Image Reconstruction from Double Random Projection
Qiang Zhang, Robert J. Plemmons

Abstract—We present double random projection methods for reconstruction of imaging data. The methods draw upon recent results in the random projection literature, particularly on low-rank matrix approximations, and the reconstruction algorithm has only two simple and non-iterative steps, while the reconstruction error is close to the error of the optimal low-rank approximation by the truncated singular-value decomposition. We extend the often-required symmetric distributions of entries in a random-projection matrix to asymmetric distributions, which can be more easily implementable on imaging devices. Experimental results are provided on the subsampling of natural images and hyperspectral images, and on simulated compressible matrices. Comparisons with other random projection methods are also provided.

Index Terms—random projection, random matrix, compressive sensing, compressible matrices, natural images, hyperspectral images.

I. INTRODUCTION

THE random projection problem is a linear inverse problem stated as,

\[ b = P^T a, \]

where \( a \in \mathbb{R}^n \) is the true signal to recover, \( P \in \mathbb{R}^{n \times k} \) is the projection matrix, with independent and identically distributed (i.i.d.) random entries, and \( b \in \mathbb{R}^k \) is the observed signal. Sometimes columns of \( P \) can be orthogonalized and normalized. A theorem by Candès, Rumberg and Tao [1] states that if \( k \geq cn \log n \), where \( c \) is a constant and \( m \) is the number of nonzeros in \( a \), then \( a \) is the unique solution to the convex optimization problem,

\[ \min \|a\|_1, \text{ subject to } P^T a = b, \]

for an overwhelming percentage of column vector sets, \( \Omega \), with cardinality \( k \).

In the compressive imaging literature, there have been several efforts on applying random projection methods. For example, the double-disperser coded-aperture snapshot imager (DD-CASSI) [2], and the single-disperser coded-aperture snapshot imager (SD-CASSI) [3], can snapshot a hyperspectral image cube, and both systems can be formulated as (1), where the random entries in \( P \) are calibrated from coded apertures. For these systems, because problem (1) is generally highly underdetermined with \( k \ll n \), extensive efforts are needed for reconstruction with regularizations such as \( l_1 \) or total variation [4]. Marcia and Willet [5] have also used a coded aperture to superresolve a low-resolution image, and considered a system matrix decomposed into

\[ P^T = DF^{-1}C_HF, \]

where \( D \) is the downsampling matrix, \( F \) is the Fourier matrix, and \( C_H \) is a diagonal matrix with entries from the Fourier transform of a point spread function (PSF). Here, the matrix \( F^{-1}C_HF \) has a block-circulant-with-circulant-blocks (BCCB) structure. Marks and Brady [6] have considered the compressive coding of focal tomography, where the projection matrix is decomposed into

\[ P^T = HT, \]

and here, \( H \) is the optical system matrix and \( T \) is a random matrix with i.i.d. entries drawn from the uniform distribution on \([0, 1]\). Notice that this distribution is asymmetric around the origin, and hence does not preserve Euclidean length, i.e.,

\[ E(||P^T f||_2^2) \neq ||f||_2^2, \]

since a symmetric distribution is a necessary condition for the Euclidean length preservation [7]. Also, \( P \) does not satisfy the restricted isometry property (RIP).

When the true signal is an image, rather than vectorizing it in a lexicographical order before the random projection, Fowler [8] applied random projections to the columns of a hyperspectral image matrix, i.e.,

\[ B = P^T A, \]

where \( A \in \mathbb{R}^{m \times n} \) is the matrix form of an image. This approach, called compressive-projection principal component analysis (CPPCA), not only enjoys the clear advantage of a smaller projection matrix, \( m \times k \) rather than \( mn \times k \), and hence faster computation, but it also preserves row structures, specifically the range of row vectors, which would be lost by vectorization. In the same line of thought, Eftekhari, Babaie-Zadeh and Moghaddam [9] applied projections on both columns and rows simultaneously, i.e.,

\[ B = P_1^T AP_2, \]

where \( P_1 \in \mathbb{R}^{m \times k_1} \) and \( P_2 \in \mathbb{R}^{n \times k_2} \) are random matrices satisfying RIP. All the approaches mentioned above involve iterative reconstruction algorithms.

To ensure an accurate or even exact reconstruction of \( f \), studies by Candès et al., [1], [10], [11] have pointed out the necessity of a parsimonious representation of \( f \) on a certain basis. However, finding the optimal prespecified basis,
e.g., wavelets, is often problem-specific and tested empirically [12]. As is well known, many natural images often have a sharp-dropping spectrum of singular values, and hence, an image could likely have a parsimonious representation on the basis composed of its singular vectors, which means that the chosen basis is data-driven, rather than prespecified. Fowler [8] actually used this idea that if one projects along the column dimension of an image, the eigenvectors of \(BB^T\) in (6) will provide a good approximation to the eigenvectors of \(AA^T\), especially when the number of columns is far greater than the number of rows, i.e., \(m \ll n\). However, this would not be true for the \(B\) in (7), which leads us to the idea that if we project along both dimensions individually, i.e.,

\[
B_1 = P_1^T A, \quad B_2 = AP_2,
\]

we may be able to use the preserved singular vectors to reconstruct the original \(A\) in a direct, non-iterative way. Several more recent compressed-sensing algorithms, e.g., smoothed, projected Landweber reconstruction (BCS-SPL) [13], its multiscale variant (MS-BCS-SPL) [14], and the multiple-hypothesis-prediction variant (MH-MS-BCS-SPL) [15], and a similar algorithm using collaborative sparsity (RCoS) [16], all involve random projections on the 2D wavelet space, which can be more compressive than the column- or row-based projections for natural images.

In [17], [18], the authors presented a randomized singular-value decomposition (rSVD) method for the purposes of lossless compression, reconstruction, classification, and target detection with hyperspectral image (HSI) data. Recent work on low-rank matrix approximations obtained from one-sided random projections suggests that these approximations are well-suited for randomized dimensionality reduction. In these papers, approximation errors for the rSVD are evaluated and compared to deterministic techniques, as well as to other randomized low-rank matrix approximation methods involving compressive principal component analysis, such as CPPCA.

Our paper is organized in the following way. In Sec. 2, we discuss the main concepts of double random projection (DRP) methods along with theorems to provide bounds for reconstruction, and we also discuss the feasibility of using random matrices with asymmetrically distributed entries by introducing an interpretation of a compressible image. In Sec. 3, numerical examples are provided to demonstrate the effectiveness of the method, and we conclude in Sec. 4 with discussions on implementing the DRP using point spread functions.

II. METHOD

A. Double Random Projection Algorithm

We first describe a fast low-rank approximation algorithm using random projections [7], from which we have drawn inspiration. Let \(P_1 = (p_{ij})_{m \times k_1}\) be a random matrix such that \(p_{ij}\) are drawn i.i.d. from \(\mathcal{N}(0, 1)\). The algorithm has three simple steps, i.e.,

1) compute a random projection of an image, \(A\), as

\[
B_1 = \frac{1}{\sqrt{k_1}} P_1^T A,
\]

where \(A \in \mathbb{R}^{m \times n}\), \(P_1 \in \mathbb{R}^{m \times k_1}\), \(B_1 \in \mathbb{R}^{k_1 \times n}\).

