Современные проблемы дистанционного зондирования Земли из космоса. 2014. Т. 11. № 2. С. 9-17

# Применение рандомизированного метода главных компонент для сжатия данных гиперспектральной съемки

**С.И. Смирнов, В.В. Михайлов, В.Н. Остриков**

*Санкт-Петербургский филиал ОАО «КБ «Луч»*
*Санкт-Петербург 197376, Россия*
*E-mail: luchspbf@rambler.ru*

Работа посвящена исследованию вопросов, связанных со сжатием данных гиперспектральной съемки. Среди ряда известных выбран метод главных компонент, что обусловлено возможностью проведения последующей спектральной идентификации непосредственно в пространстве сжатия с экономией времени на обработку за счет его значительно меньшей размерности. Кроме того, универсальность снижения вычислительных затрат на основе использования именно такого метода обоснована необходимостью решать, на основе получаемых данных, широкий круг задач спектральной идентификации. Вместе с тем, классическая реализация метода главных компонент имеет достаточно высокую вычислительную сложность, в ходе которой наибольшее количество операций в силу обычно высокого пространственного разрешения современных приборов тратится на вычисление ковариационной матрицы. В работе исследуются методы снижения объема ее определения за счет применения рандомизации.
В литературе существует несколько подходов реализации такой рандомизации. В части работ предлагают использовать случайную выборку из пикселей гиперкуба, в других – проецировать осредненный куб в пространство меньшей размерности с применением преобразования Джонсона - Линденштрауса. В работе исследованы оба подхода на предмет применимости к данным гиперспектральной съемки, полученным с авиационных носителей. Оба метода опробованы на нескольких реальных гиперкубах. Приведены результаты сравнения используемых подходов.

**Ключевые слова**: МГК, сжатие, гиперспектральные данные, рандомизация, преобразование Джонсона-Линденштрауса.

## Введение

В настоящее время возрастает интерес к дистанционному зондированию с использованием аппаратуры гиперспектральной съемки. Области применения включают детектирование очагов лесных пожаров, идентификацию мест вырубок леса, обнаружение областей загрязнения окружающей среды (мониторинг свалок, выделение зон разливов нефти), поиск зараженных вредителями сельскохозяйственных угодий и многое другое. При этом актуальным является использование гиперспектральной аппаратуры в системах “реального времени”, когда результаты передаются к месту обработки в процессе съемки или эта обработка проводится непосредственно на борту носителя, что требует минимизации вычислительных затрат для реализации сжатия данных. Применение, например, таких методов сокращения избыточности, как пошаговые методы возможного объединения каналов для снижения размерности «куба» данных, в рассматриваемом случае не представляется возможным, поскольку при этом качество решения задачи, контролируемое ошибкой спектрального распознавания конкретных объектов, также на конкретных фонах, находится в существенной зависимости от типа решаемой тематической задачи. Такой подход особенно неприемлем, когда на основе

одних и тех же данных решается значительная совокупность тематических задач. В этой ситуации более целесообразно использование универсальных методов сжатия, качество функционирования которых контролируется, в конечном итоге, влиянием вносимых искажений в спектральные векторы идентифицируемых поверхностей.

Известно несколько широко используемых универсальных методов сжатия данных гиперспектральной съемки (ГСС), в частности, метод главных компонент (МГК, в англоязычной литературе Principal Component Analysis, PCA), 3-D JPEG 2000, 3-D Wavelet. Поскольку МГК позволяет проводить последующую тематическую обработку непосредственно в базисе сжатия, целесообразно рассмотреть вопросы применения МГК и некоторых его модификаций, нацеленных на прямое ускорение его функционирования.

Современная бортовая гиперспектральная система может охватывать рабочий спектральный диапазон, включающий видимую и инфракрасную области, иметь в этом диапазоне сотни каналов со спектральным разрешением от долей нанометра, при этом число элементов разрешения (пикселей) в строке изображения может достигать нескольких тысяч. Рассмотрим ускорение процесса вычисления ковариационной матрицы и реализации компонентного разложения (общепринятая аббревиатура – SVD-разложение) в рамках МГК за счет применения рандомизации.

