квантование полученных целых.
9. Кодирование  $B$ -битовых целых без потери точности.

Отметим, что здесь используются только относительные меры близости, поскольку абсолютные меры (1–3) мало информативны.

Выходными параметрами являются величины.

1. Число сингулярных чисел  $M$ .
2. Значения сингулярных чисел  $S_m, m=1, \dots, M$ .

3. Левые  $\mathbf{q}_m \in R^L$  и правые  $\mathbf{v}_m \in R^N$  сингулярные вектора,  $m=1, \dots, M$ .
4. Достигнутая мера близости  $E$ .
5. Достигнутая степень сжатия  $C$ .

Отметим, что чаще метод главных компонент используется для задач кластеризации изображений, см. например, [14], а не для задач сжатия.

Заметим, что, согласно равенству (27), степень сжатия  $C$  тем больше, чем больше число спектральных каналов  $L$ . Таким образом, алгоритм сжатия более эффективен для гиперспектральных, чем для многоспектральных снимков. Для многоспектральных снимков, например, при  $L=8, N \gg L$ , имеем выражение  $C \sim 8/M$  и степень сжатия оказывается близка к степени сжатия алгоритма без потерь.

### 3. Численные результаты

Тестирование предложенного алгоритма выполним на данных эксперимента Aviris, где имеется  $L=224$  спектральных канала. Каждый снимок состоит из  $N=I \times J$  пространственных пикселей, где  $I$  и  $J$  – число пикселей поперек и вдоль полосы съемки. Используются три снимка, сделанные при разных облачных ситуациях:

- f080709t01p00r09,  $I=1582, J=728$  (чистое небо, см. рис. 1а);
- f080709t01p00r16,  $I=2963, J=761$  (частичная облачность, см. рис. 2а);
- f080709t01p00r06,  $I=2158, J=724$  (сплошная облачность, см. рис. 3а).

Сжатие применяется к значениям интенсивности отраженного излучения ( $\text{Вт}/\text{см}^2$ ). Меры близости  $E_4, E_5, E_6, E_7$  являются безразмерными величинами.

Каждый снимок разделяется на  $N_0=20$  непересекающихся фрагментов. Каждый из фрагментов сжимается независимо; мера близости для всего снимка вычисляется как среднее по всем  $N_0$  фрагментам.

Параметр  $\delta$ , определяющий прекращение числа итераций в методе исчерпывания, полагаем равным  $\delta=10^{-5}$ . Количество бит  $B$  в целочисленном представлении сингулярных чисел и векторов выбираем равным  $B=16$ .

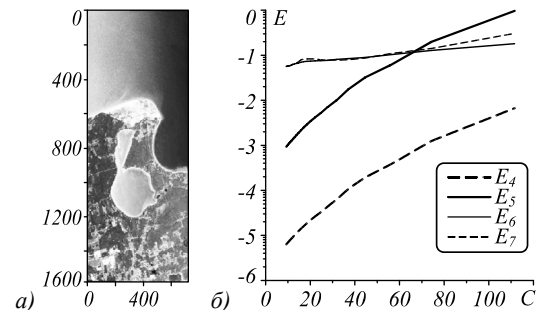


Рис. 1. Результаты работы алгоритма сжатия для снимка 1: чистое небо: снимок (а), меры близости  $E$  в зависимости от степени сжатия  $C$  (б)

На рис. 1–3 представлены обрабатываемые снимки и меры близости  $E_4, E_5, E_6, E_7$  в зависимости от степени сжатия  $C$ . Можно видеть, что

- 1) при уменьшении степени сжатия меры близости  $E_4, E_5, E_6$  убывают монотонно, а мера  $E_7$  – немонотонно;
- 2) из всех мер близости при фиксированной степени сжатия  $C$  наименьшей является мера  $E_4$ ;
- 3) наибольшими являются меры близости  $E_6, E_7$ ;
- 4) меры близости  $E_6, E_7$  соразмерны.

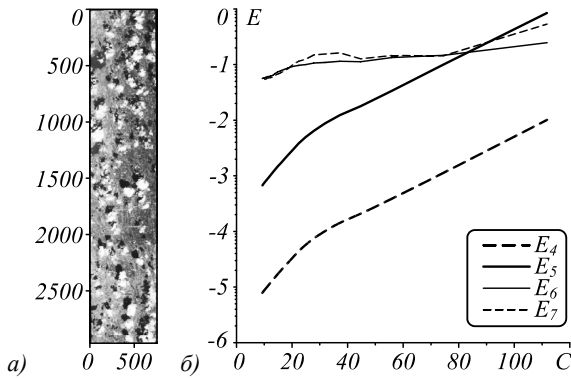


Рис. 2. Результаты работы алгоритма сжатия для снимка 2: частичная облачность: снимок (а), меры близости  $E$  в зависимости от степени сжатия  $C$  (б)

Таким образом, выбор формулы для меры близости существенно влияет на значение погрешности восстановления. Поэтому пороговое значение  $\Delta$  для меры близости следует задавать с учетом того, по какой формуле находится эта мера близости. Также следует учитывать возможный локальный рост меры  $E_7$  при уменьшении степени сжатия  $C$ . Рост объясняется тем, что в PCA-методе минимизируется абсолютная мера (2) [17], тогда как мера  $E_7$  – относительная.