2) compute the singular-value decomposition (SVD) of \(B_1\), \(B_1 = \sum_{i=1}^{k} \lambda_i \hat{u}_i \hat{v}_i^T\),

3) return:

\[
\hat{A}_k \leftarrow A \left( \sum_{i=1}^{k} \hat{v}_i \hat{v}_i^T \right) = A \hat{V}_k \hat{V}_k^T.
\]

It is shown in [7] that with a high probability, the approximation error of \(\hat{A}_k\) in Frobenius norm is bounded by,

\[
\|A - \hat{A}_k\|_F \leq \|A - A_k\|_F + 2c\|A_k\|_F^2,
\]

where \(A_k\) is the truncated SVD, i.e.,

\[
A = USV = \sum_{i=1}^{n} \sigma_i u_i v_i^T, \quad \text{and} \quad A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^T.
\]

By the Eckart-Young theorem [19], the optimal low-rank approximation in terms of Frobenius norm error, is \(A_k\), but the approximation \(\hat{A}_k\) is not too far away from \(A_k\) as shown by the bound above.

If we have a second random projection along the rows, i.e.,

\[
B_2 = AP_2,
\]

where \(P_2 \in \mathbb{R}^{n \times k_2}\) and \(B_2 \in \mathbb{R}^{m \times k_2}\). Then we can use \(\hat{A}_k\) to approximate \(A\) as,

\[
B_2 \approx A \hat{V}_k \hat{V}_k^T P_2.
\]

Notice that because \(\hat{V}_k^T P_2\) has a full row rank, we can take its Moore-Penrose pseudo-inverse to change (14) as

\[
B_2(\hat{V}_k^T P_2)^\dagger \approx A \hat{V}_k.
\]

Multiplying both sides of the equation above by \(\hat{V}_k^T\), we have

\[
\hat{A}_k = B_2(\hat{V}_k^T P_2)^\dagger \hat{V}_k \approx A \hat{V}_k \hat{V}_k^T \approx A_k.
\]

The matrix \(\hat{A}_k\) in (16) will be our reconstruction formula of \(A\) from \(B_1\) and \(B_2\). Note that directly taking the pseudo-inverse of \(P_2\) to reconstruct \(A\) from \(B_2\) will likely give poor results, because \(P_2 P_2^T\) is rank-deficient. We describe the method in Algorithm 1.

**Algorithm 1**: A double random projection (DRP) reconstruction algorithm

**Input**: two projected images, \(B_1 = P_1^T A\), and \(B_2 = AP_2\), where \(A \in \mathbb{R}^{m \times n}\), \(P_1 \in \mathbb{R}^{m \times k_1}\), \(B_1 \in \mathbb{R}^{k_1 \times n}\), \(P_2 \in \mathbb{R}^{n \times k_2}\), \(B_2 \in \mathbb{R}^{m \times k_2}\).

**Output**: \(\hat{A}_k\), a rank-k approximation of \(A\).

1. Compute the SVD of \(B_1\),

\[
B_1 = \hat{U} \Sigma \hat{V}^T.
\]

2. Select only the first \(k\) columns of \(\hat{V}\) as \(\hat{V}_k\) and reconstruct an approximation \(\hat{A}_k\) to \(A\),

\[
\hat{A}_k = B_2(\hat{V}_k^T P_2)^\dagger \hat{V}_k^T.
\]

Given this simple algorithm, we would like to know how well it performs in terms of error norm and what properties the projection matrices need to satisfy.
B. Approximation Error

With the bound on the approximation error of \( \hat{A}_k \) in (11), we only need to compare our reconstruction \( A_e \) with \( \hat{A}_k \). Let \( A = A_k + A_e \), where we regard \( A_e \) as an error term, and then we can write in exact terms,

\[
B^2 = (P_2^T \hat{V}_k) \hat{V}_k^T A_e + P_2^T A_e^T. \tag{19}
\]

The error term in (19) is \( P_2^T A_e^T \), and we can seek a least-squares solution for \( \hat{V}_k^T A_e \), i.e., \( \min_A \| B^2 - (P_2^T \hat{V}_k) A \|_2^2 \). If \( P_2 \) is a random matrix with entries drawn from a sub-Gaussian distribution, e.g., Gaussian, Bernoulli or uniform distributions, asymptotically there is a probability of 1 that \( P_2 \) has a full rank [20]. Thus, because \( \hat{V}_k \) is orthonormal, we know that Eq. (19) has a unique least-squares solution.

**Proposition 1.** If \( k_2 \geq k \), \( P_2 \) is a random matrix with entries drawn from a sub-Gaussian distribution, and \( \hat{V}_k \) is any orthonormal matrix, then when \( n \to \infty \), with a probability of 1, Eq. (19) has a unique least-squares solution, \( (P_2^T \hat{V}_k) \hat{V}_k^T \).

**Proof.** Because \( k_2 \geq k \), Eq. (19) is over-determined and because \( P_2^T \hat{V}_k \) has a full rank \( k \) with a probability close to 1, the solution is unique, given by \( (P_2^T \hat{V}_k) \hat{V}_k^T \).

In the following discussions, the assumptions in Proposition 1, we also assume the random matrices \( P_1 \) and \( P_2 \), to be full rank in the asymptotic sense. For the difference between two rank-\( k \) approximations, \( \hat{A}_k - \hat{A}_k \), the following proposition states that it is exactly the error term \( A_e \), multiplied by a square matrix, \( P = P_2(\hat{V}_k^T P_2)^{\dagger} \hat{V}_k \).

**Proposition 2.** The difference between two rank-\( k \) approximations of \( A \) is

\[
\hat{A}_k - \hat{A}_k = A_e P, \tag{20}
\]

where \( P = P_2(\hat{V}_k^T P_2)^{\dagger} \hat{V}_k \).

**Proof.** Notice that \( \hat{A}_k = AP \), and

\[
\hat{A}_k P = A \hat{V}_k \hat{V}_k^T P_2(\hat{V}_k^T P_2)^{\dagger} \hat{V}_k = A \hat{V}_k (\hat{V}_k^T P_2 P_2^T \hat{V}_k)(\hat{V}_k^T P_2 P_2^T \hat{V}_k)^{\dagger} \hat{V}_k \tag{21}
\]

Hence, we have

\[
\hat{A}_k - \hat{A}_k = AP - \hat{A}_k P = A_e P. \tag{22}
\]

From Proposition 2, we know the approximation error of \( \hat{A}_k \) depends solely on the properties of \( P \). Specifically, if \( P \) would enlarge the norm of \( A_e \), our approximation would be rather poor. However, we will show that if certain conditions on \( n, k_2 \) and \( k \) are met, the expected Frobenius norm of \( A_k - \hat{A}_k \) will not exceed that of \( A_k - A \). We first state some basic facts of \( P \) in the following:

1. \( P \) is a square matrix, \( n \times n \), and has a rank \( k \).
2. \( \hat{A}_k P = \hat{A}_k \) and \( \hat{A}_k = AP \).
3. \( P \hat{V}_k = 0 \), for \( i = k + 1, \ldots, n \).
4. \( \hat{V}_k^T P = \hat{V}_k^T \) and hence \( \hat{V}_k^T P \hat{V}_k = \hat{V}_k^T P \hat{V}_k = I_k \).
5. \( P^2 = P \), and hence, \( P \) is idempotent, and \( \text{trace}(P) = k \).

