Compressive-Projection Principal Component Analysis

James E. Fowler, *Senior Member, IEEE*

***Abstract*—Principal component analysis (PCA) is often central to dimensionality reduction and compression in many applications, yet its data-dependent nature as a transform computed via expensive eigendecomposition often hinders its use in severely resource-constrained settings such as satellite-borne sensors. A process is presented that effectively shifts the computational burden of PCA from the resource-constrained encoder to a presumably more capable base-station decoder. The proposed approach, compressive-projection PCA (CPPCA), is driven by projections at the sensor onto lower-dimensional subspaces chosen at random, while the CPPCA decoder, given only these random projections, recovers not only the coefficients associated with the PCA transform, but also an approximation to the PCA transform basis itself. An analysis is presented that extends existing Rayleigh–Ritz theory to the special case of highly eccentric distributions; this analysis in turn motivates a reconstruction process at the CPPCA decoder that consists of a novel eigenvector reconstruction based on a convex-set optimization driven by Ritz vectors within the projected subspaces. As such, CPPCA constitutes a fundamental departure from traditional PCA in that it permits its excellent dimensionality-reduction and compression performance to be realized in an light-encoder/heavy-decoder system architecture. In experimental results, CPPCA outperforms a multiple-vector variant of compressed sensing for the reconstruction of hyperspectral data.**

***Index Terms*—Hyperspectral data, principal component analysis (PCA), random projections, Rayleigh–Ritz theory.**

I. Introduction

PRINCIPAL component analysis (PCA) has long played a central role in dimensionality reduction and compression of multidimensional datasets in myriads of signal-processing applications. However, PCA—also known as the Karhunen–Loève transform—is a data-dependent transform arising from the eigendecomposition of the covariance matrix of the signal in question. Thus, in traditional compression and communication applications using PCA, the encoder must calculate the PCA transform before it can be applied to the data. Unfortunately, the computational burden that this process entails may well exceed the limited capabilities of many encoding platforms.

For example, there has traditionally been substantial interest in applying PCA for the decorrelation and dimensionality reduction of spectral bands in hyperspectral imagery—PCA has been employed in hyperspectral compression (e.g., [1] and [2]) as well as prior to various processing such as enhancement/denoising and classification (e.g., [3]–[5]). Yet, many hyperspectral sensing platforms are often severely resource-constrained, e.g., satellite-borne devices. For such hyperspectral sensors, as well as similar sensors in other application areas, it would be greatly beneficial if PCA-based dimensionality reduction and compression could be accomplished without the heavy encoder-side burden entailed by PCA.

In this paper, we present a process that effectively shifts the computational burden of PCA from the resource-constrained encoder to the decoder which presumably resides on a significantly more powerful "base-station" system. Our approach, compressive-projection PCA (CPPCA), is driven by projections at the signal sensor onto lower dimensional subspaces chosen at random. The CPPCA decoder, given only these random projections, recovers not only the coefficients associated with the PCA transform, but also an approximation to the PCA transform basis itself.

Coupling random projections at the sensor with simple scalar quantization and entropy coding yields a lightweight CPPCA encoder. On the other hand, the bulk of the computation resides at the CPPCA decoder which consists of a novel eigenvector-reconstruction process based on a projections-onto-convex-sets (POCS) optimization. CPPCA constitutes a fundamental departure from the traditional use of PCA in that it permits the excellent dimensionality-reduction and compression performance of PCA to be realized in an light-encoder/heavy-decoder system architecture. We know of no other approach to "decoder-side" PCA that accomplishes anything similar.

The primary contributions of this paper are twofold. As the first contribution, we provide an extensive analysis to justify our CPPCA approach. Specifically, we invoke Rayleigh–Ritz theory [6] which defines Ritz vectors in projected subspaces. Our analysis provides insight into the relation between these Ritz vectors and orthonormal projections of eigenvectors and argues that the former can be used to approximate the latter. This analysis then leads to the second contribution, the CPPCA reconstruction algorithm itself, wherein we use Ritz vectors to drive a POCS optimization to recover approximations to eigenvectors of the PCA transform directly from the projected subspaces.

We note that, in its reliance on encoder-side random projections, CPPCA bares some similarity to the emerging mathematical paradigm of compressed sensing[1] (CS) (e.g., [7]–[11]).

Manuscript received June 18, 2008; revised May 25, 2009. First published June 10, 2009; current version published September 10, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Kai-Kuang Ma.

The author is with the Department of Electrical and Computer Engineering, Geosystems Research Institute, Mississippi State University, Mississippi State, MS 39762 USA (e-mail: fowler@ece.msstate.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2009.2025089

[1] Also known as *compressive sampling*.

1057-7149/$26.00 © 2009 IEEE

Although both CS and CPPCA consist of lightweight projection-based encoding, their decoder-side reconstructions differ significantly—CS assumes a fixed basis ensuring sparsity while CPPCA determines a data-specific PCA basis directly from the random projections. Experimental results presented below reveal that CPPCA achieves reconstruction performance substantially superior to that of a multiple-vector CS variant when applied to hyperspectral data.

Below, we detail our CPPCA approach. We start with Section II which surveys relevant background surrounding PCA and Rayleigh–Ritz theory. The first main contribution of the paper, our central analysis that argues that Ritz vectors can approximate projected eigenvectors, follows as Section III. Section IV in turn describes the CPPCA algorithm, the second main contribution of the paper. Section V briefly relates CPPCA to other work, including CS, while experimental results on hyperspectral data are presented in Section VI. Finally, we make some concluding remarks in Section VII. We note that preliminary descriptions of the CPPCA algorithm and analysis appeared in [12] and [13], respectively. We also note that source code to reproduce all the experimental results to follow can be found at http://www.ece.msstate.edu/~fowler/CPPCA.

II. Background

A. PCA and Projections

Consider a dataset of M vectors $\mathbf{X} = [\mathbf{x}_1 \ \cdots \ \mathbf{x}_M]$, where each $\mathbf{x}_m \in \mathbb{R}^N$; we assume that the vectors have zero mean. The covariance matrix of $\mathbf{X}$ is $\mathbf{\Sigma} = \mathbf{X}\mathbf{X}^T/M$. For a given vector, $\mathbf{x}_m$, in $\mathbf{X}$, the PCA of $\mathbf{x}_m$ results from the application of a linear transform, $\check{\mathbf{x}}_m = \mathbf{W}^T\mathbf{x}_m$, where $N \times N$ transform matrix $\mathbf{W}$ emanates from the eigendecomposition of $\mathbf{\Sigma}$; i.e.,

$$\mathbf{\Sigma} = \mathbf{W}\mathbf{\Lambda}\mathbf{W}^T \tag{1}$$

where $\mathbf{W}$ contains the N unit eigenvectors of $\mathbf{\Sigma}$ column-wise.

In the sequel, we will be interested in the effect that projection onto a subspace has on PCA. Specifically, suppose we have K orthonormal vectors $\mathbf{p}_k$ that form the basis of K-dimensional subspace $\mathcal{P}$ such that $\mathbf{P} = [\mathbf{p}_1 \ \cdots \ \mathbf{p}_K]$ provides an orthogonal projection onto $\mathcal{P}$. Using the terminology of [6], we call $\mathbf{P}$ an *orthonormal* $N \times K$ matrix. Then, the orthogonal projection of $\mathbf{x}_m$ onto $\mathcal{P}$ is $\mathbf{y}_m = \mathbf{P}\mathbf{P}^T\mathbf{x}_m$; expressed with respect to the basis $\{\mathbf{p}_k\}$, we have $\widetilde{\mathbf{y}}_m = \mathbf{P}^T\mathbf{x}_m$, such that $\mathbf{y}_m = \mathbf{P}\widetilde{\mathbf{y}}_m$. The projected vectors $\widetilde{\mathbf{Y}} = [\widetilde{\mathbf{y}}_1 \ \cdots \ \widetilde{\mathbf{y}}_M]$ then have covariance

$$\widetilde{\mathbf{\Sigma}} = \widetilde{\mathbf{Y}}\widetilde{\mathbf{Y}}^T/M = \mathbf{P}^T\mathbf{X}\mathbf{X}^T\mathbf{P}/M = \mathbf{P}^T\mathbf{\Sigma}\mathbf{P}. \tag{2}$$

In the next section, we consider the relation between the eigenvectors of $\mathbf{\Sigma}$ and those of $\widetilde{\mathbf{\Sigma}}$.

B. Rayleigh–Ritz Procedure

In the classic problem of the calculation of eigenvalues and eigenvectors of a matrix, a number of solutions proceed by finding a sequence of subspaces containing approximations to the eigenvectors that increase in accuracy with each subsequent subspace [14]. Consequently, a crucial issue in any of these solutions methods is the production of approximations to eigenvectors within a given subspace. Perhaps the best-known method for such subspace approximation is the Rayleigh–Ritz procedure which we discuss briefly below. We note that our focus is PCA; consequently, the matrix in question is the real, symmetric covariance matrix $\mathbf{\Sigma}$. We, therefore, follow the treatment of the symmetric eigenvalue problem given by Parlett [6].