В литературе рассматривается два основных подхода к проблеме рандомизации SVD-разложения. Так, в статье (Drineas, Kannan, Mahoney, 2006) рекомендуется случайный выбор строк матрицы, тогда как в работе (Halko, Martinsson, Tropp, 2011) рекомендуется производить редукцию размерности пространства за счет преобразования Джонсона-Линденштрауса. Ниже проводится анализ этих подходов на предмет применимости к данным ГСС.

## Постановка задачи и описание алгоритма сжатия по МГК

Исходными данными для обработки являются результаты ГСС с авиационного носителя (в дальнейшем – гиперкуб). На *рис. 1* отображены особенности представления этих данных.

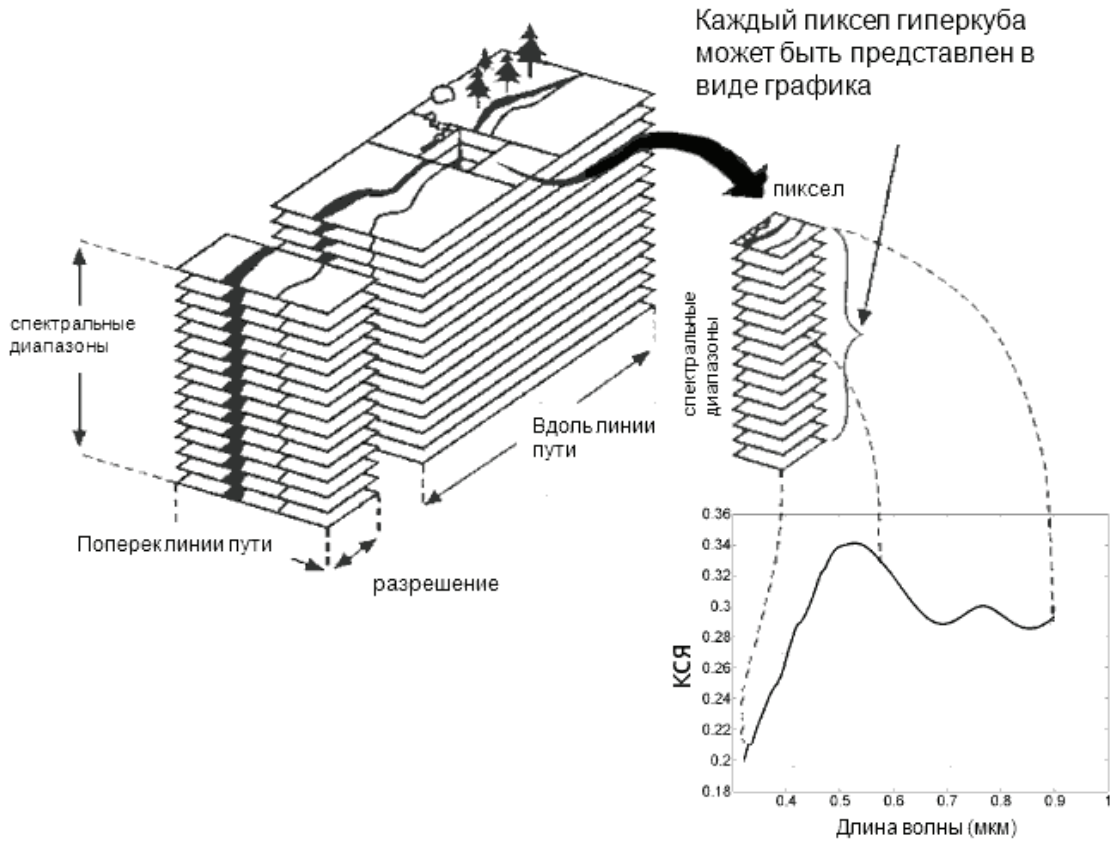



*Рис. 1. Особенности представления данных гиперспектральной съемки*

Задача сжатия заключается в проецировании гиперкуба из исходного спектрального пространства в пространство меньшей размерности. Опишем суть МГК (Jollife, 2002).

**Вход**: Гиперкуб $H(i,j)$ размера $N \times d$, где $N = w \times h$, $w,h$ - пространственное разрешение в пикселях по горизонтали и вертикали соответственно, $d$ - количество спектральных каналов.

**Выход**: Сжатый гиперкуб $G(i,\tilde{j})$ размера $N \times \tilde{d}$, где $\tilde{d}$ - количество каналов в базисе сжатия.