6) \( P = Q_1 Q_2 \), where \( Q_1 = P_2 P_2^T \) and \( Q_2 = \hat{V}_k(\hat{V}_k^T P_2 P_2^T \hat{V}_k)^{-1} \hat{V}_k^T \). Both \( Q_1 \) and \( Q_2 \) are symmetric, though \( P \) is not symmetric.

**Proposition 3.** Matrix \( P \) preserves the length of vectors in a subspace \( S \) of \( \mathbb{R}^n \), where \( S = \{ x | x = \hat{V}_k \bar{x}, x \in \mathbb{R}^n, \bar{x} \in \mathbb{R}^k \} \).

**Proof.** From the fourth property of \( P \), we know

\[
\| P^T x \|_2 = \| \hat{V}_k^T P^T \|_2 = \| \hat{V}_k^T P \hat{V}_k \|_2 = \| \hat{V}_k^T \|_2 = \| x \|_2. \tag{23}
\]

**Proposition 4.** Matrix \( P \) can be decomposed into the sum of two matrices, i.e.,

\[
P = \hat{V}_k^T P_2 + P_2^T \hat{V}_k \tag{24}
\]

where \( P_2 \) has rank \( k \), and its columns are spanned by vectors perpendicular to \( \hat{V}_k \).

**Proof.** Let \( P_k = P_2 \hat{V}_k^T \hat{V}_k + P_2^T \hat{V}_k - 1 \), and hence \( P = \hat{V}_k^T P_2 \hat{V}_k \). Then from the fourth property of \( P \), we know

\[
\hat{V}_k P = I_k, \quad \text{or} \quad \hat{V}_k P_k = 0. \tag{25}
\]

Let \( P_2 = P_2 - \hat{V}_k \), and clearly its span is orthogonal to \( \hat{V}_k \).

One deterministic bound on \( \| P^T x \|_2 \), as opposed to probabilistic bounds, is to use the 2-norm of \( P \), or its largest singular value, i.e.,

\[
\| P^T x \|_2 \leq \| P \|_2 \| x \|_2 = \lambda_1 \| x \|_2, \tag{26}
\]

where \( \lambda_1 \) is the largest singular value of \( P \). This will provide us a rough idea on whether \( P \) would significantly alter the length of a vector. By intuition, we know the inverse term in \( Q_2 \) might effectively cancel any expansion caused by \( Q_1 \), and hence the overall expansion may not be too large. By the following theorem, we will show that indeed \( \lambda_1 \) is not too large.

**Theorem 1.** As \( k, k_2, n \to \infty \) while \( k/k_2 \) and \( k_2/n \) remain constants, the largest singular value of \( P \) satisfies

\[
\lambda_1 \leq \frac{\left( \frac{\sqrt{n} + \sqrt{k_2}}{\sqrt{k_2} - \sqrt{k}} \right)^2}{2}. \tag{27}
\]

**Proof.** Since \( P = Q_1 Q_2 \), and by the fact that

\[
\| P \| \leq \| Q_1 \| \| Q_2 \|, \tag{28}
\]

we know \( \lambda_1 \) would be bounded by the product of the largest singular values of \( Q_1 \) and \( Q_2 \). By the Marčenko-Pastur distribution [21], we know the asymptotic largest singular value of \( P \) is

\[
\lambda_1(2) = n \sqrt{2} \left( 1 + \sqrt{\frac{k_2}{n}} \right)^2. \tag{29}
\]
where \( \sigma \) is the variance of entries in \( P_2 \), and the \( k_2^{th} \) singular value is,
\[
\lambda_{k_2}^{(2)} = n\sigma^2 \left( 1 - \sqrt{\frac{k_2}{n}} \right)^2 .
\] (30)

Similarly, the asymptotic maximum and minimum singular values of \( \hat{V}_k^T P_2 P_2^T \hat{V}_k \) are respectively,
\[
\lambda_1^{(3)} = k_2\sigma^2 \left( 1 + \sqrt{\frac{k}{k_2}} \right)^2 , \quad \lambda_k^{(3)} = k_2\sigma^2 \left( 1 - \sqrt{\frac{k}{k_2}} \right)^2 .
\] (31)

Then the first \( k \) singular values of \( \hat{V}_k (\hat{V}_k^T P_2 P_2^T \hat{V}_k)^{-1} \hat{V}_k^T \) would simply be the reciprocal of the \( \lambda_i^{(3)} \), \( i = 1, \ldots, k \). Hence we have the largest singular value of \( P \) bounded as
\[
\lambda_1 \leq \frac{\lambda_{k_2}^{(2)}}{\lambda_{k_1}^{(3)}} = \frac{n}{k_2} \left( 1 + \frac{k_2}{\sqrt{k_2}} \right)^2 = \left( \frac{\sqrt{n} + \sqrt{k}}{\sqrt{k_2} - \sqrt{k}} \right)^2 .
\] (32)

**Remark 1.** Asymptotically, the distribution of the singular values of \( P \) does not depend on the variance of entries in \( P_2 \). The theorem remains true for asymmetric (nonzero-mean) distributions of entries in \( P_2 \), though these distributions do not satisfy RIP.

Next, by setting \( n = 5,000, k_2 = 2,000 \) and \( k = 1,000 \), we simulate \( P_2 \) using both a Gaussian distribution, \( \mathcal{N}(0,1) \), and a uniform distribution on \([0,1] \). For simulating \( \hat{V}_k \), we run the Gram-Schmidt orthogonalization [19] on a random Gaussian matrix to make it orthonormal. We then compute the singular values of matrices, \( Q_1, \hat{V}_k^T P_2 P_2^T \hat{V}_k, (\hat{V}_k^T P_2 P_2^T \hat{V}_k)^{-1} \) and \( P \), and plot their normalized histograms in Fig. 1. For the first two histograms, we also plot in red lines the Marchenko-Pastur distributions of eigenvalues of a covariance matrix, \( C = HH^T \), where \( H \in \mathbb{R}^{P \times N} \) is a random matrix, entries of which have variance \( \sigma^2 \). The explicit form of the distribution is,
\[
Pr(\lambda) = \frac{1}{2\pi\sigma^2 r} \sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)} ,
\] (33)

where \( \lambda_+ \) and \( \lambda_- \) are the maximum and minimum eigenvalues respectively, and they are
\[
\lambda_\pm = \sigma^2 (1 \pm r)^2 ,
\] (34)

where \( r = p/N \).