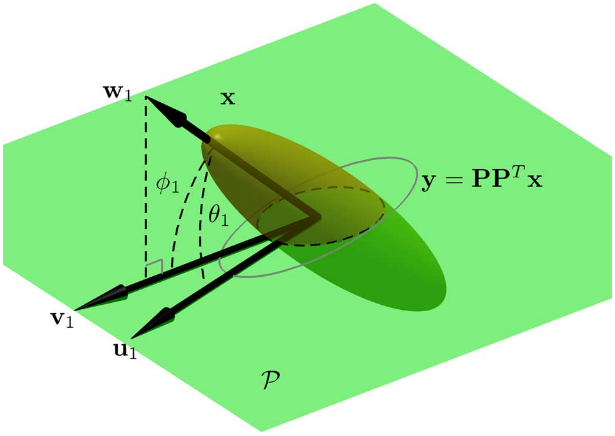

Fig. 1. Data distribution of $\mathbf{x}$ in $\mathbb{R}^3$ is projected onto 2-D subspace $\mathcal{P}$ as $\mathbf{y}$; the first Ritz vector, $\mathbf{u}_1$, lies close to the normalized projection, $\mathbf{v}_1$, onto $\mathcal{P}$ of the first eigenvector, $\mathbf{w}_1$, of $\mathbf{x}$.

Rayleigh–Ritz theory [6] describes the relation between the eigenvectors of $\mathbf{\Sigma}$ and those of $\widetilde{\mathbf{\Sigma}}$ as given by (2). Assume covariance matrix $\mathbf{\Sigma}$ has spectrum $\lambda(\mathbf{\Sigma}) = \{\lambda_1(\mathbf{\Sigma}), \ldots, \lambda_N(\mathbf{\Sigma})\}$, where the eigenvalues satisfy $\lambda_1(\mathbf{\Sigma}) \geq \cdots \geq \lambda_N(\mathbf{\Sigma})$. The corresponding unit eigenvectors are $\mathbf{w}_n$. Thus, $\mathbf{\Sigma} = \mathbf{W}\mathbf{\Lambda}\mathbf{W}^T$, where $\mathbf{W} = [\mathbf{w}_1 \ \cdots \ \mathbf{w}_N]$, $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1(\mathbf{\Sigma}), \ldots, \lambda_N(\mathbf{\Sigma}))$, and $\|\mathbf{w}_n\|_2 = 1$. The eigendecomposition of $\widetilde{\mathbf{\Sigma}} = \mathbf{P}^T\mathbf{\Sigma}\mathbf{P}$ is $\widetilde{\mathbf{\Sigma}} = \widetilde{\mathbf{U}}\widetilde{\mathbf{\Lambda}}\widetilde{\mathbf{U}}^T$, where $\widetilde{\mathbf{U}} = [\widetilde{\mathbf{u}}_1 \cdots \widetilde{\mathbf{u}}_K]$, $\widetilde{\mathbf{\Lambda}} = \operatorname{diag}(\lambda_1(\widetilde{\mathbf{\Sigma}}), \ldots, \lambda_K(\widetilde{\mathbf{\Sigma}}))$, $\|\widetilde{\mathbf{u}}_k\|_2 = 1$, and $\lambda_1(\widetilde{\mathbf{\Sigma}}) \geq \cdots \geq \lambda_K(\widetilde{\mathbf{\Sigma}})$. The K eigenvalues $\lambda_k(\widetilde{\mathbf{\Sigma}})$ are called *Ritz values*; additionally, there are K vectors, known as *Ritz vectors*, defined as

$$\mathbf{u}_k = \mathbf{P}\widetilde{\mathbf{u}}_k, \quad 1 \leq k \leq K \tag{3}$$

where $\widetilde{\mathbf{u}}_k$ are the eigenvectors of $\widetilde{\mathbf{\Sigma}}$. Note that $\|\mathbf{u}_k\|_2 = 1$. Finally, we define *normalized projection* $\mathbf{v}_n$ as the orthogonal projection of $\mathbf{w}_n$ onto $\mathcal{P}$, normalized to unit length; i.e.,

$$\mathbf{v}_n = \frac{\mathbf{P}\mathbf{P}^T\mathbf{w}_n}{\|\mathbf{P}\mathbf{P}^T\mathbf{w}_n\|_2}. \tag{4}$$

These vectors are illustrated for an example distribution in the simple case of $N = 3$ and $K = 2$ in Fig. 1.

The Rayleigh–Ritz procedure postulates that the pairs $(\lambda_k(\widetilde{\mathbf{\Sigma}}), \mathbf{u}_k)$, $1 \leq k \leq K$, are a reasonable approximation to some K eigenpairs $(\lambda_n(\mathbf{\Sigma}), \mathbf{w}_n)$ of the original matrix $\mathbf{\Sigma}$ as $K \rightarrow N$, and the "gap theorem" (Theorem 11.7.1 of [6]) bounds the difference between the Ritz vectors and the eigenvectors. That is, suppose that, for a given Ritz vector $\mathbf{u}_k$, the eigenpair $(\lambda_{n^*}(\mathbf{\Sigma}), \mathbf{w}_{n^*})$ has the eigenvalue closest to the Rayleigh quotient of $\mathbf{u}_k$; i.e.,

$$\lambda_{n^*}(\mathbf{\Sigma}) = \arg \min_{\lambda \in \lambda(\mathbf{\Sigma})} |\lambda - \rho(\mathbf{u}_k)| \tag{5}$$

where the Rayleigh quotient is $\rho(\mathbf{u}_k) = \mathbf{u}_k^T \mathbf{\Sigma} \mathbf{u}_k$. The gap theorem holds that, if the eigenvalues are sufficiently well-separated from one another, then the angle between $\mathbf{u}_k$ and $\mathbf{w}_{n^*}$ can be bounded. Specifically, suppose that $\rho(\mathbf{u}_k)$ is separated from the other eigenvalues by at least γ_k; i.e.,

$$\gamma_k = \min_{\substack{\lambda \in \lambda(\mathbf{\Sigma}) \\ \lambda \neq \lambda_{n^*}(\mathbf{\Sigma})}} |\lambda - \rho(\mathbf{u}_k)| . \tag{6}$$

Then

$$|\sin \phi_{n^*}| \leq |\sin \theta_k| \leq \frac{\|\mathbf{\Sigma}\mathbf{u}_k - \mathbf{u}_k \rho(\mathbf{u}_k)\|_2}{\gamma_k} \tag{7}$$

where θ_k is the angle between $\mathbf{u}_k$ and $\mathbf{w}_{n^*}$, and ϕ_{n^*} is the angle between $\mathbf{w}_{n^*}$ and its corresponding normalized projection $\mathbf{v}_{n^*}$ (see Fig. 1). The right inequality in (7) is the gap theorem from [6]; the left inequality arises simply from the fact that the orthogonal projection is the closest vector in $\mathcal{P}$ to $\mathbf{w}_{n^*}$ (i.e., has the smallest angle). We note that, contrary to what one might expect to be the case, the Ritz vectors do not generally align with the orthogonal projections of any of the eigenvectors [6]; i.e., $\mathbf{u}_k \neq \mathbf{v}_n$ and $\theta_k \neq \phi_n$ in general.

We now depart from traditional Rayleigh–Ritz theory in that we hold K fixed rather than have it increase towards N. However, we observe that a large γ_k for a particular Ritz vector will tend to drive the upper bound close to the lower bound in (7), suggesting that θ_k is close to ϕ_{n^*} in this case. In other words, the Ritz vector $\mathbf{u}_k$ is a close approximation to a normalized projection $\mathbf{v}_{n^*}$ into the subspace $\mathcal{P}$ of some eigenvector $\mathbf{w}_{n^*}$ if γ_k is large. Unfortunately, to the best of our knowledge, existing theory does not tell us which $\mathbf{w}_{n^*}$ out of the N eigenvectors possesses the close normalized projection. Nor does the theory guarantee that γ_k is large for all (or even any) of the K Ritz vectors.

CPPCA is built on the idea that, if subspace $\mathcal{P}$ is chosen randomly, and the distribution of the vectors in $\mathbf{X}$ is highly eccentric in that eigenvalue $\lambda_k(\mathbf{\Sigma})$ is sufficiently separated in value with respect to the other eigenvalues, then it is likely that its corresponding normalized projection, $\mathbf{v}_k$, will be quite close to the Ritz vector, $\mathbf{u}_k$, corresponding to the Ritz value $\lambda_k(\widetilde{\mathbf{\Sigma}})$. Of course, it is possible for $\mathcal{P}$ to be oriented such that this is not the case (i.e., if $\mathbf{w}_k$ happens to be close to being orthogonal to $\mathcal{P}$); however, for a randomly chosen $\mathcal{P}$ and highly eccentric $\mathbf{X}$ distribution, such an occurrence is rare. The analysis in the next section verifies the validity of this conjecture.