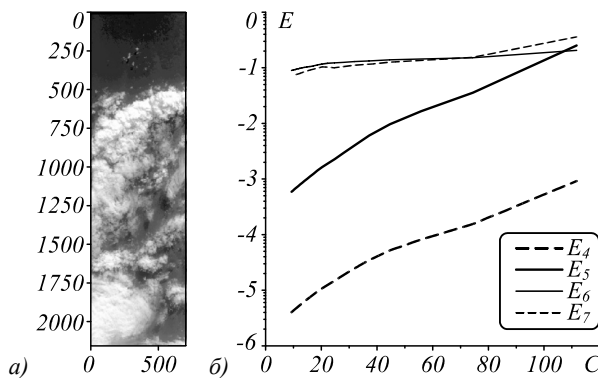


Рис. 3. Результаты работы алгоритма сжатия для снимка 3: сплошная облачность: снимок (а), меры близости  $E$  в зависимости от степени сжатия  $C$  (б)

Рассмотрим вопрос о том, как формируется мера близости  $E_7$ . Представим величину  $E_7$  в виде

$$E_7 = \frac{1}{\Lambda(\mathbf{X})} \sum_{\ell=1}^L \Lambda_{\ell}(\mathbf{X}) e_{7,\ell},$$

$$e_{7,\ell} = \frac{1}{\Lambda_{\ell}(\mathbf{X})} \sum_{n=1, x_{n,\ell} > 0}^N |1 - x'_{n,\ell} / x_{n,\ell}|. \quad (37)$$

Здесь  $e_{7,\ell}$  – относительная мера близости в  $\ell$ -м канале,  $\Lambda_{\ell}(\mathbf{X})$  – число положительных элементов в  $\ell$ -м

столбце матрицы  $\mathbf{X}$ . На рис. 4 приведена зависимость величин  $e_{7,\ell}$  от длины волны  $\lambda_{\ell}$  для каждого снимка при степени сжатия  $C = 15$ .

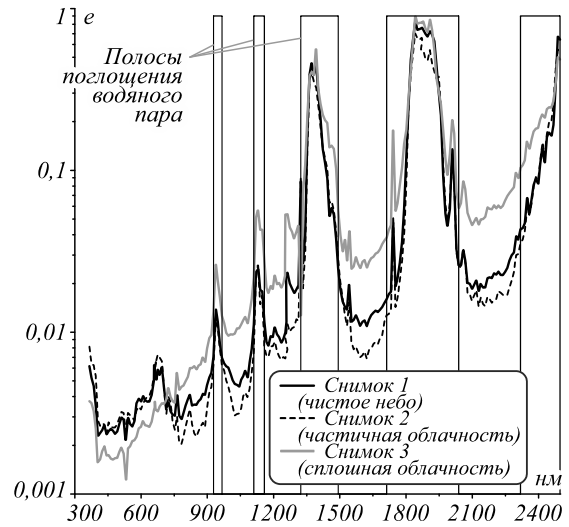


Рис. 4. Зависимость величин  $e_{7,\ell}$  от длины волны  $\lambda_{\ell}$

Там же указано расположение полос поглощения водяного пара. Можно видеть, что вне полос поглощения мера  $e_{7,\ell}$  на один-два порядка меньше, чем в полосах поглощения.

Большие значения меры  $e_7$  в полосах поглощения водяного пара проявляются в том, что восстановленный спектр именно в этих полосах сильно отличается от исходного, см. рис. 5.