From Fig. 1, we see the singular values of \( Q_1 \) and \( \hat{V}_k^T P_2 P_2^T \hat{V}_k \) follow closely to the Marchenko-Pastur distributions and the singular values of \( P \) are all smaller than 6, suggesting it will not greatly expand the length of a vector it is multiplied with. Not only, after multiplying the simulated \( P \) with 10,000 random vectors, in Fig. 2, we plot the histograms of ratios, \( \|P^T x\|_2/\|x\|_2 \), with \( P_{25} \) having entries following both the normal and the uniform distributions. Figure 2 suggests that \( P \) tightly preserves the length of vectors it is multiplied with. Now we move on to study a probabilistic bound of \( \|P^T x\|_2^2 \).
Fig. 2. The histograms of ratios, $\|P^T x\|_2^2/\|x\|_2^2$.

**Theorem 2.** If $\tilde{V}_k$ is any random orthonormal matrix, and $P_2$ has i.i.d. entries drawn from a distribution with a finite variance, then the asymptotic expectation of the ratio $\|P^T x\|_2^2/\|x\|_2^2$ is,

$$
\lim_{n, k_2 \to \infty, k_2/n = \text{const}} \frac{k}{n} \left( 1 + \frac{n-k}{k_2 - k - 1} \right).
$$

**Remark 2.** Theorem 2 indicates that the asymptotic expectation of $\|P^T x\|_2^2/\|x\|_2^2$ is independent of the scale of the entries in $P_2$, or specifically $\sigma^2$, due to the cancelation by the inverse term in $P$. Hence it is not necessary to multiply a random matrix by a constant, e.g., $1/\sqrt{k_2}$ in (9).

**Remark 3.** Theorem 2 allows a much larger family of distributions for entries in $P_2$, than the symmetric distributions usually required for RIP. For example, many asymmetric distributions such as a chi-square distribution, a uniform distribution on $[0, 1]$ or a Bernoulli distribution on $\{0, 1\}$, all satisfy Eq. (35).

Theorem 2 provides us with a guideline on how to choose $k_2$ and $k$ to make sure at least the expectation of the ratio is no greater than one, i.e., $E(\|P^T x\|_2^2/\|x\|_2^2) \leq 1$. Interestingly, in the example in Fig. 2, it is purely by chance that we choose $k_2 = 2k$ that makes the expectation of the ratio exactly 1, and if to choose $k_2 > 2k$, we can be sure both expectations above will always be below 1, since $n > k_2$.

**Corollary 1.** If $\tilde{V}_k$ is any random orthonormal matrix, $P_2$ has i.i.d. entries drawn from a distribution with a finite variance, and $k_2 \geq 2k$, then

$$
E \left( \|\hat{A}_k - \tilde{A}_k\|_F^2 \right) \leq \|A - \tilde{A}_k\|_F^2.
$$

One important contribution of the theorem above is the relaxation from the often-required symmetric (zero-mean) distribution of entries in $P_2$, which serves as a necessary condition to guarantee $E(\|P^T x\|_2^2) = \|x\|_2^2$. The two most popular symmetric distributions used in the compressive sensing literature are the Gaussian and the Rademacher distributions, but both of which will inevitably draw negative entries that can be hard to implement in imaging devices. For example, the coded aperture is one of the popular techniques to implement random projection, but the resulting distribution of entries in the projection matrix can be far from symmetric. Figure 3 shows the distribution of entries in the system matrix of DD-CASSI [2], which is nonnegative and bears more resemblance to a chi-square or a Gamma distribution.

To prove that the approximation error is tightly distributed around the expectation, we first state a result by Davenport [22].

**Theorem 3** (Davenport, 2011). Suppose that $\Phi$ is an $m \times n$ matrix whose entries $\phi_{ij}$ are i.i.d. and drawn from a strictly sub-Gaussian distribution with $E(\phi_{ij}^2) = 1/M$. Let $Y = \Phi x$ for $x \in \mathbb{R}^n$. Then $\forall \epsilon > 0, \forall x \in \mathbb{R}^n$,

$$
E(\|Y\|_2^2) = \|x\|_2^2,
$$

and

$$
Pr(\|Y\|_2^2 - \|x\|_2^2 \geq \epsilon \|x\|_2^2) \leq 2 \exp\left(-\frac{M\epsilon^2}{k^*} \right),
$$

with $k^* = 2/\left(1 - \log(2)\right) \approx 6.52$. 

![Fig. 3. The distribution of entries in the system matrix of DD-CASSI due to the coded aperture.](image-url)
Let
\[ \Phi = \frac{n}{k} + \frac{1}{k} \frac{n-k-1}{2-k}. \] (39)

Then we know \( E(\phi_{ij}) = 0, E(\phi_{ij}^2) = 1/k, \) and \( E(\|\Phi^T x\|_2^2) = \|x\|_2^2. \) Since \( p_{ij} \) is normally distributed, it is strictly sub-Gaussian. Hence we have the following theorem.

**Theorem 4.** Given any vector \( x \in \mathbb{R}^n, \) \( \|P^T x\|_2^2 \) is tightly distributed around its expectation, and the probabilistic bound can be stated as,
\[ P_r \left( \|P^T x\|_2^2 \geq (1 + \epsilon) E(\|P^T x\|_2^2) \right) \leq 2 \exp \left( -\frac{k\epsilon^2}{2} \right). \] (40)

**C. Relaxation of the Symmetric Distribution Requirement of Entries in \( P_1 \)**

By now, we have only considered the distributions of entries in \( P_2, \) and in this subsection, we will discuss relaxing the required symmetric distribution of entries in \( P_1, \) and extend the bound in (11) to asymmetric distributions. We first state a theorem on the error bound of \( A_k \) in [7].

**Theorem 5** (Vempala, 2004). The Frobenius norm of \( \tilde{A}_k - A \) is bounded by,
\[ \|A - \tilde{A}_k\|_F^2 \leq \|A - A_k\|_F^2 + 2\epsilon\|A_k\|_F^2, \] (41)
where \( A_k \) is the truncated SVD and \( 0 \leq \epsilon \leq 1/2. \) satisfies RIP as
\[ (1 - \epsilon)\|x\|_2^2 \leq \|P_1^T x\|_2^2 \leq (1 + \epsilon)\|x\|_2^2, \] (42)
with a success probability of \( 1 - 2\exp[-(\epsilon^2 - \epsilon^3)k_1/4] \) for \( k_1 > 9 \log(n)/(\epsilon^2 - \epsilon^3). \)

In its proof, we find the following statement which imposes no requirement on the distribution of entries in \( P_1, \) i.e.,
\[ \|A - \tilde{A}_k\|_F^2 = \|A - A_k\|_F^2 + \left( \|A_k\|_F^2 - \sum_{i=1}^k \|A_{\tilde{v}_i}\|_2^2 \right). \] (43)

Hence, if \( \|A_k\|_F^2 \) is close to \( \sum_{i=1}^k \|A_{\tilde{v}_i}\|_2^2, \) we should have a good approximation of \( A \) by \( \tilde{A}_k. \) Here, we introduce a definition of compressible images, equivalent to what Donoho [23] defined as compressible signals, i.e. coefficients of a signal on a certain basis decay as a power law.