III. Analytical Motivations

In a general setting, the normalized projections, $\mathbf{v}_n$, of the eigenvectors, $\mathbf{w}_n$, of $\mathbf{\Sigma}$ do not typically align with any of the Ritz vectors, $\mathbf{u}_k$, of $\widetilde{\mathbf{\Sigma}} = \mathbf{P}^T\mathbf{\Sigma}\mathbf{P}$. However, under the imposition of additional structure—specifically, an eccentric $\mathbf{X}$ distribution that results in highly separated eigenvalues for $\mathbf{\Sigma}$—existing Rayleigh–Ritz theory can be substantially enhanced. In fact, in this section, we show that, for a single-spike covariance model wherein one eigenvalue is large while the rest are small and identical, we have that the first Ritz vector is exactly identical to the first normalized projection; i.e., $\mathbf{u}_1 = \mathbf{v}_1$. This result is then extended to show that, under a more general, but still eccentric, covariance (one eigenvalue large, the rest small but not necessarily identical), the angle of deviation between $\mathbf{u}_1$ and $\mathbf{v}_1$ is bounded, and, with $\mathbf{P}$ selected randomly, this bound is expected to be small.

The remainder of this section is as follows. We first focus on the first eigenvector corresponding to the largest eigenvector in Section III-A. We then consider the remaining eigenvectors in Section III-B. Finally, we discuss the practical relevance of the eccentricity assumptions that underlie the analysis in Section III-C. Throughout, only the main results are presented while detailed proofs are relegated to the appendices.

A. First Eigenvector

Theorem 1: Let $\mathbf{\Sigma}$ be a single-spike covariance matrix; that is, $\mathbf{\Sigma}$ is an $N \times N$ symmetric, positive-definite matrix with spectrum $\lambda_1(\mathbf{\Sigma}) > \lambda_2(\mathbf{\Sigma}) = \lambda_3(\mathbf{\Sigma}) = \ldots = \lambda_N(\mathbf{\Sigma}) > 0$. Let $\mathbf{w}_1$ be the first eigenvector of $\mathbf{\Sigma}$ associated with the first eigenvalue, $\lambda_1(\mathbf{\Sigma})$. For orthonormal $N \times K$ matrix $\mathbf{P}$ such that $\mathbf{P}^T\mathbf{w}_1 \neq \mathbf{0}$, the first eigenvector of $\widetilde{\mathbf{\Sigma}} = \mathbf{P}^T\mathbf{\Sigma}\mathbf{P}$ is $\tilde{\mathbf{u}}_1 = \mathbf{P}^T\mathbf{w}_1 / \|\mathbf{P}^T\mathbf{w}_1\|_2$. The corresponding first eigenvalue is

$$\lambda_1(\widetilde{\mathbf{\Sigma}}) = \delta_1 \|\mathbf{P}^T\mathbf{w}_1\|_2^2 + \lambda_N(\mathbf{\Sigma}) \tag{8}$$

where $\delta_1 = \lambda_1(\mathbf{\Sigma}) - \lambda_N(\mathbf{\Sigma})$.

Proof: See Appendix II. ■

Theorem 1 shows that, under the extreme structure of perfect eccentricity (only a single large eigenvalue in the covariance), Rayleigh–Ritz theory can be substantially strengthened—in this case, we are guaranteed perfect alignment between the first Ritz vector, $\mathbf{u}_1 = \mathbf{P}\tilde{\mathbf{u}}_1$, and the first normalized projection, $\mathbf{v}_1$, except when the projection $\mathbf{P}$ happens to be orthogonal to the first eigenvector $\mathbf{w}_1$. We note that, if we are choosing $\mathbf{P}$ at random, then this exceptional situation will almost never occur.

We now consider the case of a more general covariance matrix that is eccentric but not perfectly so; that is, eigenvalues $\lambda_2(\mathbf{\Sigma}), \ldots, \lambda_N(\mathbf{\Sigma})$ are small but not necessarily identical to one another. We first establish Theorem 2 which provides a general bound on the angle between $\mathbf{u}_1$ and $\mathbf{v}_1$. We then analyze the expected value of this bound when $\mathbf{P}$ is selected randomly in Theorem 3.

Theorem 2: Let $\mathbf{\Sigma}$ be a general $N \times N$ positive-definite covariance matrix with spectrum $\lambda_1(\mathbf{\Sigma}) \geq \lambda_2(\mathbf{\Sigma}) \geq \ldots \geq \lambda_N(\mathbf{\Sigma})$, and let $\delta_n = \lambda_n(\mathbf{\Sigma}) - \lambda_N(\mathbf{\Sigma})$ for $1 \leq n \leq N-1$ such that $\delta = \sum_{n=2}^{N-1} \delta_n$. Let $\mathbf{w}_1$ be the first eigenvector of $\mathbf{\Sigma}$ associated with the first eigenvalue, $\lambda_1(\mathbf{\Sigma})$. Let $\mathbf{P}$ be an orthonormal $N \times K$ matrix such that $\mathbf{P}^T\mathbf{w}_1 \neq \mathbf{0}$. Then, if $\delta_1 > 0$ and

$$\delta \leq \frac{\delta_1}{5} \|\mathbf{P}^T\mathbf{w}_1\|_2^2 \tag{9}$$

the first eigenvector, $\tilde{\mathbf{u}}_1$, of $\widetilde{\mathbf{\Sigma}} = \mathbf{P}^T\mathbf{\Sigma}\mathbf{P}$ satisfies

$$\sin \omega_1 \leq \frac{4\delta}{\delta_1 \|\mathbf{P}^T\mathbf{w}_1\|_2^2} \tag{10}$$

where $\omega_1 = \angle(\tilde{\mathbf{u}}_1, \mathbf{P}^T\mathbf{w}_1)$.

Proof: See Appendix III. ■

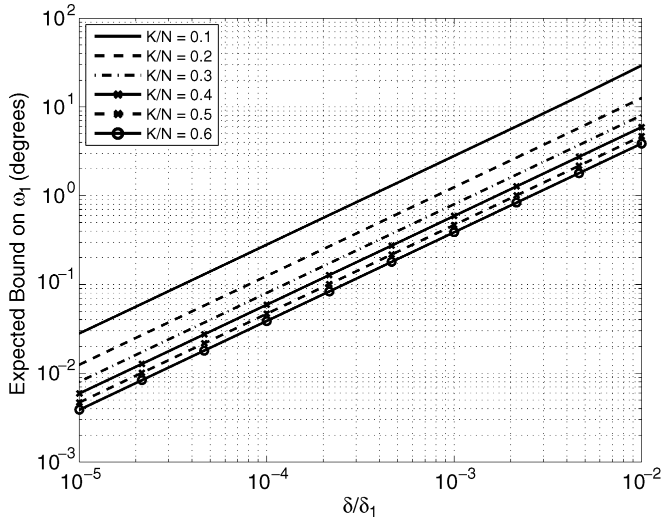

Fig. 2. Expected bound on ω_1 as given by (11) of Theorem 3 (for $N = 100$).

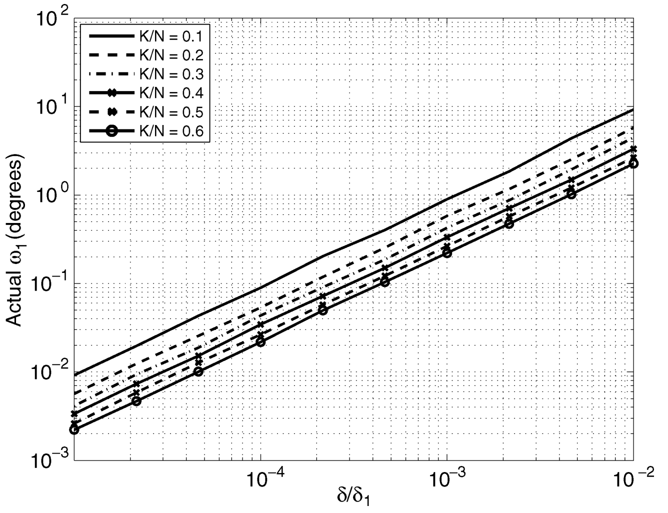

Fig. 3. Experimental evaluation of ω_1 for $\mathbf{\Sigma} = \mathbf{W}\mathbf{\Lambda}\mathbf{W}^T$ with $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, 1, 1, \ldots, 1)$ and $N = 100$.

Theorem 3: For a fixed $\mathbf{\Sigma}$ and, therefore, fixed δ and δ_1 as defined in Theorem 2, the expected value of the upper bound in (10) is

$$E\left[\frac{4\delta}{\delta_1 \|\mathbf{P}^T\mathbf{w}_1\|_2^2}\right] = \frac{4\delta(N-2)}{\delta_1(K-2)}. \tag{11}$$

Proof: See Appendix IV. ■

We note that, if $\lambda_1(\mathbf{\Sigma}) \gg \lambda_2(\mathbf{\Sigma}) \geq \lambda_N(\mathbf{\Sigma})$, then δ_1 will be large. Assuming further that $\delta_1 \gg \delta$, (11) implies that ω_1 in Theorem 2 is likely to be small. Thus, if the covariance matrix is highly eccentric in the direction of the first eigenvector $\mathbf{w}_1$, we expect that the first eigenvector of $\widetilde{\mathbf{\Sigma}} = \mathbf{P}^T\mathbf{\Sigma}\mathbf{P}$ will be close to aligning with the projection of the eigenvector, $\mathbf{P}^T\mathbf{w}_1$. That is, we expect that the first Ritz vector, $\mathbf{u}_1$, will be close to the first normalized projection, $\mathbf{v}_1$.