**Алгоритм:**

1. Находится ковариационная матрица спектральных каналов $C$ размером $d \times d$, каждый элемент которой $C(k,l) = \mathrm{cov}(H(:,k), H(:,l))$ (выборочной ковариации каналов $k$ и $l$).

2. Получаем ее SVD-разложение в виде $C = U^T DU$, где $U^T$ - матрица, транспонированная к ортогональной матрице $U$, $D$ - диагональная матрица, $D$ (1,1)≥ $D$ (2,2)…≥ $D$ ($d$,$d$).

3. В зависимости от уровня шума выбирается $m$ старших компонент, тогда базис пространства сжатия $V = \left[ U(1)^T, U(2)^T, \ldots, U(m)^T \right]$, где $U(i)$ - $i$-я строка матрицы $U$.

4. Рассчитываем $G(i,j) = H(i,j) - \overline{H(:,j)}$, $\overline{H(:,j)}$ - выборочное среднее для канала j.

5. По формуле $Q = GV$ получаем сжатый куб.

### *Алгоритм декомпрессии*

**Вход**: гиперкуб $G\left(i,\tilde{j}\right)$ размера $N \times \tilde{d}$, где $\tilde{d}$ - количество каналов в базисе сжатия, матрица преобразований $V$, среднее значение по каналам $\overline{H(:,j)}$, $j \in 1,...,d$.

**Выход**: Гиперкуб $\tilde{H}\left(i,j\right)$ размера $N \times d$, где $N = w \times h$, $w,h$ - пространственное разрешение в пикселях по горизонтали и вертикали соответственно, $d$ - количество спектральных каналов.

**Алгоритм**: пересчет по формуле $\tilde{H}\left(i,j\right) = \left(GV^T\right)(i,j) + \overline{H(:,j)}$, $\forall i,j$.

Оценим вычислительные затраты приведенного алгоритма. Для гиперкуба $H\left(i,j\right)$ размера $N \times d$ вычисление ковариационной матрицы требует операций порядка $O\left(N \times d^2\right)$, вычисление SVD-разложения - порядка $O\left(d^3\right)$. При условии $N >> d$, основные затраты составляют вычисление матрицы $C$.

Вычисления ускоряются за счет «подмены» ковариационной матрицы её несмещенной оценкой по подмножеству пикселей исходного гиперкуба значительно меньшей мощности.

Первый подход использует рандомизированный метод, где вероятность выбора пикселя связана с его спектральной характеристикой. В рамках этого подхода строится функция распределения вероятностей, на основе которой и выбираются элементы множества. (Drineas, Kannan, Mahoney, 2006).

Второй подход связан с преобразованием каждого канала в пространство существенно меньшей размерности по лемме Джонсона-Линденштраусса (один из вариантов можно найти в статье (Halko, Martinsson, Tropp, 2011), в статье (Zhang et al., 2012) - с реализацией для гиперспектральных изображений).

Для рассматриваемых подходов имеются оценки изменчивости нормы матрицы (фробениусова и спектральная). Однако в нашем случае важно, чтобы алгоритм “не портил” спектральные характеристики пикселей гиперспектрального снимка в смысле специальной метрики, введенной в (Остриков, Смирнов, Михайлов, 2013).

Введем критерий качества алгоритма сжатия: потребуем, чтобы сжатие сохраняло форму спектральной кривой, то есть, если рассматривать пиксель сжатого по нашему алгоритму снимка (РМГК) и по стандартному методу главных компонент, то нормализован-

ные спектральные кривые $c_{РМГК}$ и $c_{МГК}$ не должны существенно отличаться в терминах метрики Теребижа (Теребиж, 2005):

$$T\left(\rho_{МГК},\rho_{РМГК}\right)=\sum_{i=1}^{d}\frac{\left(\rho_{МГК}\left(\lambda_i\right)-\rho_{РМГК}\left(\lambda_i\right)\right)^2}{\rho_{МГК}\left(\lambda_i\right)}\ . \qquad (1)$$

Таким образом, если ввести порог $\varepsilon_T$, то качество сжатия будет определяться долей тех пикселей, которые лежат вне «$\varepsilon_T$ - коридора» точки 0 ($T\left(\rho_{МГК},\rho_{РМГК}\right)(i,j)>\varepsilon_T$). Ниже приведены алгоритмы для рандомизированного метода и метода, основанного на лемме Джонсона-Линденштраусса.