**Definition 1.** A matrix \( A \) is compressible if there exists \( s > 1/2, \) such that the singular values of \( A \) are bounded by a power law, i.e., \( \sigma_k/\sigma_1 \leq k^{-s}. \) And the compressibility of a matrix is defined as,
\[ s(A) = \inf_{s} \frac{\sigma_k}{\sigma_1} \leq k^{-s}. \] (44)

If we also have \( \sigma_1 = 1, \) we call it a normalized compressible matrix. Next, we look at a couple of properties of a compressible matrix.

**Proposition 5.** If matrix \( A \) is compressible, as \( k \to \infty, \) then \( \|A\|_F^2 = \text{tr}(A^T A) = \sum_{k=1}^\infty \sigma_k^2, \) is bounded.

**Proof.** Since the sum of a power sequence is bounded when the power is greater than 1, we know \( \sum_{k=1}^\infty \sigma_k^2 \leq \sum_{k=1}^\infty k^{-2s}, \) and hence is bounded.

The bounded sum of squared singular values serves as the key in our following proof of the low-rank approximation of \( A \) by \( A_k \) using random projection matrices with asymmetric distributions.

**Proposition 6.** The norm error of the \( k^{th} \) truncated SVD of a normalized compressible matrix is bounded by,
\[ \|A\|_F^2 - \|A_k\|_F^2 \leq \zeta(2s, k+1) - \zeta(2s, n+1), \] (45)
where \( \zeta(x, y) \) is the Hurwitz Zeta function defined as
\[ \zeta(x, y) = \sum_{i=0}^\infty (i + x)^y. \] (46)

**Proof.**
\[ \|A\|_F^2 = \|A_k\|_F^2 = \sum_{i=k+1}^n \sigma_i^2 = \sum_{i=1}^n i^{-2s}, \] (47)
and by the definition of the Hurwitz Zeta function, we have (45).

A slightly less accurate bound on the norm error of the truncated SVD is to use the integral of a power function, i.e.,
\[ \sum_{i=k+1}^n i^{-2s} \approx \int_{k+1}^n \frac{1}{x^{2s}} dx = \frac{1}{1 - 2s} [n^{1-2s} - (k+1)^{1-2s}]. \] (48)

In the numerical results section, we will show that the relationship between norm errors and \( s \) approximately follows a power law. Next, we state a deterministic upper-bound on \( \|A_k\|_F^2 - \sum_{i=1}^k \|A_{\tilde{v}_i}\|_2^2. \)

**Theorem 6.** When \( n \to \infty, k_1 \to \infty, \) and \( r = \sqrt{k_1/n}, \) for a random matrix \( P_1 = (p_{ij})_{n \times k_1}, \) where \( \text{var}(p_{ij}) = \sigma_i^2, \) we have
\[ \|A_k\|_F^2 - \sum_{i=1}^k \|A_{\tilde{v}_i}\|_2^2 \leq \frac{4r}{(1+r)^2} \|A\|_F^2 \|A_k\|_F^2. \] (49)

However, the bound in Theorem 6 is not too impressive. For example, if \( r = .1, \) a subsampling rate of 10, the relative norm error would have a rather large upper bound at .33. Now we seek a probabilistic bound which may not always be true, but is true with a large probability.

**Theorem 7.** If matrix \( A \) is compressible, and if random matrix \( P_1 \) has entries drawn from a finite-variance distribution, the following approximation bound stands with a large probability,
\[ \|A - \tilde{A}_k\|_F^2 \leq \|A - A_k\|_F^2 + 2\epsilon\|A_k\|_F^2. \] (50)

Except the compressibility requirement, Theorem 7 is the same as a result found [7]. To prove Theorem 7, we first need to look at a basic isometry property of random matrices.

**Proposition 7.** For a given vector \( x \in \mathbb{R}^n, \) and a random matrix \( P_1 = (p_{ij})_{n \times k_1}, \) where \( E(p_{ij}) = \mu_p, \text{var}(p_{ij}) = \sigma_p^2, \)
let \( y = \frac{1}{\sqrt{k_1}} P^T x \). As \( n \to \infty \), the expectation of \( \| P^T x \|^2 \) is:

\[
E(\|y\|^2) = (\mu_p^2 + \sigma_p^2)\|x\|^2 + 2\mu_p^2 \sum_{i_1,i_2=1 \atop i_1 \neq i_2}^{n} x_{i_1} x_{i_2}.
\]  

(51)

Since the proof only involves basic algebra, we leave it to readers to check. Clearly, if \( \mu_p = 0 \), the cross terms can be removed, but it is retained in the asymmetric case and hence the expectation depends on the individual entries of \( x \).

**Remark 4.** If \( x \in \mathbb{R}^n \) is also a normalized random vector with zero expectation and variance \( 1/n \), as \( n \to \infty \), by the central limit theorem, \( y_i = \sum_{j=1}^{n} p_{ij} x_j \) is normally distributed with a zero mean and its variance is \( \sigma^2_y = (\mu_p^2 + \sigma_p^2)k_1 \). Since \( y_i \) is strictly sub-Gaussian, we know [22]

\[
E(\|y\|^2) = k_1 \sigma^2_y = \mu_p^2 + \sigma_p^2,
\]

and for \( \epsilon > 0 \),

\[
Pr\left( \|y\|^2 - k_\epsilon \sigma^2_y \geq \epsilon k_1 \sigma^2_y \right) \leq 2 \exp\left( -\frac{k_1 \epsilon^2}{k^*} \right).
\]

(53)

We then have the following lemma.

**Lemma 8.** Let \( \{x_i \in \mathbb{R}^n, i = 1, \ldots, k\} \), be normalized random vectors with a zero expectation and variance \( 1/n \), and let \( \{\sigma_i, i = 1, \ldots, k\} \), be the singular values of a compressible matrix. As \( k, n \to \infty \), we have

\[
E(\|\sum_{i=1}^{k} \sigma_i^2 P^T x_i \|^2) = (\mu_p^2 + \sigma_p^2) \sum_{i=1}^{k} \sigma_i^2,
\]

(54)

and it is finite.

With this lemma, we are ready to prove Theorem 7, and readers can find details in the appendix. Now, we have extended the same bound proved by Vempala [7] to asymmetric distributions with the requirement of a compressible matrix.