To experimentally evaluate the ramifications of Theorems 2 and 3, consider Figs. 2 and 3. In Fig. 2, we plot the bound of (11) for varying values of K/N and δ/δ_1. Fig. 2 predicts that we will achieve a small ω_1 angle between $\mathbf{u}_1$ and $\mathbf{v}_1$ if δ/δ_1 is small; for example, with $K/N = 30\%$, we will have $\omega_1 \leq 1^\circ$ for $\delta/\delta_1 \leq 0.001$. To see if this accuracy holds up in practice, we consider $N \times N$ matrix $\mathbf{\Sigma} = \mathbf{W}\mathbf{\Lambda}\mathbf{W}^T$, where $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, 1, 1, \ldots, 1)$, and $\mathbf{W}$ is an arbitrary $N \times N$ orthonormal matrix. We use $N = 100$, set $\lambda_1 = 1001$ (i.e., $\delta_1 = 1000$), and vary λ_2 between 1.00001 and 11 (i.e., $\delta \in [0.00001, 10]$). We generate 1000 random orthonormal projection matrices $\mathbf{P}$ and average the resulting ω_1 angles measured between $\mathbf{u}_1$ and $\mathbf{v}_1$ for these projections. The results are shown in Fig. 3. We observe that the behavior of the curves in Fig. 3 is fairly similar to that of the curves in Fig. 2, except that the actual ω_1 values are somewhat lower than those predicted by the bound of Theorem 3. This is likely due to the fact that the bound of Lemma 2 (see Appendix I) is not particularly tight, resulting in a somewhat loose bound in Theorem 2. Nevertheless, Figs. 2 and 3 affirm that, if $\delta_1 \gg \delta$, we will have the first Ritz vector $\mathbf{u}_1$ lie close to the first normalized projection $\mathbf{v}_1$ as our analysis suggests.

B. Subsequent Eigenvectors

The analysis of the previous section focused exclusively on the first normalized projection $\mathbf{v}_1$ and its corresponding Ritz vector $\mathbf{u}_1$. However, the following result argues that we can extend this approximation beyond the first Ritz vector to subsequent Ritz vectors. That is, we can use the other Ritz vectors, $\mathbf{u}_2$, $\mathbf{u}_3$, etc., to successfully approximate the normalized projections $\mathbf{v}_2$, $\mathbf{v}_3$, etc.

Approximation 1: Let $\mathbf{\Sigma}$ be a general covariance matrix as in Theorem 2. Assume that $\lambda_1(\mathbf{\Sigma}) \gg \lambda_2(\mathbf{\Sigma})$ such that $\mathbf{P}^T\mathbf{w}_1$ successfully aligns with the first eigenvector, $\widetilde{\mathbf{u}}_1$, of $\widetilde{\mathbf{\Sigma}} = \mathbf{P}^T\mathbf{\Sigma}\mathbf{P}$; that is, ω_1 in (10) is small such that

$$\mathbf{P}^T\mathbf{w}_1 / \|\mathbf{P}^T\mathbf{w}_1\|_2 \approx \widetilde{\mathbf{u}}_1 \tag{12}$$

and

$$\lambda_1(\widetilde{\mathbf{\Sigma}}) \approx \left(\lambda_1(\mathbf{\Sigma}) - \lambda_N(\mathbf{\Sigma})\right) \|\mathbf{P}^T\mathbf{w}_1\|_2^2 + \lambda_N(\mathbf{\Sigma}) \tag{13}$$

in (8). Then, if $\lambda_2(\mathbf{\Sigma}) \gg \lambda_3(\mathbf{\Sigma})$

$$\mathbf{P}^T\mathbf{w}_2 / \|\mathbf{P}^T\mathbf{w}_2\|_2 \approx \widetilde{\mathbf{u}}_2 \tag{14}$$

where $\mathbf{w}_2$ and $\widetilde{\mathbf{u}}_2$ are eigenvectors of $\mathbf{\Sigma}$ and $\widetilde{\mathbf{\Sigma}}$, respectively, corresponding to the second-largest eigenvalues $\lambda_2(\mathbf{\Sigma})$ and $\lambda_2(\widetilde{\mathbf{\Sigma}})$.

Justification: See Appendix V. ■

Approximation 1 argues that, if our distribution is sufficiently eccentric in the direction of its first eigenvector to yield close alignment between the first normalized projection and the first Ritz vector, we will also have close alignment between the second normalized projection and the second Ritz vector, as long as we also have sufficient eccentricity in the direction of the second eigenvector. It is a straightforward process to extend this argument to the other eigenvectors as well via successive application of the deflation operation [see (51) in Appendix V] for $\mathbf{w}_2$, $\mathbf{w}_3$, etc. Approximation 1 implies that, if the separation between the $\lambda_k(\mathbf{\Sigma})$ the other eigenvalues is large, then $\mathbf{u}_k$ will be close to $\mathbf{v}_k$.

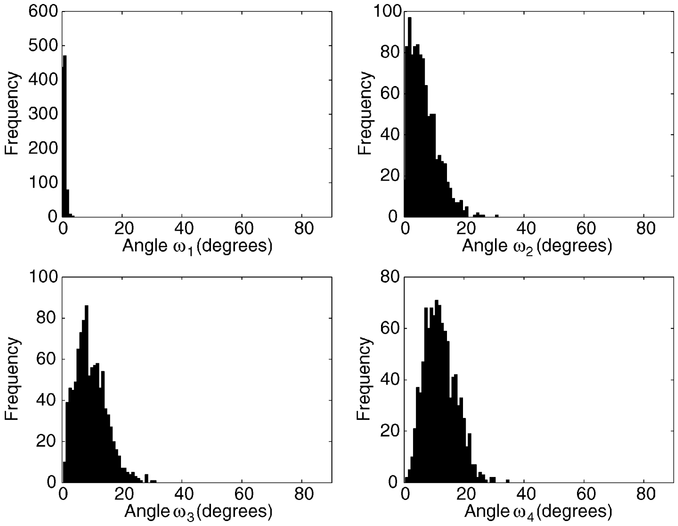

Fig. 4. Histogram of angle ω_i between Ritz vector $\mathbf{u}_i$ and normalized projection $\mathbf{v}_i$ for $i = 1, \ldots, 4$; averages: $\omega_1 = 0.7°$, $\omega_2 = 6.6°$, $\omega_3 = 9.6°$, $\omega_3 = 12.0°$.

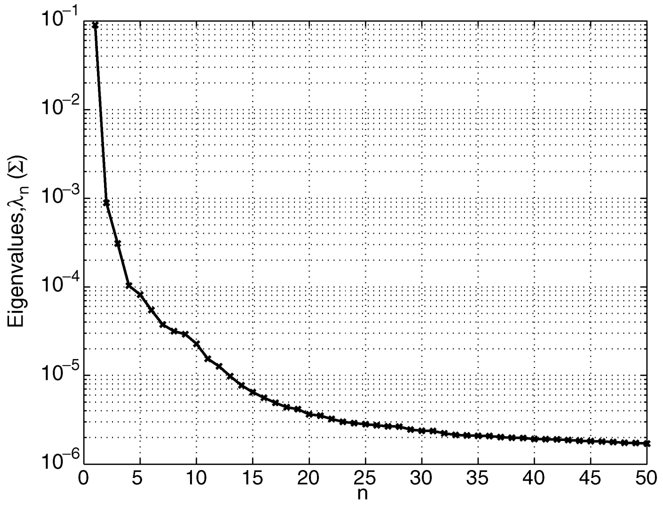

Fig. 5. LEV plot of the first 50 eigenvalues of the "Cuprite" hyperspectral dataset.

To experimentally evaluate Approximation 1 and its implications, we consider $N \times N$ matrix $\mathbf{\Sigma} = \mathbf{W}\mathbf{\Lambda}\mathbf{W}^T$, where

$$\mathbf{\Lambda} = \mathrm{diag}(10000, 1000, 100, 10, 1, \ldots, 1) \qquad (15)$$

and $\mathbf{W}$ is an arbitrary $N \times N$ orthonormal matrix. We generate 1000 random orthonormal projection matrices $\mathbf{P}$; we arbitrarily set $N = 100$ and $K = 40$. Fig. 4 illustrates the histograms of the angles ω_i for $i = 1, \ldots, 4$, where ω_i is the angle between the Ritz vector $\mathbf{u}_i$ and the corresponding normalized projection $\mathbf{v}_i$. We see that, as expected, ω_1 is typically close to 0°, indicating that, indeed, $\mathbf{u}_1$ is very close to $\mathbf{v}_1$ as predicted by Theorems 2 and 3. Additionally, as foreseen by Approximation 1, we have fairly close alignment between $\mathbf{u}_k$ and $\mathbf{v}_k$ for $k = 2$, 3, and 4. We note that ω_i increases for increasing i—this is in line with the exponentially decreasing spectrum of $\mathbf{\Lambda}$, since, as the gap between successive eigenvalues decreases, we expect that the approximations in Approximation 1 will become successively less accurate. Below, we argue that we will often encounter decaying spectra of this nature in many applications; thus, we expect to see this phenomenon of increasing ω_i angle for real data in practice.