### *Алгоритм рандомизированного метода*

**Вход**: Гиперкуб $\tilde{H}(i,j)=H(i,j)-\overline{H(:,j)}$ размера $N\times d$, где $N=w\times h$, $w,h$ - пространственное разрешение в пикселях по горизонтали и вертикали соответственно, $d$ - количество спектральных каналов.

**Выход**: Оценка для ковариационной матрицы $C$ (далее по стандартному алгоритму проделываем шаги 2-5).

**Алгоритм:**

1. Вычисляем вероятности для моделирования и средние значения по каналам.

2. Моделируем нужное количество индексов строчек r в соответствии с таблицей распределения $P$:

$$P\left(i\right)=\frac{\tilde{H}\left(i,:\right)\cdot\left(\tilde{H}(i,:)\right)^T}{\sum_{j=1}^{N}\tilde{H}\left(j,:\right)\cdot\left(\tilde{H}(j,:)\right)^T}\ . \qquad (2)$$

3. Получаем оценку ковариационной матрицы по формуле:

$$C=\frac{1}{r}\sum_{i=1}^{r}\frac{1}{P\left(\tau_i\right)}\cdot\left(\tilde{H}\left(\tau_i,:\right)\right)^T\cdot\tilde{H}\left(\tau_i,:\right)\ . \qquad (3)$$

### *Алгоритм рандомизации с использованием преобразования Джонсона-Линденштрауса*

**Вход**: Гиперкуб $\tilde{H}(i,j)=H(i,j)-\overline{H(:,j)}$ размера $N\times d$, где $N=w\times h$, $w,h$ - пространственное разрешение в пискелях по горизонтали и вертикали соответственно, $d$ - количество спектральных каналов, $s$ - параметр разреженности преобразования Джонсона-Линденштрауса (в примерах берем $s=5$).

**Выход**: Матрица $V = \left[ U(1)^T, U(2)^T, \ldots, U(k)^T \right]$.

1. Получаем матрицу $R(i, j)$, каждый элемент которой есть независимая реализация случайной величины

$$\xi = \begin{cases} -1, & p = \frac{1}{2s} \\ 0, & p = 1 - \frac{1}{s} \\ 1, & p = \frac{1}{2s} \end{cases}.$$

2. Получаем матрицу $\tilde{B} = \tilde{H}R$.

3. Производим ортогонализацию столбцов матрицы B, получаем матрицу $Q = orth(B)$.

4. Получаем матрицу $B = Q^T \tilde{B}$.

5. Из SVD матрицы $B$ получаем матрицу $\tilde{U}$ (первые $k$ компонент) такую, что $B = \tilde{U}\Sigma V^T$.

6. Получаем матрицу $\hat{U} = Q\tilde{U}$.

7. Получаем матрицу компрессии $V = \tilde{H}\hat{U}$.

Оценим вычислительную сложность приведенных выше алгоритмов. Для первого алгоритма при разумном выборе параметра $r$ преобразование сжатия рассчитывается за $O(kd^2)$ операций. Для второго – за $O(kdN)$ операций ($k$ – количество главных компонент). Для приведенных выше порядков величин $d$ и $N$ следует, что первый алгоритм будет существенно быстрее второго. Проецирование в пространство сжатия осуществляется за одинаковое время для всех алгоритмов.

Следует особо отметить, что многие метрики исходного спектрального пространства могут быть без проблем перенесены в базис сжатия. В частности, рассмотрим класс метрик вида $\rho_D(\xi, \eta) = (\xi - \eta)^T D(\xi - \eta)$.

В том случае, если используется алгоритм сжатия по методу главных компонент (или его рандомизированный аналог), можно представить каждый из спектральных векторов в виде $\tilde{\xi} = U(\xi - \mu)$ и $\tilde{\eta} = U(\eta - \mu)$. При подстановке в формулу для вычисления метрики соотношений декомпрессии $\xi = U^T\tilde{\xi} + \mu$ и $\eta = U^T\tilde{\eta} + \mu$ получим, что метрика в исходном базисе может быть выражена в виде

$$\rho_D(\xi, \eta) = \left(U^T\tilde{\xi} + \mu - U^T\tilde{\eta} - \mu\right)^T D\left(U^T\tilde{\xi} + \mu - U^T\tilde{\eta} - \mu\right) = \left(\tilde{\xi} - \tilde{\eta}\right)^T UDU^T\left(\tilde{\xi} - \tilde{\eta}\right).$$