**D. Randomized Singular Value Decomposition from DRP**

Here, we will discuss a combined algorithm of rSVD [24] and DRP, which will enable us to compute an approximate SVD using only the projected data. The rSVD algorithm finds the matrix \( Q \) with the same range as the projected data \( B_2 \), and a low-rank approximation of \( A \) is given by \( QQ^T A \), from which we have the following approximation,

\[
B_1 = P^T_1 A \approx P^T_1 QQ^T A.
\]

(55)

Multiplying both sides by the pseudo-inverse of \( P^T Q \), we have,

\[
(P^T_1 Q)^\dagger B_1 \approx Q^T A.
\]

(56)

The rSVD algorithm proceeds by computing an SVD of \( Q^T A \),

\[
Q^T A = \hat{U} \hat{\Sigma} \hat{V}^T,
\]

(57)

and

\[
\hat{U} = Q \hat{U} \approx U, \hat{V} \approx V,
\]

(58)

where \( U \) and \( V \) are from the SVD of \( A \) as \( A = U\Sigma V^T \). Rather than using \( Q^T A \), we can approximate it by \( (P^T_1 Q)^\dagger B_1 \) and proceed as the rest of the rSVD. This approach will allow us to collect only \( B_1 \) and \( B_2 \) without knowing \( A \) while still being able to compute its approximate SVD. Algorithm 2 summarizes this procedure.

**Algorithm 2: A double random projection rSVD algorithm**

**Input**: two projected images, \( B_1 = P^T_1 A \) and \( B_2 = AP_2 \), where \( A \in \mathbb{R}^{m \times n}, P_1 \in \mathbb{R}^{m \times k_1}, B_1 \in \mathbb{R}^{k_1 \times n}, P_2 \in \mathbb{R}^{n \times k_2}, B_2 \in \mathbb{R}^{n \times k_2} \).

**Output**: \( \hat{U}, \hat{\Sigma} \) and \( \hat{V} \), an approximate singular-value decomposition of \( A \).

1. Find the range \( Q \in \mathbb{R}^{m \times k} \) of \( B_2 \) by Alg. 4.2 in [24].
2. Compute an SVD of \( (P^T_1 Q)^\dagger B_1 \),

\[
(P^T_1 Q)^\dagger B_1 = \hat{U} \hat{\Sigma} \hat{V}.
\]

(59)

3. Return \( \hat{U} = Q \hat{U}, \hat{\Sigma}, \) and \( \hat{V} \).

**III. Numerical Experiments**

In this section, we present three examples: commonly-studied natural images, a real hyperspectral data set, and simulated compressible matrices with power-law decaying singular values.

**A. Test Example: Natural Images**

We choose four commonly-studied natural images, shown in Fig. 4 with names ‘Lena’, ‘Barbara’, ‘Mandrill’ and ‘Camera-man’ from left to right, to compare the DRP method with several compressed-sensing algorithms, i.e., MS-BCS-SPL [14]. MH-MS-BCS-SPL [15], RCoS [16], and the truncated SVD (TSVD). All algorithms are tested on the same platform, an 8-core 3.0GHz desktop computer. We choose five subsampling rates (subrates) from 0.1 to 0.5 as in [15]. In our method, both projection matrices, \( B_1 \) and \( B_2 \), are saved in the 8-bit unsigned integer format, and because of the square images, we choose \( k_1 \) to be same as \( k_2 \) in the set, \{26, 52, 76, 102, 128\}, corresponding to the five subrates from 0.1 to 0.5. In Table I, we compare all four algorithms in terms of peak signal-to-noise ratio (PSNR) and computation time. Clearly, we see the DRP runs much faster than all other algorithms, though its accuracy of using 2D wavelet transforms here is that the basis functions can capture both horizontal and vertical correlations at different spatial scales, while column- or row-based singular vectors used by DRP and TSVD can only capture one such correlation. We can see from Table I that the wavelet-based
methods can sometimes outperform TSVD, see e.g. the PSNR of the MS-BCS-SPL at subrate 0.5 for the Lena image. The DRP method is based on TSVD, and hence its best PSNR can only be equal to the PSNR of TSVD. In terms of computing time, the wavelet-based methods are understandably slower due to the wavelet transformations. It is worth noting that DRP also runs faster than TSVD.

In our next experiment with hyperspectral data, we will reshape a 3D hyperspectral datacube into a 2D matrix, with pixels on rows and spectral bands on columns. In our computations, this proves to be more efficient and accurate than 2D wavelet transforms, because the singular vectors spanning all pixels would capture correlations beyond horizontal, vertical, radial or any predefined directions as in wavelet transforms, but in all directions and scales.

B. Test Example II: Hyperspectral Images

We use a well-studied real hyperspectral dataset, Indian Pines dataset [25], collected by the AVIRIS sensor over a 25 × 6 mile portion of Northwest Tippecanoe County, Indiana, on June 12, 1992. The sensor has a spectral range of 0.45 to 2.5 microns over 220 bands, and the chosen subset consists of a 512 × 512 × 220 image cube of reflectance data stored as double precision. The data cube, is unfolded into a 262,144 × 220 matrix and projected along both dimensions. Figure 5(a) shows the image at the 200th spectral band, and Fig. 5(b) shows the sharply-decaying singular-value spectrum of A, indicating a low-rank structure of the matrix.

In this analysis, we use SNR to evaluate reconstruction quality rather than PSNR, because PSNR, by incorporating the signal’s peak value, effectively attempts a rudimentary contrast scaling to increase perceptual relevance of the measure. However, because hyperspectral imagery is primarily the subject of analysis (classification or anomaly detection) algorithms rather than human observation, perceptual quality is not as important. Consequently, our use of SNR is more meaningful for hyperspectral imagery and is consistent with the hyperspectral imagery literature, e.g., [8].

Here, we vary both $k_1 \in \{20, 40, 60, 80, 100, 200, 400, 600, 800, 1,000, 10,000\}$ and $k_2 \in \{2, 4, 6, \ldots, 20\}$, while setting $k = k_2 / 2$ to see their impacts on SNR. For each combination, because of the random projections involved, we run the DRP method ten times and plot the average SNR and computing time in Fig. 6. Interestingly, except the smallest $k_1 = 20$, all the rest choices of $k_1$ result in quite similar SNRs given any $k_2$, and the SNRs tend to level off after $k_2 = 10$, which is consistent with the singular value spectrum in Fig. 5(b). The computing time is largely related to $k_1$, due to the generation of the large matrix $P_1$, and hence in Fig. 6(b), we used $k_1$ as the x-axis. Because of the small sensitivity in $k_1$, in the following experiment, we fix $k_1$ at 100 while varying $k_2$ to test four different random matrices’ impact on SNR. Figure 7 shows very close performances by all four random matrices, i.e., uniform, Bernoulli, orthonormal and Gaussian, validating the claims that random matrices whose entries have an asymmetric distribution can perform as well as those with symmetric distributions.

To compare with other subsampling algorithms involving random projections, we compare DRP with CPPCA [8], which is particularly designed and tested on hyperspectral data. Here, we fix $k_1 = 100$, and vary $k_2$. Figure 8 shows the comparisons in SNR and in computing time. Clearly, DRP outperforms CPPCA in both categories. One important observation is that
A matrix $C$. Simulation on Compressible Matrices e.g., [26].

Drp can finish processing large amount of data in less than 2 seconds, suitable for onboard processing, where batches of scanned hyperspectral data can be processed in real time, see, e.g., [26].