C. Eccentricity in Practice

The preceding analysis—as well as the CPPCA technique proposed in the next section—depends on the distribution of $\mathbf{X}$ being sufficiently eccentric. That is, if we wish to approximate the first L normalized projections using the first L Ritz vectors, we need that the first L eigenvalues are sufficiently distinct from not only the other eigenvalues but also from one another. One might question how reliably such eccentricity arises in practice.

The prime objective in many applications of PCA is to reduce dimensionality, and, for this, one must select a certain number of principal components to retain. Consequently, there is great interest in determining "intrinsic dimensionality"—the number of components that account for most of the variation in $\mathbf{X}$. Jolliffe [15] and Jackson [16] discuss extensively in their classic texts on PCA that, while this problem of intrinsic dimensionality has been studied for many decades, several simple and *ad hoc* heuristics are widely used despite attempts to develop more formal methods, simply because the *ad hoc* approaches seem intuitively plausible and are indeed effective in practice. Foremost of these techniques is the scree[2] plot [17] whose very name suggests the ubiquity of eccentric distributions in practice. In the scree plot, which is a graph of $\lambda_n(\mathbf{\Sigma})$ against n, the intrinsic dimensionality is selected to be the point of a sharp "elbow" between a steep descent in principal eigenvalues and a relative shallow slope of the remaining lower eigenvalues. In fact, in some applications, the dynamic range of the principal eigenvalues is so great that $\lambda_n(\mathbf{\Sigma})$ is plotted on a log scale in what is known as a log-eigenvalue (LEV) plot [18], [19]. Such is the case for the hyperspectral data we consider in Section VI; this is exemplified by the LEV plot of Fig. 5.

The bottom line is that, while it is trivial to construct distributions with closely spaced—or worse, equal—principal eigenvalues, such distributions are extremely unlikely to arise in practical PCA applications. On the contrary, as witnessed by the widespread use of scree and LEV plots across numerous fields of application and as exemplified here in Fig. 5, we expect to find distributions that are sufficiently eccentric in their primary principal components so as to permit the analysis above, as well as the CPPCA technique developed next, to be broadly applicable.

IV. CPPCA Algorithm

The analysis of the previous section establishes that Ritz vectors form suitable approximations to orthonormal projections of eigenvectors. We now use this analysis as the basis for a system that uses random projections at the signal sensor as a lightweight encoder. The corresponding decoder then uses the insight from the analysis in the preceding section to implement recovery of not only the PCA coefficients for the transmitted dataset, but also an approximation to the PCA transform basis itself. In this

[2] **Scree**: an accumulation of loose stones or rocky debris lying on a slope or at the base of a hill or cliff (Merriam-Webster).

sense, the resulting CPPCA system in effect shifts the computational complexity of PCA from the encoder to the decoder.

Specifically, in the CPPCA encoder, the M vectors of $\mathbf{X} = [\mathbf{x}_1 \ \cdots \ \mathbf{x}_M]$ are merely each subjected to random projection. On the other hand, the CPPCA decoder then must recover not only the PCA transform coefficients, but also the basis vectors of the transform itself, all from the projections. We assume that the decoder knows only the projection operator and its resulting projections, but not $\mathbf{X}$ or its statistics (e.g., covariance). Below, we present an overview of the CPPCA approach which expands on the initial description put forth in [12]. We initially consider recovery of the eigenvectors and the PCA coefficients alone without the effects of quantization (Sections IV-A and IV-B); Section IV-C then examines issues arising when quantization is then inserted into the process.

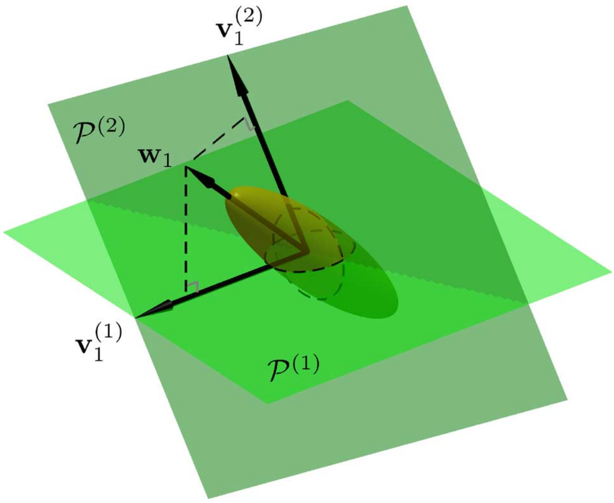

Fig. 6. Two 2-D subspaces $\mathcal{P}^{(1)}$ and $\mathcal{P}^{(2)}$ with corresponding normalized projections $\mathbf{v}_1^{(1)}$ and $\mathbf{v}_1^{(2)}$.

A. Eigenvector Recovery

Traditional design methods for PCA produce the transform $\mathbf{W}$ via the eigendecomposition given by (1); however, in the CPPCA decoder, one has access to merely $\widetilde{\mathbf{\Sigma}}$ and not $\mathbf{\Sigma}$ as required in (1). The goal of CPPCA is thus to approximate $\mathbf{W}$ from $\widetilde{\mathbf{\Sigma}}$ without knowledge of $\mathbf{\Sigma}$, given that $\widetilde{\mathbf{\Sigma}}$ results from random projection. The CPPCA decoder first recovers an approximation to the PCA transform basis by recovering approximations to the first L eigenvectors of $\mathbf{\Sigma}$ from random projections. We observe that, if we knew the true normalized projection $\mathbf{v}$ of eigenvector $\mathbf{w}$ in subspace $\mathcal{P}$, we could form subspace $\mathcal{Q}$ as

$$\mathcal{Q} = \mathcal{P}^{\perp} \oplus \operatorname{span}\{\mathbf{v}\} \tag{16}$$

the direct sum of the orthogonal complement of $\mathcal{P}$ with a 1-D space containing $\mathbf{v}$. Clearly, $\mathbf{w}$ would lie in $\mathcal{Q}$. Suppose then that we produce J distinct random K-dimensional subspaces, $\mathcal{P}^{(1)}$ through $\mathcal{P}^{(J)}$, each containing a normalized projection, $\mathbf{v}^{(1)}$ through $\mathbf{v}^{(J)}$, respectively, produced via (4) using the corresponding projection matrices, $\mathbf{P}^{(1)}$ through $\mathbf{P}^{(J)}$. We could then form subspaces $\mathcal{Q}^{(1)}$ through $\mathcal{Q}^{(J)}$ via (16) using $\mathcal{P}^{(1)}, \ldots, \mathcal{P}^{(J)}$ and $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(J)}$. The eigenvector $\mathbf{w}$ would thus be in the intersection $\mathcal{Q}^{(1)} \cap \ldots \cap \mathcal{Q}^{(J)}$. This situation is illustrated in Figs. 6 and 7 for the case of $N = 3$, $K = 2$, $J = 2$, and the eigenvector in question being $\mathbf{w}_1$.

In the CPPCA decoder, though, we do not have access to the true normalized projections; instead, we can form Ritz vectors in each subspace $\mathcal{P}^{(j)}$ via an eigendecomposition of the corresponding projected covariance matrix $\widetilde{\mathbf{\Sigma}}^{(j)}$. Motivated by the analysis in Section III, we use these Ritz vectors to approximate normalized projections; i.e., we use $\mathbf{u}_k^{(j)}$ instead of $\mathbf{v}_k^{(j)}$ to form the spaces $\mathcal{Q}^{(j)}$. Since the Ritz vectors will differ slightly from the true normalized projections, the intersection $\mathcal{Q}^{(1)} \cap \ldots \cap \mathcal{Q}^{(J)}$ is almost certain to be empty. However, since the $\mathcal{Q}^{(j)}$ are closed and convex, a parallel implementation of POCS will converge to a least-squares solution minimizing the average distance to the subspaces $\mathcal{Q}^{(j)}$ [20]; this POCS solution can then be used to approximate $\mathbf{w}$. Specifically, for iteration $i = 1, 2, \ldots$, we form an estimate of the eigenvector as

$$\widehat{\mathbf{w}}^{(i)} = \frac{1}{J} \sum_{j=1}^{J} \mathbf{Q}^{(j)} \mathbf{Q}^{(j)^T} \widehat{\mathbf{w}}^{(i-1)} \tag{17}$$

where projection onto $\mathcal{Q}^{(j)}$ is performed by the matrix $\mathbf{Q}^{(j)}$, and we initialize $\widehat{\mathbf{w}}^{(0)}$ to the average of the Ritz vectors; (17) will converge to $\widehat{\mathbf{w}}$; normalizing this $\widehat{\mathbf{w}}$ will approximate the desired normalized eigenvector $\mathbf{w}$ (up to sign).