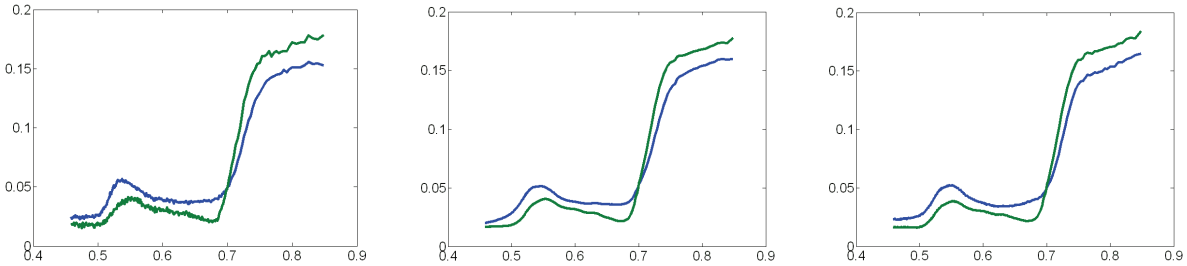
Приведенные выше цепочки равенств позволяют утверждать, что для любой метрики в исходном пространстве найдется эквивалентная ей метрика в пространстве сжатия, и формула расчета этой метрики известна. В данном случае наиболее интересен тот факт, что евклидова метрика сохраняется при переходе в сжатый базис. Кроме того, метрика Теребижа, приведенная выше, полностью удовлетворяет этому случаю, если в качестве $D$ взять матрицу, диагональные элементы которой составлены из величин, обратных к $\rho_{min}$ (Остриков, Смирнов, Михайлов, 2013).

Эти соображения могут быть использованы для создания систем бортовой компрессии данных.

Рассмотренные алгоритмы реализованы в комплексе функциональных программ (Остриков, Плахотников, Кириенко, 2013) и протестированы на нескольких наборах данных. Исследования показали, что рандомизированный алгоритм и алгоритм с использованием преобразования Джонсона – Линденштрауса незначительно меняют форму спектральных кривых. По введенному качеству сжатия (1) метод с использованием преобразования Джонсона - Линденштрауса чуть лучше, однако различия незначительны.

Сравнительное время выполнения алгоритмов для используемого вычислителя на кубе ГСС размером 238*465*289 (289 - число спектральных каналов) составило: 17 секунд для стандартной реализации МГК, 4 секунды для метода с использованием преобразования Джонсона - Линденштрауса и около 1 секунды для рандомизированного метода. Такая оценка по времени близка к асимптотической при том условии, что в расчет берется время применения преобразования ко всему гиперкубу, и свидетельствует, что оба алгоритма существенно превосходят стандартную реализацию метода главных компонент по времени быстродействия при незначительном ухудшении качества сжатия в смысле формы спектральной кривой. Вместе с тем, первый рандомизированный алгоритм значительно превосходит второй по быстродействию.

В качестве примера влияния сжатия на форму спектральной кривой можно привести различия двух близких спектральных типов поверхностей одного из гиперкубов тестового набора (*рис. 2*).




а) б) в)

*Рис. 2. Спектральные кривые двух близких типов исходного куба (а), сжатого по методу главных компонент (б) и сжатого по рандомизированному методу главных компонент (в)*

Как видно из рисунка, при разумном выборе количества главных компонент сжатие не влияет в рассматриваемом случае на форму спектральной кривой.

## Выводы

Исследуемый в работе подход к рандомизации вычисления преобразования метода главных компонент может быть применен к процессу сжатия данных на борту носителя. Применение метода как к данным в пространстве коэффициентов спектральной яркости, так и в пространстве спектральной плотности энергетической яркости показало, что сжатие в целом не ухудшает дальнейшую спектральную идентификацию кривых, например, с использованием алгоритма из (Остриков, Смирнов, Михайлов, 2013). При разумном выборе числа компонент в некоторых случаях происходит также снижение количества ложных срабатываний при классификации (эффект фильтрации случайных шумов). Вместе с тем, остается ряд открытых вопросов по границам применимости метода и уточнению оценок для погрешностей. Кроме того, требует проработки вопрос реализации всего цикла обработки данных в рамках базиса сжатия.