C. Simulation on Compressible Matrices

We examine the low-rank approximation of a compressible matrix $A$ by $A_k = A\hat{V}_k\hat{V}_k^T$ in (10). From [7], we know the symmetric distribution is a necessary condition for the success of the DRP. And indeed, our empirical experience shows that in some cases, the approximation can be rather poor with asymmetric distributions. However, here we will show empirically how asymmetric distributions can perform quite well for compressible images.

Given a power $s$, we simulate a compressible matrix by first simulating the orthonormal singular vectors in $U$ and $V$, and then the singular values in $\Sigma$, whose diagonal entries obey the power law, i.e.,

$$A = U\Sigma V^T , \sigma_k = ck^{-s}.$$  \hfill (60)

We then project $A$ with four different random matrices, i.e., normal, uniform distribution in $[0, 1]$ and Bernoulli, and perform a SVD on the projected image, $B = P^TA$, before computing $A_k$. This time we compute the PSNR between the optimal approximation $A_k$ and $\hat{A}_k$. The simulation is repeated for 1,000 times and the mean PSNRs are plotted against the power $s$ in Fig 9, which shows a clear linear relationship between PSNR and power $s$ for all three random-projection matrices, and we observe major differences between them.

For each $s$, we repeat the simulation 1,000 times and compare PSNRs across three random projections, i.e., normal, Bernoulli and uniform. Figure 10 shows the histograms of the percentage of times either Bernoulli or uniform projection matrices outperforms the normal projection matrices among all variations of $s$. If the histogram is centered at $.5$, this indicates that the two compared projections are equivalent. However, we do see the medians are both slightly less than $.5$, indicating the normal projection matrices perform slightly better than the other two types, but not by much. Hence, we have demonstrated that even asymmetric distributions such as uniform or Bernoulli also work well on compressible images.

More simulations by varying both $k$ and $s$ indicate that the error norms approximately follow a power rule, as anticipated by Proposition 6 and (48), i.e.,

$$\|A_k - A\hat{V}_k\hat{V}_k^T\|_F \approx c_1k^{-(s+c_2)},$$  \hfill (61)

where $c_1$ and $c_2$ are two constants. Hence it is not hard to see that if the singular-value spectrum of an image is upper-bounded by a power law, i.e.,

$$\sigma_k \leq Ck^{-s},$$  \hfill (62)

then the error norms from the low-rank approximation would also be upper-bounded by the power law above, a result similar to Candes and Tao [11], namely

$$\|A_k - A\hat{V}_k\hat{V}_k^T\|_F \leq c_1k^{-(s+c_2)}.$$  \hfill (63)
D. Approximated Singular Vectors and Singular Values

We numerically examine the differences between singular vectors computed by our Alg. 2 and those by Alg. 4.3 in [24]. We simulate a $5,000 \times 1,000$ random matrix with singular values satisfying $\sigma_k/\sigma_1 = k^{-2}$, and set $k_1 = 2,000, k_2 = 400$, and $k = 100$. Fixing matrix $A$, we draw random matrices $P_1$ and $P_2$ for $B_1$ and $B_2$, from which we run Alg. 2 to compute $\hat{U}, \hat{V}$ and $\hat{\Sigma}$. The same process is repeated 1,000 times. At each time, the same $P_2$ is used as $\Omega$ for Alg. 4.3 in [24]. The estimated singular vectors are then compared with the true ones using their cosine angles, while the estimated singular values are compared with the true ones using their relative absolute differences. Figure 11 shows the histograms of cosine angles between $u_i$ and $\hat{u}_i$ and between $v_i$ and $\hat{v}_i$, and the histogram of the relative differences between $\sigma_i$ and $\hat{\sigma}_i$, of all 1,000 simulations. For the first 40 singular vectors, Alg. 2 gives good approximations while the approximated singular values are also quite close.
IV. Conclusions

In summary, we have proposed a simple and non-iterative algorithm for reconstructing a true image from two randomly projected images, one along the column dimension (9) and the other along the row dimension (13). Empirical results show the good quality of reconstruction in terms of the PSNR, even with random projection matrices with entries drawn from asymmetric distributions. For compressible images, we have shown empirically that the low-rank approximation by $\tilde{A}_k$ given in (10) is not restricted to symmetric distributions, and that the error norms of reconstruction are bounded by a power law, if the singular-value-spectrum of the true image is bounded by a power law. We have also proposed an approximate singular-value decomposition (SVD) method using two randomly projected matrices and numerical results demonstrates its close proximity to the randomized SVD method in [24].

The main drawback of the proposed framework is that two sensors are required, one to implement Eq. (9) and another to implement Eq. (13). This implies a doubling of the hardware required for a sensor platform and may be difficult to implement with hyperspectral imagery. Specifically, Eq. (13) will require a sensor that is capable of imaging the full spatial extent of the scene, e.g., using a single-pixel hyperspectral sensor [27]. In remote-sensing applications, hyperspectral imagery is often acquired with whiskbroom or pushbroom sensors that capitalize on the sensor-platform motion, e.g., airborne or satellite-borne sensors, it may not be possible to image the full spatial extent all at once. While Eq. (9) can possibly support whiskbroom/pushbroom acquisition, Eq. (13) cannot. However, the proposed DRP method can be applied onboard computationally, rather than at the sensor level, after a certain amount of data is acquired by whiskbroom/pushbroom sensors, as illustrated in [26].

REFERENCES

APPENDIX A

PROOF OF THEOREM 2

Proof. First note that $\|P_k x\|_2 = \|P_k^T y\|_2$, and hence we need only consider the properties of $P_k$. Denote entries of $P_k$ as $p_{ij}$. To prove Eq. (35) it is equivalent to prove that

$$
\operatorname{var}(p_{ij}) = \frac{1}{n} \left(1 + \frac{n - k}{k_2 - k - 1}\right),
$$

(64)

if $E(p_{ij}) = 0$ and all $p_{ij}$ are i.i.d. We leave this to readers to check since it only involves some basic algebra.

Let $X = P_k^T \hat{V}_k$, then entries of $X$ are normally distributed with mean 0 and variance $\mu^2 + \sigma^2$. Then $X^T X$ has a Wishart distribution, i.e.,

$$
X^T X \sim W_k(\sigma^2 I_k, k_2),
$$

(65)

and its inverse has an inverse Wishart distribution,

$$
(X^T X)^{-1} \sim W_k^{-1}(\sigma^2, (k_2 - k - 1) I_k, k_2).
$$

(66)

Let $Y = X^T = (y_{ij})$ then we know, [28], [29], that

$$
\operatorname{var}(y_{ij}) = \frac{1}{\sigma^4 (k_2 - k - 1)}.
$$

(67)

Since $P_k = \hat{V}_k + P_k^-$, and $\hat{V}_k^T = (v_{ij})$, the variance of $p_{ij}$ can be split into two parts. Because $\operatorname{var}(v_{ij}) = \frac{1}{n}$, we only need the variance of $P_k^-$, which can be further split into two parts, i.e.,

$$
P_k^- = (I_n - \hat{V}_k \hat{V}_k^T) P_2 Y.
$$

(68)