In order to avoid producing multiple random projections for each vector in our dataset, the CPPCA encoder splits the dataset of M vectors $\mathbf{X} = [\mathbf{x}_1 \ \cdots \ \mathbf{x}_M]$ into J partitions $\mathbf{X}^{(j)}$, each associated with its own randomly chosen projection $\mathbf{P}^{(j)}$, $1 \leq j \leq J$. It is assumed that the dataset splitting is conducted such that each $\mathbf{X}^{(j)}$ closely resembles the whole dataset $\mathbf{X}$ statistically and so has approximately the same eigendecomposition.[3] The encoder transmits the projected data $\widetilde{\mathbf{Y}}^{(j)} = \mathbf{P}^{(j)} \mathbf{X}^{(j)}$ to the decoder which is assumed to know the projection operators $\mathbf{P}^{(j)}$ *a priori*. In the CPPCA decoder, $\widetilde{\mathbf{\Sigma}}^{(j)}$ is calculated from $\widetilde{\mathbf{Y}}^{(j)}$, a set of Ritz vectors $\mathbf{u}_k^{(j)}$ is produced from $\widetilde{\mathbf{\Sigma}}^{(j)}$, and then the Ritz vectors are used in place of the normalized projections to drive the POCS recovery of (17). The CPPCA decoder repeats this POCS procedure using the first L Ritz vectors to approximate the first L principal eigenvectors which are assembled into $N \times L$ matrix $\mathbf{\Psi}$, an approximation to the L-component PCA transform, $L \leq K$.

To empirically evaluate the performance of the proposed POCS recovery of eigenvectors, let us again consider $N \times N$ matrix $\mathbf{\Sigma} = \mathbf{W} \mathbf{\Lambda} \mathbf{W}^T$, with $\mathbf{\Lambda}$ as in (15) and $\mathbf{W}$ an arbitrary $N \times N$ orthonormal matrix. We apply the POCS iteration of (17) using the Ritz vectors $\mathbf{u}^{(j)}$ in the place of the normalized projections $\mathbf{v}^{(j)}$. Let ξ_i be the angle between the true eigenvector $\mathbf{w}_i$ and its approximation $\widehat{\mathbf{w}}_i$ that results from

[3] Dataset subsampling is commonly used to expedite covariance-matrix calculation in traditional applications of PCA, e.g., [2], [21]; we suggest modulo partitioning such as $\mathbf{X}^{(j)} = \{\mathbf{x}_m \in \mathbf{X} | (m-1) \bmod J = j-1\}$.

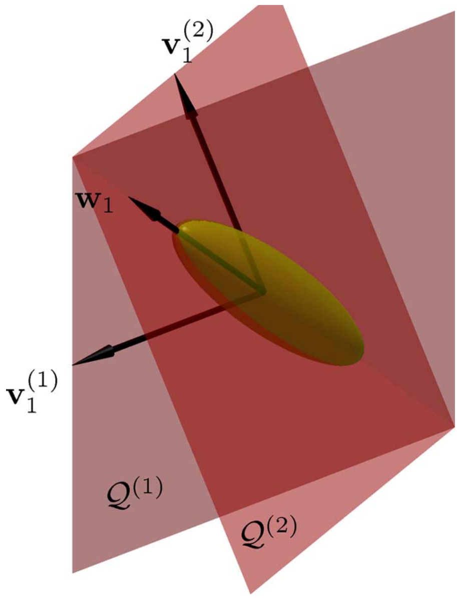

Fig. 7. Subspaces $\mathcal{Q}^{(1)}$ and $\mathcal{Q}^{(2)}$ whose intersection uniquely determines eigenvector $\mathbf{w}_1$ up to a sign.

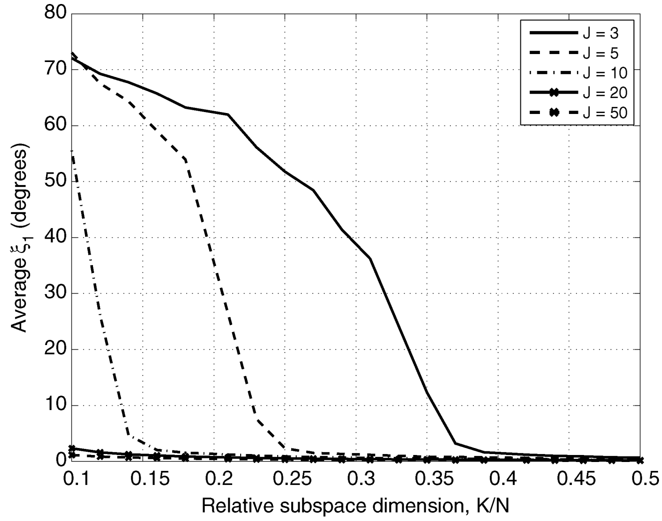

Fig. 8. Average error angle ξ_1 between eigenvector $\mathbf{w}_1$ and its approximation $\widehat{\mathbf{w}}_1$ resulting from POCS-based eigenvector recovery.

(17). We generate sets of random matrices, each set containing J $N \times K$ projection matrices $\mathbf{P}^{(1)}$ through $\mathbf{P}^{(J)}$ for $N = 100$; we average results over 100 trials. Fig. 8 shows the average error angle ξ_1 as both the number of projections J and the dimensionality of the projections K (relative to N) vary. Fig. 9 shows similar results for the error angle ξ_2 associated with the second eigenvector. We see that, when the dimensionality of the projection spaces is small (K/N small), then a larger number, J, of distinct projections is needed to produce a small error between $\mathbf{w}_i$ and $\widehat{\mathbf{w}}_i$. However, there is an aspect of diminishing returns—the amount of reduction in ξ_i decreases for each increase in J. Subsequently, we focus on $J = 20$ which appears to give a reasonable tradeoff between approximation accuracy and computational complexity for our data.

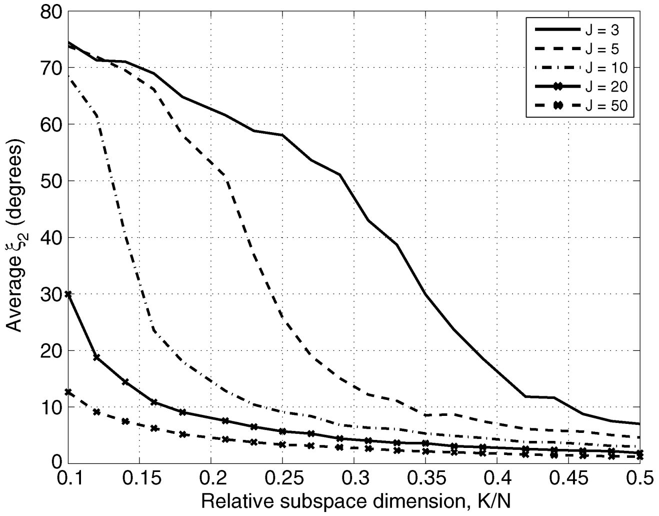

Fig. 9. Average error angle ξ_2 between eigenvector $\mathbf{w}_2$ and its approximation $\widehat{\mathbf{w}}_2$ resulting from POCS-based eigenvector recovery.

B. *Coefficient Recovery*

Once obtaining $\boldsymbol{\Psi}$, the CPPCA decoder then proceeds to recover the PCA coefficients by solving $\widetilde{\mathbf{Y}}^{(j)} = \mathbf{P}^{(j)^T}\boldsymbol{\Psi}\check{\mathbf{X}}^{(j)}$ for PCA coefficients $\check{\mathbf{X}}^{(j)}$ in the least-squares sense for each j. This linear reconstruction can be accomplished in several ways, for example, by using the pseudoinverse

$$\check{\mathbf{X}}^{(j)} = \left(\mathbf{P}^{(j)^T}\boldsymbol{\Psi}\right)^+ \widetilde{\mathbf{Y}}^{(j)}. \qquad (18)$$

C. *Quantization Issues*

True data compression, and not just dimensionality reduction, must necessarily involve some form of quantization. In CPPCA, quantization of the projections will produce distortions in both the coefficient-recovery process as well as in the eigenvector recovery used to approximate the PCA transform. However, known results from perturbation theory argue that the eigenvector-recovery procedure central to CPPCA is robust under quantization.