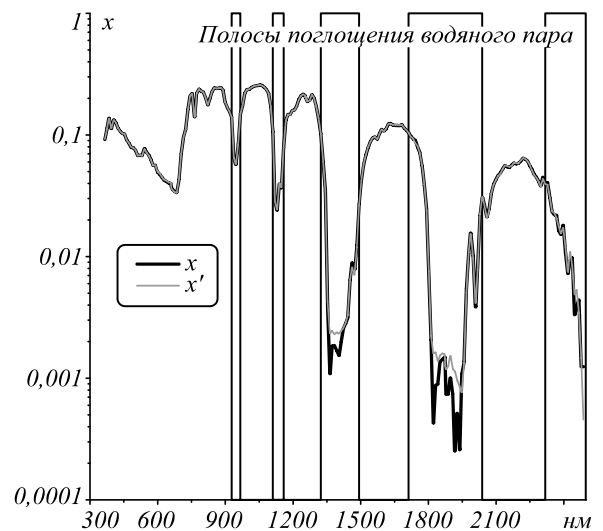


Рис. 5. Исходный и восстановленный спектры (коэффициенты яркости), степень сжатия  $C = 15$ , пиксель из снимка 2

На рис. 5 представлены коэффициенты яркости в двух пространственных пикселях снимка 2 в зависимости от длины волны. Коэффициенты яркости находятся по формулам

$$R_{n,\ell} = \frac{\pi}{\cos \theta_0} \tilde{x}_{\ell} x_{n,\ell}, \quad R'_{n,\ell} = \frac{\pi}{\cos \theta_0} \tilde{x}_{\ell} x'_{n,\ell},$$

где  $\tilde{x}_\ell$  – интенсивность внеатмосферного излучения,  $\theta_0$  – зенитный угол Солнца. Коэффициенты яркости представляют собой отношение интенсивности отраженного атмосферой солнечного излучения к интенсивности излучения, которое было бы отражено атмосферой, если бы ее верхняя граница представляла собой белый ортотропный экран.

Чтобы прояснить причину больших расхождений исходного и восстановленного сигналов в полосах газового поглощения, обратимся к рис. 6 и 7. Величины  $R_{n,\ell}$  безразмерные и являются монотонно убывающими функциями сечений газового поглощения  $\sigma(\lambda_\ell)$  [18]. Поэтому минимумы коэффициента яркости  $R_{n,\ell}$  должны совпадать с максимумами сечений газового поглощения  $\sigma(\lambda_\ell)$ , а максимумы  $R_{n,\ell}$  – с минимумами  $\sigma(\lambda_\ell)$ .

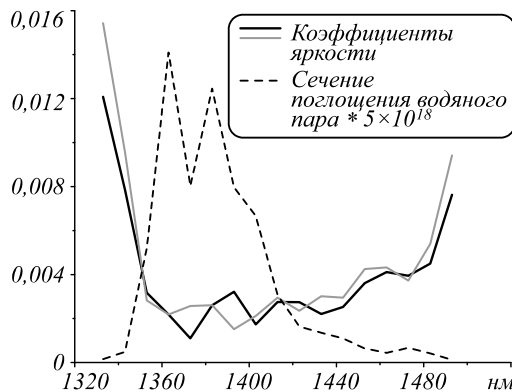


Рис. 6. Исходные спектры (коэффициенты яркости) в полосе поглощения водяного пара [1333, 1492] нм для двух пикселей из снимка 2 и сечение поглощения пара

Это свойство выполнено в полосе несильного поглощения водяного пара, см. рис. 7, но не выполнено в полосе сильного поглощения, см. рис. 6. Более того, при несильном поглощении спектральный ход коэффициентов яркости в двух разных пространственных пикселях подобен, а при сильном поглощении – нет, ср. рис. 6 и рис. 7.

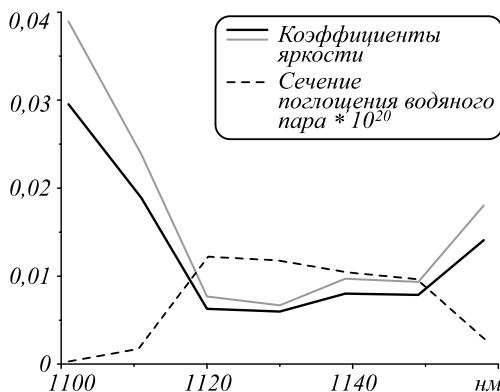


Рис. 7. Исходные спектры (коэффициенты яркости) в полосе поглощения водяного пара [1101, 1158] нм для двух пикселей из снимка 2 и сечение поглощения водяного пара. Взятые те же пиксели, что и на рис. 6

Можно предположить, что такой эффект возникает вследствие больших погрешностей измерения слабого сигнала в полосе сильного поглощения.

В таблице 1 приведены значения среднего косинусного расстояния

$$\rho = \frac{1}{N(N-1)} \sum_{n=1}^N \sum_{n'=1, n \neq n'}^N (\mathbf{x}_n \cdot \mathbf{x}_{n'}) / (\|\mathbf{x}_n\| \|\mathbf{x}_{n'}\|)$$

между значениями сигнала в различных пространственных пикселях отдельно для длин волн в полосах сильного поглощения и вне этих полос.