Let $R = (I_n - \hat{V}_k \hat{V}_k^T) P_2$, and notice $V_k^- = I_n - \hat{V}_k \hat{V}_k^T = (v_{ij})$ is independent of $P_2$, i.e.,

$$
\operatorname{var}(r_{ij}) = \operatorname{var}(v_{ij}) \operatorname{var}(\tilde{p}_{ij}) = \sigma^2 \frac{n - k}{n}.
$$

(69)

Further $R$ is asymptotically independent of $Y$, and both have a zero mean. Hence, we know that

$$
\operatorname{var}(p_{ij}) = \operatorname{var}(r_{ij}) \operatorname{var}(y_{ij}) = \sigma^2 \frac{n - k}{n} = \frac{1}{n} \frac{\sigma^2}{\sigma^4(k_2 - k - 1)} = \frac{n - k}{n k_2 - k - 1}.
$$

(70)

Putting together $\operatorname{var}(v_{ij})$ and $\operatorname{var}(p_{ij})$, we have established Eq. (35). \qed
Appendix B
Proof of Theorem 6
Proof. Let \(\{\lambda_i|i=1,\ldots,k\}\) be the singular values of \(B_1\) and \(\{\sigma_i|i=1,\ldots,k\}\) be the singular values of \(A\). By the Marcenko-Pastur distribution, we have the following inequalities,
\[
\sum_{i=1}^{k} \lambda_i^2 = \sum_{i=1}^{k} \tilde{v}_i^T A^T P_i P_i^T A \tilde{v}_i \\
\leq n \sigma^2 (1 + r)^2 \sum_{i=1}^{k} \|A \tilde{v}_i\|_2^2,
\]
(71)
and
\[
\sum_{i=1}^{k} \lambda_i^2 \geq \sum_{i=1}^{k} v_i^T A^T P_i P_i^T A v_i \\
= \sum_{i=1}^{k} \sigma^2 u_i^T P_i P_i^T u_i \\
\geq n \sigma^2 (1 - r)^2 \|A_k\|_F^2.
\]
(72)
Putting the two inequalities above together, we have,
\[
\sum_{i=1}^{k} \|A \tilde{v}_i\|_2^2 \geq \frac{(1 - r)^2}{(1 + r)^2} \|A_k\|_F^2,
\]
(73)
and thus (49) holds.

Appendix C
Proof of Theorem 7
Proof. We follow the same line of proof as in [7]. Let \(\{\lambda_i|i=1,\ldots,k\}\) be the singular values of \(B_1\) and \(\{\sigma_i|i=1,\ldots,k\}\) be the singular values of \(A\). And we assume a normal prior distribution on the entries of \(v_i\) and \(u_i\) with mean zero and variance \(1/n\). Let \(x_i = A v_i\) and \(y_i = P_i^T x_i\). Denote the \(j^{th}\) row vector of matrix \(A\) as \(a_j\). It is not hard to check that
\[
E(x_{ij}) = 0, \ E(x_{ij}^2) = \frac{1}{n} \|a_j\|_2^2, \text{ and } E(\|x_i\|_2^2) = \frac{1}{n} \|A\|_F^2.
\]
(74)
Since \(A\) is compressible, we know that \(n \to \infty, \|A\|_F^2\) remains finite, and so does \(\|a_j\|_2^2\). Similarly for \(y_i\), we have
\[
E(y_{ij}) = 0, \ E(y_{ij}^2) = \frac{1}{n} (\mu_p^2 + \sigma_p^2) \|a_j\|_2^2,
\]
(75)
and
\[
E(\|y_i\|_2^2) = \frac{1}{n} (\mu_p^2 + \sigma_p^2) \|A\|_F^2.
\]
(76)
To ensure \(\|y_i\|_2^2\) is highly concentrated around its expectation, by Corollary 1 in [22], we know that \(\{y_{ij}|j=1,\ldots,n\}\) needs to be i.i.d., strictly sub-Gaussian (SSub), and with the equal variance \(c^2\). Here, we slightly relax the equal variance condition, by assuming \(y_{ij} \sim SSub(c^2)\). The same proof of Lemma 1 in [22] still follows through, except by replacing \(c^2\) with
\[
c_m^2 = \max_j c_j^2 = \frac{1}{n} (\mu_p^2 + \sigma_p^2) \max_j \|a_j\|_2^2,
\]
(77)
and \(k^*\) [22, Eq. (19)] would be changed accordingly. Thus (53) still stands and we have the following two probabilistic bounds:
\[
\sum_{i=1}^{k} \lambda_i^2 = \sum_{i=1}^{k} \|y_i\|_2^2 \leq (1 + \epsilon) \frac{1}{n} (\mu_p^2 + \sigma_p^2) \sum_{i=1}^{k} \|A \tilde{v}_i\|_2^2, \text{ w.p.},
\]
(78)
and
\[
\sum_{i=1}^{k} \lambda_i^2 \geq \sum_{i=1}^{k} \|P_i^T A v_i\|_2^2 = \sum_{i=1}^{k} \sigma^2 \|P_i^T u_i\|_2^2 \\
\geq (1 - \epsilon) \frac{1}{n} (\mu_p^2 + \sigma_p^2) \|A_k\|_F^2, \text{ w.p.}
\]
(79)
Hence, we have,
\[
(1 + \epsilon) \sum_{i=1}^{k} \|A \tilde{v}_i\|_2^2 \geq (1 - \epsilon) \|A_k\|_F^2, \text{ w.p.}
\]
(80)
\[\square\]

Acknowledgments
Research by R. Plemmons and Q. Zhang is supported by the U.S. Air Force Office of Scientific Research (AFOSR), under Grant FA9550-11-1-0194. We thank the referees for their very useful input on this paper.

Qiang Zhang is a quantitative analyst with BB&T Bank and previously he is a Senior Research Associate with Department of Biostatistical Sciences in Wake Forest School of Medicine. He has a Bachelor’s and a Master’s degree in Mechanical Engineering from Tsinghua University, a Master’s degree in Architecture from University of California, Berkeley, and a Master’s degree in Mathematics from Wake Forest University. His research interests include statistical modeling, financial data analysis, remote sensing and image processing. Email: qizhang@wakehealth.edu.

Robert J. Plemmons (Z. Smith Reynolds Professor, Departments of Mathematics and Computer Science, Wake Forest University). Previous positions: Professor, Departments of Mathematics and Computer Science, North Carolina State University (1981-91); Associate Professor and Professor, Departments of Mathematics and Computer Science, University of Tennessee (1967-81); Research Scientist, National Security Agency (1965-67); Visiting Professor: Duke University (1998), University of Minnesota (1992), University of Illinois (1986), and Stanford University (1978); Education: Ph.D. in Applied Mathematics, Auburn University; B.S. in Mathematics and Physics, Wake Forest University. Professional baseball player, Baltimore Orioles Farm System (1961-65). Current research interests include scientific computation, specifically: ill-posed inverse problems, image processing and hyperspectral methods in remote sensing, with funding from the Air Force Office of Scientific Research. Email: plemmons@wfu.edu, web: http://users.wfu.edu/plemmons/