Specifically, in CPPCA, original vector $\mathbf{x} \in \mathbb{R}^N$ is projected into a K-dimensional subspace $\mathcal{P}$ as $\widetilde{\mathbf{y}} = \mathbf{P}^T\mathbf{x}$. Assume uniform scalar quantization (USQ) is applied to the components of $\widetilde{\mathbf{y}}$. In order to analyze the effect of this quantization process on the performance of CPPCA, we adopt a simplified, high-resolution model [22] of USQ as additive noise of variance $q_k^2/12$, where q_k is the quantizer stepsize for component k of $\widetilde{\mathbf{y}}$. That is, $\widetilde{\mathbf{y}}$ is quantized as $\widehat{\mathbf{y}} = \widetilde{\mathbf{y}} + \mathbf{n}$ where noise $\mathbf{n}$ has covariance $\mathbf{N} = E[\mathbf{n}\mathbf{n}^T] = \operatorname{diag}(q_k^2)/12$ and zero mean. Let $q = \max_k q_k$ and note $\lambda_1(\mathbf{N}) = q^2/12$. The covariance of $\widehat{\mathbf{y}}$ is then

$$\widehat{\boldsymbol{\Sigma}} = E[\widehat{\mathbf{y}}\widehat{\mathbf{y}}^T] = E\left[(\widetilde{\mathbf{y}}+\mathbf{n})(\widetilde{\mathbf{y}}+\mathbf{n})^T\right] = \widetilde{\boldsymbol{\Sigma}} + \mathbf{N}. \qquad (19)$$

CPPCA will recover both the PCA coefficients as well as the basis vectors of the PCA transform itself from the quantized projections $\widehat{\mathbf{y}}$. Suppose $\widetilde{\mathbf{u}}_1$ is the first eigenvector of $\widehat{\boldsymbol{\Sigma}}$, and its gap is $\widetilde{\delta}_1 = \lambda_1(\widetilde{\mathbf{\Sigma}}) - \lambda_2(\widetilde{\mathbf{\Sigma}})$. Then, if $q^2/12 \leq \widetilde{\delta}_1/5$, from Proposition 3 (see Appendix I), we have

$$\sin \zeta_1 \leq \frac{q^2}{3\widetilde{\delta}_1} \tag{20}$$

where ζ_1 is the angle between $\widetilde{\mathbf{u}}_1$ and $\widehat{\mathbf{u}}_1$. We note that similar bounds result for $\zeta_k = \angle(\widetilde{\mathbf{u}}_k, \widehat{\mathbf{u}}_k)$ for $k \geq 2$ via Theorem 8.1.12 of [23].

We note that the central idea driving CPPCA in the first place—that Ritz vectors form reasonable approximations to normalized projections of eigenvectors—relies on the distribution of $\mathbf{X}$ being eccentric, i.e., the eigenvalues being sufficiently distinct from one another. We see thus that the same phenomenon that permits eigenvector recovery—mutually distinct eigenvalues—also encourages stability of the result under scalar quantization. The bound of (20) ensures a graceful degradation in the accuracy of the CPPCA eigenvector-recovery procedure as the maximum quantizer stepsize q increases. Furthermore, eigenvector recovery is more robust to quantization distortion as the gap between the eigenvalues and, consequently, the eccentricity of the distribution, increases.

We observe that, despite the stability to quantization discussed here, CPPCA, like CS, is unlikely to be competitive with traditional source-coding algorithms in terms of rate-distortion performance. On the contrary, the strength of CPPCA lies in random projections that provide encoder simplicity and universality, advantages that may be of great use in certain applications.

V. Connections to Other Work

CPPCA bears some similarity to CS in that both effectuate a recovery from random projections; however, there are some significant differences. In brief, CS (e.g., [7]–[10]) produces a sparse signal representation directly from a small number of projections onto another basis, recovering the sparse transform coefficients via nonlinear reconstruction. The main tenet of CS theory holds that, if signal $\mathbf{x} \in \mathbb{R}^N$ can be sparsely represented (i.e., using only L nonzero coefficients) with some basis, then we can recover $\mathbf{x}$ from K-dimensional projections $\widetilde{\mathbf{y}} = \mathbf{P}^T\mathbf{x}$ under certain conditions; here $\mathbf{P} = [\mathbf{p}_1 \cdots \mathbf{p}_K]$, and $K < N$. For recovery of a set of multiple, possibly correlated vectors $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_M]$, there have been proposals for multivector extensions of CS under the name of "multitask" [24] or "distributed" [25] CS; these, in turn, link closely to a larger body of literature on "simultaneous sparse approximation" (e.g., [26]–[30]). In experimental results below, we compare the performance of CPPCA to that of Multi-Task Bayesian Compressive Sensing (MT-BCS) [24] which introduces a hierarchical Bayesian framework into the multivector CS-recovery problem to share prior information across the multiple vectors. We note that, on the surface, although CPPCA and MT-BCS appear somewhat similar in their functionality, there exist some crucial differences. MT-BCS, like other CS techniques, operates under an assumption of sparsity in a *known* basis $\mathbf{\Psi}$, but the pattern of sparsity (i.e., which L components are nonzero) is *unknown*. On the other hand, CPPCA reconstruction operates under a *known* sparsity pattern (i.e., the first L principal components), but the transform $\mathbf{\Psi}$ itself is *unknown*. Additionally, while MT-BCS can recover the M vectors of $\mathbf{X}$ from the same set of projections $\widetilde{\mathbf{Y}}^{(j)} = \mathbf{P}^{(j)}\mathbf{X}^{(j)}$ which drive the CPPCA recovery process, it can also function on arbitrarily small numbers of vectors, even down to $M = 1$ (in which case, MT-BCS becomes the special case of "single-task" CS recovery). CPPCA, on the other hand, requires M to be sufficiently large to enable covariance-matrix calculation in the J subspaces.

Finally, we note that the analysis in Section III that underlies CPPCA is consistent with the celebrated Johnson–Lindenstrauss Lemma [31]–[33] which holds that L points in $\mathbb{R}^N$ can be projected into a K-dimensional subspace while approximately maintaining pairwise distances as long as $K \geq O(\log L)$. If we project L eigenvectors into K-dimensional subspace $\mathcal{P}$, we have that the L points represented by the eigenvectors will roughly maintain pairwise distances, and the L eigenvectors will approximately maintain their lengths. Since the eigenvectors are mutually orthogonal in $\mathbb{R}^N$, by the Pythagorean theorem, the projected eigenvectors will also be approximately mutually orthogonal in $\mathcal{P}$, if K is sufficiently large with respect to L. Since the Ritz vectors used to drive CPPCA reconstruction are necessarily mutually orthogonal in $\mathcal{P}$, it is plausible that they could approximate the projected eigenvectors. In this respect, CPPCA is consistent with the Johnson–Lindenstrauss Lemma.

VI. Experimental Results

We now examine the performance of CPPCA reconstruction in the form of (17) and (18). Since both CPPCA and MT-BCS feature lightweight encoding via random projections, and there is increasing interest in integrating CS methodology directly into hyperspectral sensors (e.g., [34]–[36]), our comparisons focus on CPPCA performance relative to that of MT-BCS for real hyperspectral data.

We use hyperspectral images cropped spatially to size 100 × 100 (i.e., $M = 10,000$); we use the popular "Cuprite" and "Jasper Ridge" images, AVIRIS datasets with $N = 224$ spectral bands. The mean vector has been removed from the vectors to impose a zero-mean condition. For CPPCA, we use $J = 20$ projection partitions as discussed above while L ranges between 3 and 30, depending on the specific K used. For MT-BCS, we consider several orthonormal bases commonly used with hyperspectral data: an N-point DCT (MT-BCS-DCT) as well as DWTs using both the Haar basis (MT-BCS-Haar) and the length-4 Daubechies basis (MT-BCS-D4). We apply the same random projections as used for CPPCA. We use the MT-BCS implementation provided by its authors.[4]

Clearly, the performance of both CPPCA and MT-BCS will depend on the degree of dataset reduction inherent in the projections; this quantity is characterized as a relative projection dimensionality in the form of K/N expressed as a percentage. We see from Figs. 10 and 11 that CPPCA yields average SNR substantially higher than that of the fixed-basis MT-BCS approaches over a broad range of practical K/N values.

[4] http://www.people.ee.duke.edu/~lihan/cs/.

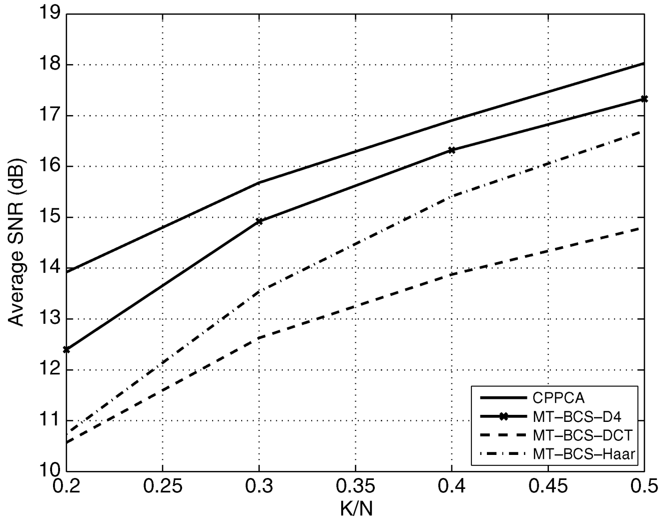

Fig. 10. Reconstruction performance for the "Cuprite" hyperspectral dataseta—average SNR for varying dimensionality K/N.