Табл. 1. Среднее косинусное расстояние между значениями измеренного сигнала в различных пространственных пикселях для каналов, лежащих в полосах сильного поглощения [1333, 1494] нм, [1721, 2028] нм, [2318, 2497] нм и вне их

	В полосах поглощения	Вне полос поглощения
Снимок 1	0,728	0,92
Снимок 2	0,746	0,949
Снимок 3	0,777	0,953

Значения  $\rho$  показывают, что спектры сигнала в разных пространственных пикселях вне полос сильного поглощения близки по форме, в полосах – нет.

Оценим, как присутствие каналов с большой погрешностью измерения влияет на результат работы алгоритма сжатия. Удалим из снимков 1, 2, 3 каналы, лежащие в полосах сильного поглощения в водяном паре: [1333, 1492] нм, [1721, 2028] нм, [2318, 2497] нм.

Применим к получившимся снимкам алгоритм сжатия. Найдем меры близости  $E'_4, E'_5, E'_6, E'_7$ . Их зависимость от степени сжатия  $C$  приведена на рис. 8–10.

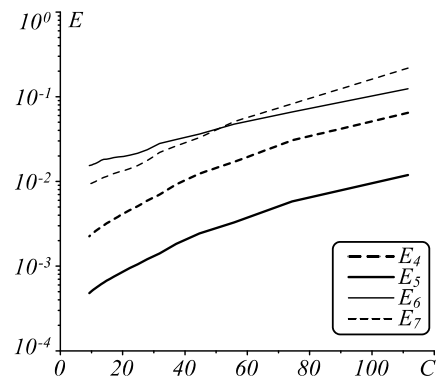


Рис. 8. Меры близости  $E'$  в зависимости от степени сжатия  $C$ , снимок 1

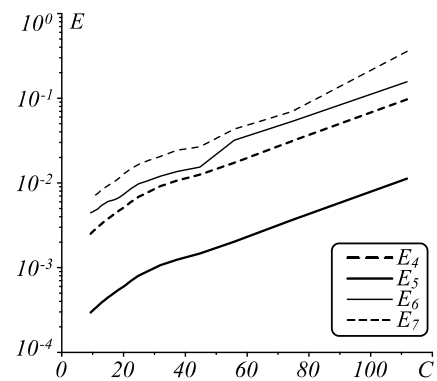


Рис. 9. Меры близости  $E'$  в зависимости от степени сжатия  $C$ , снимок 2

Сравнивая попарно рис. 1 и 8, 2 и 9, 3 и 10, можно видеть, что выбрасывание каналов с сильным поглощением приводит к уменьшению мер близости  $E_5, E_6, E_7$ , но к увеличению меры  $E_4$ :

$$E'_5 < E_5, E'_6 < E_6, E'_7 < E_7, E'_4 > E_4.$$

Рост меры  $E_4$  связан с тем, что в определяющих ее формулах (12), (13) присутствует деление на значения квадрата восстановленного сигнала. В полосах газового поглощения сигнал восстанавливается с большой ошибкой и имеет значительные по величине «выбросы», которые и уменьшают меру  $E_4$ .

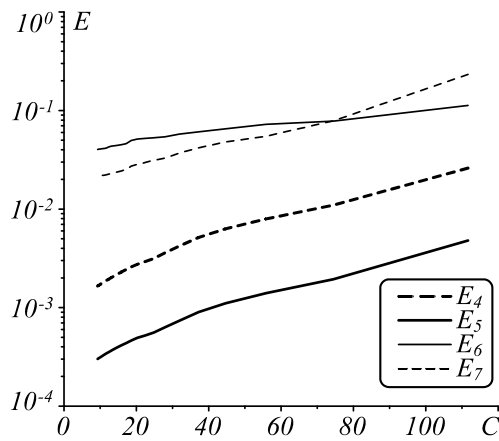


Рис. 10. Меры близости  $E'$  в зависимости от степени сжатия  $C$ , снимок 3

С другой стороны, меры близости  $E_5, E_6, E_7$  при выбрасывании каналов с сильным поглощением уменьшаются более чем на порядок. Кроме того, мера  $E'_7$  с уменьшением степени сжатия  $C$  уменьшается монотонно, в отличие от меры  $E_7$ .

При этом наиболее «жесткими» по-прежнему являются меры  $E'_6$  (мера структурного сходства) и  $E'_7$  (относительное отклонение). Наиболее «мягкой» уже является мера  $E'_5$  (отношение шума к сигналу), а не  $E'_4$  (среднее квадратичное отклонение).