## Литература

1. *Остриков В.Н., Плахотников О.В., Кириенко А.В.* Обработка гиперспектральных данных, получаемых с авиационных и космических носителей // Современные проблемы дистанционного зондирования земли из космоса. 2013. Т. 10. № 2. С. 243-252.
2. *Остриков В.Н., Смирнов С.И., Михайлов В.В.* Алгоритм двухэтапной классификации гиперспектральных данных в пространстве коэффициентов спектральной яркости по результатам авиационной съемки // Современные проблемы дистанционного зондирования земли из космоса. 2013. Т. 10. № 3. С. 75-84.
3. *Drineas P., Kannan R., Mahoney M.V.* Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix // SIAM J. Comput. 2006. Vol. 36. No. 1. P. 158–183.
4. *Jollife I.T.* Principal component analysis. Second edition // Springer. NY. 2002.
5. *Halko N., Martinsson P. G., Tropp J. A.* Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions // SIAM Review. 2011. Vol. 53. No. 2. P. 217–288.
6. *Zhang J., Erway J., Hu X., Zhang Q., Plemmons R.* Randomized SVD Methods in Hyperspectral Imaging // Journal of Electrical and Computer Engineering. 2012. Vol. 2012.

# Application of randomized principal component analysis for compression of hyperspectral data

**S.I. Smirnov, V.V. Mikhailov, V.N. Ostrikov**

*Joint-Stock Company «Luch», branch in St.-Petersburg*
*Saint-Petersburg 197376, Russia*
*E-mail: luchspbf@rambler.ru*

The paper is devoted to issues related to compression of hyperspectral data. Among a number common methods, principal component analysis (PCA) is chosen because it allows to produce spectral identification of data in compressed space with economy of processing time because of much smaller dimension of this space. Moreover, this decreasing of computational costs is universal for a wide class of spectral identification problems. However, the classical implementation of the PCA method has a relatively high computational complexity, in which the highest number of operations because of high spatial resolution of modern sensors is spent on the calculation of the covariance matrix. In this paper, methods for reducing the computational complexity via randomization are discussed.
In the literature, there are several approaches to this problem. Some authors suggest to use random sampling of hypercube pixels, some other - project an averaged cube to the space of lower dimension via Johnson - Lindenstrauss transform. Two these approaches are studied for implementation on hyperspectral data obtained from aircraft sensors. Both methods were tested on several sets of real data. A comparison of these approaches is given.

**Keywords**: PCA, compression, hyperspectral data, randomization, Johnson - Lindenstrauss transform.

## References

1. Ostrikov V.N., Plakhotnikov O.V., Kirienko A.V., Obrabotka giperspektral'nykh dannykh, poluchaemykh s aviatsionnykh i kosmicheskikh nositelei (Hyperspectral images processing from aircraft and satellite instruments), *Sovremennye problemy distantsionnogo zondirovaniya zemli iz kosmosa*, 2013, Vol. 10, No. 2, pp. 243-252.
2. Ostrikov V.N., Smirnov S.I., Mikhailov V.V., Algoritm dvukhetapnoi klassifikatsii giperspektral'nykh dannykh v prostranstve koeffitsientov spektral'noi yarkosti po rezul'tatam aviatsionnoi s"emki (Two-step algorithm for the classification of hyperspectral data in the space of the spectral brightness coefficients of aerial photography), *Sovremennye problemy distantsionnogo zondirovaniya zemli iz kosmosa*, 2013, Vol. 10, No. 3, pp. 75-84.
3. Drineas P., Kannan R., Mahoney M.V., Fast Monte Carlo algorithms for matrices II: Computing a low-rank approximation to a matrix, *SIAM J. Comput.*, 2006, Vol. 36 , No. 1, pp. 158–183.
4. Jollife I.T., *Principal component analysis*. Second edition, Springer, NY, 2002.
5. Halko N., Martinsson P.G., Tropp J.A., Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions, *SIAM Review*, 2011, Vol. 53, No. 2, pp. 217–288.
6. Zhang J., Erway J., Hu X., Zhang Q., Plemmons R., Randomized SVD Methods in Hyperspectral Imaging, *Journal of Electrical and Computer Engineering*, 2012, Vol. 2012.