Отметим, что в работе [9] сжатие снимков эксперимента Aviris выполняется с помощью дискретного косинусного преобразования (ДКП) с последующим отбором и кодированием коэффициентов разложения. Мера близости исходного и восстановленного сигналов оценивается с помощью отношения сигнала к шуму. Здесь при степени сжатия  $C \approx 15$  достигается мера близости  $E_5 \approx 0,02$ . Эти результаты близки по порядку величины к результатам, представленным на рис. 1–3. Однако для некоторых пикселей в [9] наблюдается большие (до 2 раз) отклонения исходного и восстановленного сигналов, причем в видимом спектральном диапазоне. Объяснить это можно тем, что в ДКП разложение выполняется по заранее заданной системе ортогональных функций, тогда как в методе главных компонент система ортогональных функций строится на основе сжимаемых данных и более адекватна этим данным.

### Заключение

Предложен алгоритм сжатия гиперспектральных данных дистанционного зондирования, опирающийся на метод главных компонент. Для нахождения сингулярных чисел и векторов предложено использовать метод исчерпывания. Результатом работы алгоритма являются величины, предназначенные для передачи на наземный пункт приема, степень сжатия и точность восстановления данных по переданным на землю величинам.

Приведено подробное описание алгоритма. Алгоритм не содержит вычисление ковариационной матрицы – наиболее затратной операции в обычном варианте метода главных компонент. Другая важная черта представленного алгоритма – возможность выполнять сжатие с априори заданной точностью восстановления.

Представлены результаты тестирования алгоритма на находящихся в свободном доступе данных эксперимента Aviris. Приведены зависимости мер близости исходного и восстановленного сигналов от степени сжатия. Используются четыре формулы для вычисления меры близости: средняя квадратичная ошибка ( $E_4$ ), отношение шума к сигналу ( $E_5$ ), мера структурного сходства ( $E_6$ ) и среднее относительное отклонение ( $E_7$ ). Во всех случаях оцениваются относительные погрешности восстановления сжатого сигнала.

Результаты, полученные в вычислительных экспериментах, позволяют сделать следующие выводы.

1. Меры близости  $E_4, E_5, E_6$  убывают монотонно с уменьшением степени сжатия. Мера  $E_7$  в целом убывает, но локально может несколько возрастать.
2. Исключение из снимка каналов, относящихся к диапазонам [1333, 1492] нм, [1721, 2028] нм, [2318, 2497] нм, где существенно поглощение излучения в водяном паре и наблюдаются признаки низкой точности измерения слабого сигнала, приводит к уменьшению мер близости  $E_5, E_6, E_7$  более чем на порядок.
3. Наиболее строгими являются меры  $E_6, E_7$ ; они соразмерны друг другу. Мера  $E_5$  на тех же данных оказывается на 1–2 порядка меньше, чем меры  $E_6, E_7$ .
4. Среднее косинусное расстояние  $\rho$  между спектрами в разных пространственных пикселях снимка может быть использовано для оценки готовности данных к сжатию представленным алгоритмом. Чем ближе величина  $\rho$  к единице, тем меньшая мера близости исходного и восстановленного сигналов достигается при заданной степени сжатия.
5. Представленный алгоритм позволяет уменьшить объем данных эксперимента Aviris в 10–20 раз при среднем относительном отклонении исходного и восстановленного сигналов порядка  $10^{-1}$  для снимка с полным набором каналов и порядка



$10^{-2}$  для снимка с исключенными каналами в полосе поглощения водяного пара.

В целом, степень сжатия представленным алгоритмом тем больше, чем больше спектральных каналов содержится в снимке.

Представленный алгоритм может быть использован для сжатия гиперспектральных данных с контролем над мерой близости исходного и восстановленного сигналов. Найденное предварительно для снимка среднее косинусное расстояние  $\rho$  позволяет заранее оценить эффективность сжатия снимка данным алгоритмом. При  $\rho$ , близких к единице, достигается высокая степень сжатия при малой мере близости. Оптимально исключать из снимка спектральные каналы с большой погрешностью измерения; это повысит эффективность алгоритма сжатия.

### Литература

1. **Мальцев, Г.Н.** Передача гиперспектральных видеоданных дистанционного зондирования земли по радиоканалам с ограниченной пропускной способностью / Г.Н. Мальцев, И.А. Козин // Информационные каналы и среды. – 2016. – № 2. – С. 74-83. – DOI: 10.15217/issn1684-8853.2016.2.74.
2. **Гашников, М.В.** Иерархическая сеточная интерполяция при сжатии гиперспектральных изображений / М.В. Гашников, Н.И. Глузов // Компьютерная оптика. – 2014. – Т. 38, № 1. – С. 87-93. – DOI: 10.18287/0134-2452-2014-38-1-87-93.
3. **Замятин, А.В.** Алгоритм сжатия гиперспектральных аэрокосмических изображений с учетом междиапазонной корреляции / А.В. Замятин, А.Ж. Сарина // Прикладная информатика. – 2013. – Т. 47, № 5. – С. 35-42.
4. **Гашников, М.В.** Иерархическая компрессия в задаче хранения гиперспектральных изображений / М.В. Гашников, Н.И. Глузов // Компьютерная оптика. – 2014. – Т. 38, № 3. – С. 482-488. – DOI: 10.18287/0134-2452-2014-38-3-482-488.
5. **Максимов, А.И.** Адаптивная интерполяция многомерных сигналов при дифференциальной компрессии / А.И. Максимов, М.В. Гашников // Компьютерная оптика. – 2018. – Т. 42, № 4. – С. 679-687. – DOI: 10.18287/2412-6179-2018-42-4-679-687.
6. **Перцев, Д.Ю.** Параметрически настраиваемый алгоритм сжатия гиперспектральных данных с применением вейвлет-разложения / Д.Ю. Перцев, А.А. Дудкин // Доклады Белорусского государственного университета информатики и радиоэлектроники. – 2019. – Т. 119, № 1. – С. 26-31.
7. **Смирнов, С.И.** Применение рандомизированного метода главных компонент для сжатия данных гиперспектральной съемки / С.И. Смирнов, В.В. Михайлов, В.Н. Остриков // Современные проблемы дистанционного зондирования Земли из космоса. – 2014. – Т. 11, № 2. – С. 9-17.
8. **Лебедев, Л.И.** Комплексный анализ и мониторинг состояния окружающей среды на основе данных ДЗЗ / Л.И. Лебедев, Ю.В. Ясаков, Т.Ш. Утешева, В.П. Громов, А.В. Борусяк, В.Е. Турлапов // Компьютерная оптика. – 2019. – Т. 43, № 2. – С. 282-295. – DOI: 10.18287/2412-6179-2019-43-2-282-295.
9. **Чичева, М.А.** Сжатие гиперспектральных данных на основе метода кодирования с преобразованием / М.А. Чичева, Р.Р. Юзькив // Компьютерная оптика. – 2014. – Т. 38, № 4. – С. 794-803. – DOI: 10.18287/0134-2452-2014-38-4-794-803.
10. **Земляченко, А.Н.** Процедуры автоматического сжатия изображений дистанционного зондирования земли при сигнально-зависимых помехах в окрестности оптимальной рабочей точки и их эффективность / А.Н. Земляченко, Р.А. Кожемякин, С.К. Абрамов, В.В. Лукин, М.К. Чобану // Исследование Земли из космоса. – 2013. – № 3. – С. 73-90. – DOI: 10.7868/S0205961413020140.
11. **Лёзин, И.А.** Сжатие изображения с использованием многослойного персептрона / И.А. Лёзин, А.В. Соловьёв // Известия Самарского научного центра РАН. – 2016. – Т. 18, № 4(4). – С. 770-774.
12. **Шишкин, О.Г.** Сжатие с потерями мультиспектральных снимков на основе искусственной нейронной сети / О.Г. Шишкин // Современные наукоемкие технологии. – 2018. – № 12. – С. 387-391.
13. **Сирота, А.А.** Обобщённые алгоритмы сжатия изображений на фрагментах произвольной формы и их реализация с использованием искусственных нейронных сетей / А.А. Сирота, М.А. Дрюченко // Компьютерная оптика. – 2015. – Т. 39, № 5. – С. 751-761. – DOI: 10.18287/0134-2452-2015-39-5-751-761.
14. **Пестунов, И.А.** Блочный метод главных компонент для выделения информативных признаков при классификации гиперспектральных изображений / И.А. Пестунов, П.В. Мельников // Журнал Сибирского Федерального Университета. Техника и технологии. – 2015. – Т. 8, № 6. – С. 715-725. – DOI: 10.17516/1999-494X-2015-8-6-715-725.
15. **Радченко, Ю.С.** Статистика структурных изменений изображений на основе спектрального и корреляционного анализа полей / Ю.С. Радченко, В.А. Игнатов // Вестник Воронежского государственного университета. Серия: Физика. Математика. – 2014. – № 2. – С. 29-41.
16. **Wang, Zh.** Image quality assessment: From error visibility to structural similarity / Zh. Wang, A. Bovik, H. Sheikh, E. Simoncelli // IEEE Transactions on Image Processing. – 2004. – Vol. 13, Issue 4. – P. 600-612. – DOI: 10.1109/TIP.2003.819861.
17. **Gorban, A.** Principal manifolds for data visualisation and dimension reduction / A. Gorban, B. Kegl, D. Wunsch, A. Zinovyev. – Berlin, Heidelberg, New York: Springer, 2007. – 367 p.
18. **Николаева, О.В.** Малопараметрическая аппроксимация коэффициента яркости солнечного излучения в полосе газового поглощения / О.В. Николаева // Математическое моделирование. – 2020. – Т. 32, № 2. – С. 24-36. – DOI: 10.20948/mm-2020-02-02.

### Сведения об авторах

**Минкин Александр Сергеевич**, 1980 года рождения, в 2003 году окончил Московский инженерно-физический институт (государственный университет), к.ф.-м.н., с.н.с. НИЦ «Курчатовский институт». Область научных интересов: статистический анализ данных, машинное обучение, нейронные сети.

E-mail: [amink@mail.ru](mailto:amink@mail.ru).

**Николаева Ольга Васильевна**, 1971 года рождения, в 1993 году окончила факультет вычислительной математики и кибернетики Московского государственного университета им. М.В. Ломоносова, к.ф.-м.н., с.н.с. Института прикладной математики им. М.В. Келдыша РАН. Область научных интересов: прямые и обратные задачи для уравнения переноса излучения. E-mail: [nika@kiam.ru](mailto:nika@kiam.ru).

**Руссков Александр Алексеевич**, 1982 года рождения, в 2005 году окончил обучение на кафедре системной интеграции и менеджмента факультета общей и прикладной физики Московского физико-технического института, к.ф.-м.н., н.с. Института прикладной математики им. М.В. Келдыша РАН. Область научных интересов: перенос излучения, модели статистической физики, анализ временных рядов. E-mail: [sasha-rusсков@yandex.ru](mailto:sasha-rusсков@yandex.ru).

---

*ГРНТИ: 89.57.35*

*Поступила в редакцию 2 сентября 2020 г. Окончательный вариант – 11 ноября 2020 г.*

---

---

# Hyperspectral data compression based upon the principal component analysis

A.S. Minkin<sup>1</sup>, O.V. Nikolaeva<sup>2</sup>, A.A. Russkov<sup>2</sup>

<sup>1</sup>RSC Kurchatov Institute, 123182, Moscow, Russia, Kurchatov Sq 1,

<sup>2</sup>Keldysh Institute of Applied Mathematics RAS, 123047, Moscow, Russia, Miusskaya Sq 4

## Abstract

The paper is aimed at developing an algorithm of hyperspectral data compression that combines small losses with high compression rate.

The algorithm relies on a principal component analysis and a method of exhaustion. The principal components are singular vectors of an initial signal matrix, which are found by the method of exhaustion. A retrieved signal matrix is formed in parallel. The process continues until a required retrieval error is attained.

The algorithm is described in detail and input and output parameters are specified.

Testing is performed using Aviris data (Airborne Visible-Infrared Imaging Spectrometer). Three images of differently looking sky (clear sky, partly clouded sky, and overcast skies) are analyzed. For each image, testing is performed for all spectral bands and for a set of bands from which high water-vapour absorption bands are excluded.

Retrieval errors versus compression rates are presented. The error formulas include the root mean square deviation, the noise-to-signal ratio, the mean structural similarity index, and the mean relative deviation.

It is shown that the retrieval errors decrease by more than an order of magnitude if spectral bands with high gas absorption are disregarded. It is shown that the reason is that weak signals in the absorption bands are measured with great errors, leading to a weak dependence between the spectra in different spatial pixels. A mean cosine distance between the spectra in different spatial pixels is suggested to be used to assess the image compressibility.

**Keywords:** hyperspectral data, data compression, principal component analysis, proximity measure.

**Citation:** Minkin AS, Nikolaeva OV, Russkov AA. Hyperspectral data compression based upon the principal component analysis. *Computer Optics* 2021; 45(2): 235-244. DOI: 10.18287/2412-6179-CO-806.

## References

- [1] Maltsev GN, Kozinov IA. Transfer of Hyperspectral Video Data of Earth Remote Sensing via Radio Channels of Limited Capacity [In Russian]. *Information Channels and Environments* 2016; 2: 74-83. DOI: 10.15217/issn1684-8853.2016.2.74.
  - [2] Gashnikov MV, Glumov NI. Hierarchical grid interpolation for hyperspectral image compression. *Computer Optics* 2014; 38(1): 87-93. DOI: 10.18287/0134-2452-2014-38-1-87-93.
  - [3] Zamyatin AV, Sarinova AZh. Compressing hyperspectral space images accounting intra-band correlation [In Russian]. *Applied Informatics* 2013; 47(5): 35-42.
  - [4] Gashnikov MV, Glumov NI. Hierarchical compression for hyperspectral image storage. *Computer Optics* 2014; 38(3): 482-488. DOI: 10.18287/0134-2452-2014-38-3-482-488.
  - [5] Maximov AI, Gashnikov MV. Adaptive interpolation of multidimensional signals for differential compression. *Computer Optics* 2018; 42(4): 679-687. DOI: 10.18287/2412-6179-2018-42-4-679-687.
  - [6] Pertsau DY, Doudkin AA. The parametrically adjusted hyperspectral data compression algorithm based on wavelet decomposition [In Russian]. *Doklady BGUIR* 2019; 119(1); 26-31.
  - [7] Smirnov SI, Mikhailov VV, Ostrikov VN. Application of randomized principal component analysis for compression of hyperspectral data [In Russian]. *Current Problems in Remote Sensing of the Earth from Space* 2014; 11(2): 9-17.
  - [8] Lebedev LI, Yasakov YuV, Utesheva TH, Gromov VP, Borusjak AV, Turlapov VE. Complex analysis and monitoring of the environment based on Earth sensing data. *Computer Optics* 2019; 43(2): 282-295. DOI: 10.18287/2412-6179-2019-43-2-282-295.
  - [9] Chicheva MA, Yuzkiv RR. Hyperspectral data compression based on a transform coding method. *Computer Optics* 2014; 38(4): 794-803. DOI: 10.18287/0134-2452-2014-38-4-794-803.
  - [10] Zemlyachenko AN, Kozhemyakin RA, Abramov SK, Lukin VV, Tchobanou MK. Procedures for automatic image compression in remote sensing of the Earth for signal dependent noise working in the vicinity of optimal operating point and their effectiveness [In Russian]. *Research of Earth from Space* 2013; 3: 73-90. DOI: 10.7868/S0205961413020140.
  - [11] Lyozin IA, Solovyov AV. Performing an image compression by using the multilayer perceptron [In Russian]. *Izvestia of Samara Scientific Center of the Russian Academy of Sciences* 2016; 28(4): 770-774.
  - [12] Shishkin OG. Compression with losses of multispectral images that are based on artificial neural network [In Russian]. *Modern High Technologies* 2018; 12: 387-391.
  - [13] Sirota AA, Dryuchenko MA. Generalized image compression algorithms for arbitrarily-shaped fragments and their implementation using artificial neural networks. *Computer Optics* 2015; 39(5): 751-761. DOI: 10.18287/0134-2452-2015-39-5-751-761.
  - [14] Pestunov IA, Melnikov PV. Block principal component analysis for extraction of informative features for classification of hyperspectral images [In Russian]. *Journal of Siberian Federal*
-

- 
- University. Engineering & Technologies 2015; 8(6): 715-725. DOI: 10.17516/1999-494X-2015-8-6-715-725.
- [15] Radchenko YS, Ignatov VA. Statistics of structural changes of images on the basis of spectral and correlation analysis fields [In Russian]. Proceedings of Voronezh State University. Series: Physis. Mathematics 2014; 2: 29-41.
- [16] Wang Zh, Bovik A, Sheikh H, Simoncelli E. Image quality assessment: From error visibility to structural similarity. IEEE Trans Image Process 2004; 13(4): 600-612. DOI: 10.1109/TIP.2003.819861.
- [17] Gorban A, Kegl B, Wunsch D, Zinovyev A. Principal manifolds for data visualisation and dimension reduction. Berlin, Heidelberg, New York: Springer; 2007.
- [18] Nikolaeva OV. Low-parametrical approximation solar radiation's brightness coefficients gas absorption band. Math Models Comput Simul 2020; 12(5): 770-778. DOI: 10.1134/S2070048220050154.
- 

#### *Authors' information*

**Alexander S. Minkin** (b. 1980) graduated from Moscow Engineer Physics University in 2003, PhD, works for NRC "Kurchatov Institute" in position of senior researcher. Research interests: statistical data analysis, machine learning, neural networks. E-mail: [amink@mail.ru](mailto:amink@mail.ru).

**Olga V. Nikolaeva** (b. 1971) graduated from Computational Mathematic and Cybernetics department of Moscow State University, PhD, works as senior researcher in the Keldish Institute of Applied Mathematics RAS. Research interests: direct and inverse problems for radiation transport equation. E-mail: [nika@kiam.ru](mailto:nika@kiam.ru).

**Alexander A. Russkov** (b.1982) graduated from General and Applied Physics department of the Moscow Institute of Physics and Technology, PhD, works as researcher in the Keldish Institute of Applied Mathematics RAS. Research interests: radiation transport, statistic physics simulation, time series analysis. E-mail: [sasha-russkov@yandex.ru](mailto:sasha-russkov@yandex.ru).

---

*Received September 2, 2020. The final version – November 11, 2020.*

---