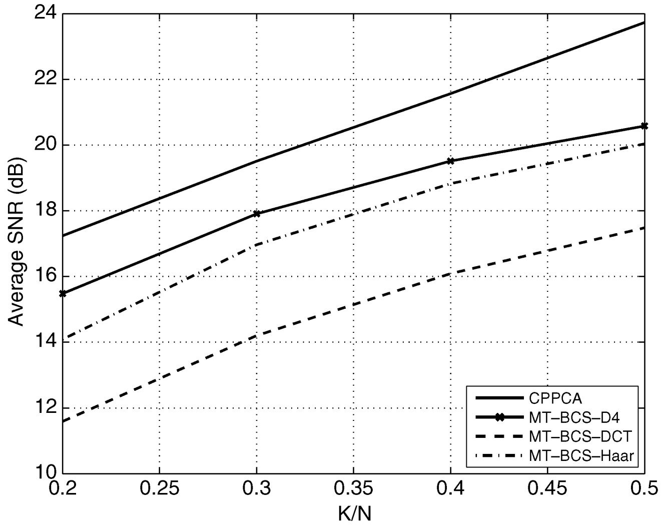

Fig. 11. Reconstruction performance for the "Jasper Ridge" hyperspectral dataset—average SNR for varying dimensionality K/N.

In terms of computational complexity, none of the implementations we employ are optimized for execution speed. Informally, however, we have observed that both the POCS-based eigenvector recovery of (17) as well as the linear coefficient recovery of (18) are quite fast. In the production of the experimental results of this section, CPPCA runs about an order of magnitude faster than MT-BCS.

VII. Conclusion

In this paper, we presented an approach that exploits encoder-side compressive projections to effectively shift the computational burden of PCA from the encoder to the decoder. This CPPCA technique coupled random projections at the encoder with a Rayleigh–Ritz process for approximating eigenvectors at the decoder. Central to this development was an extensive analysis of the relation between Ritz vectors and normalized projections of eigenvectors. Our analysis provided a bound on the angle between the first Ritz vector and the first normalized projection that was expected to be small under the conditions of widely separated eigenvalues (i.e., a highly eccentric data distribution) as well as randomly selected subspace projections. Additionally, further deflation-based analysis argued that this approximation strategy—using Ritz vectors to approximate projected eigenvectors—could be extended beyond the first eigenpair. This analysis provided the motivation for our proposed CPPCA algorithm for reconstruction in which a POCS-based optimization driven by Ritz vectors approximated the eigenvectors constituting the PCA transform. As a consequence, the CPPCA decoder, given only the random projections created by the encoder, recovered not only the coefficients associated with the PCA transform, but also an approximation to the PCA transform basis itself.

We anticipate that CPPCA will be most useful in applications in which the encoder-side random projections are not a computationally separate process but are instead integrated directly into the signal sensing and acquisition device. In this manner, with dimensionality reduction performed simultaneously with signal acquisition, one can avoid not only the computational burden of explicit dimensionality reduction, but also the production of onerous quantities of data in the first place. We expect such operation to be of great value in situations of resource-constrained signal-sensing platforms, particularly in satellite-borne remote-sensing applications. Although we have focused the experimental results of this paper on hyperspectral data, CPPCA itself is, however, algorithmically more general. Indeed, CPPCA is applicable to any dataset that takes the form of a collection of vectors, as long as it makes sense from an application perspective to apply PCA to those vectors. In such cases, CPPCA promises to be an effective strategy to provide simultaneous signal sensing and compression along with effective reconstruction.

Appendix I
Preliminaries

Proposition 1: Let $\mathbf{P}$ be an $N \times K$ orthonormal projection matrix. Denote the columns of $\mathbf{P}$ as $\mathbf{p}_k$ and the rows as $\bar{\mathbf{p}}_n^T$ such that $\mathbf{P} = [\mathbf{p}_1 \cdots \mathbf{p}_K] = [\bar{\mathbf{p}}_1 \cdots \bar{\mathbf{p}}_N]^T$. Let $N \times (N-K)$ matrix $\mathbf{R} = \text{null}(\mathbf{P}^T)$; i.e., the columns of $\mathbf{R}$ form an orthonormal basis of the nullspace of $\mathbf{P}^T$ such that $\mathbf{P}^T\mathbf{R} = \mathbf{0}$. Let the columns and rows of $\mathbf{R}$ be $\mathbf{r}_k$ and $\bar{\mathbf{r}}_n^T$, respectively. Form orthonormal $N \times N$ matrix $\mathbf{G}$ as $\mathbf{G} = [\mathbf{P}\ \mathbf{R}]$. Let the columns and rows of $\mathbf{G}$ be $\mathbf{g}_n$ and $\bar{\mathbf{g}}_n$, respectively, and define $N \times N$ matrix $\mathbf{H}$ as

$$\mathbf{H} = \mathbf{R}\mathbf{R}^T. \tag{21}$$

Then

$$\|\mathbf{g}_n\|_2^2 = \|\bar{\mathbf{g}}_n\|_2^2 = \|\bar{\mathbf{p}}_n\|_2^2 + \|\bar{\mathbf{r}}_n\|_2^2 = 1 \tag{22}$$

and the diagonal elements, h_{nn}, of $\mathbf{H}$ are

$$h_{nn} = \bar{\mathbf{r}}_n^T\bar{\mathbf{r}}_n = \|\bar{\mathbf{r}}_n\|_2^2. \tag{23}$$

■

Define the spectrum of $N \times N$ matrix $\mathbf{A}$ to be $\lambda(\mathbf{A}) = \{\lambda_1(\mathbf{A}), \ldots, \lambda_N(\mathbf{A})\}$ such that

$$\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \cdots \lambda_N(\mathbf{A}) \tag{24}$$

where $\lambda_n(\mathbf{A})$ is the n^{th} largest eigenvalue of $\mathbf{A}$. The following results from [23] characterize how the eigendecomposition of $\mathbf{A}$ is affected by a symmetric perturbation $\mathbf{E}$.

Proposition 2: If $\mathbf{A}$ and $\mathbf{E}$ are symmetric $N \times N$ matrices

$$\lambda_n(\mathbf{A}) + \lambda_N(\mathbf{E}) \leq \lambda_n(\mathbf{A} + \mathbf{E}) \leq \lambda_n(\mathbf{A}) + \lambda_1(\mathbf{E}) \tag{25}$$

for $1 \leq n \leq N$.

Proof: This is Theorem 8.1.5 of [23]. ■

Proposition 3: Suppose $\mathbf{A}$ and $\mathbf{E}$ are symmetric $N \times N$ matrices. Define the gap $\delta = \lambda_1(\mathbf{A}) - \lambda_2(\mathbf{A})$, and let the first eigenvector of $\mathbf{A}$ be $\mathbf{w}_1$. Consider $\mathbf{A} + \mathbf{E}$ with first eigenvector $\mathbf{w}'_1$. If $\delta > 0$ and $\lambda_1(\mathbf{E}) \leq \delta/5$, then

$$\sin \angle\left(\mathbf{w}_1, \mathbf{w}'_1\right) \leq \frac{4}{\delta}\lambda_1(\mathbf{E}) \tag{26}$$

where $\angle(\mathbf{w}_1, \mathbf{w}'_1)$ is the angle between $\mathbf{w}_1$ and $\mathbf{w}'_1$.

Proof: This result is an immediate consequence of Appendix A.1 of [37] which in turn derives from Theorem 8.1.12 of [23]. ■

We now consider how transformations in the form of $\mathbf{P}^T\mathbf{AP}$ affect the eigendecomposition of diagonal matrices $\mathbf{A}$. We start with the simple rank-1 case.

Lemma 1: Let $\mathbf{A}_n$ be an $N \times N$ rank-1 diagonal matrix with nonzero eigenvalue $\lambda_1(\mathbf{A}_n) = a_n > 0$. That is, $\mathbf{A}_n = \mathrm{diag}(\ldots, 0, a_n, 0, \ldots)$. Then, for orthonormal $N \times K$ matrix $\mathbf{P}$, the $K \times K$ matrix $\mathbf{P}^T\mathbf{A}_n\mathbf{P}$ has spectrum $\lambda(\mathbf{P}^T\mathbf{A}_n\mathbf{P}) = \{\lambda_1(\mathbf{P}^T\mathbf{A}_n\mathbf{P}), 0, \ldots, 0\}$, where $\lambda_1(\mathbf{P}^T\mathbf{A}_n\mathbf{P}) \leq a_n$.

Proof: Because matrix $\mathbf{A}_n$ has unit rank while $\mathrm{rank}(\mathbf{P}^T\mathbf{A}_n\mathbf{P}) \leq \min\{\mathrm{rank}(\mathbf{P}^T), \mathrm{rank}(\mathbf{A}_n), \mathrm{rank}(\mathbf{P})\}$, we have then that $\mathrm{rank}(\mathbf{P}^T\mathbf{A}_n\mathbf{P})$ is at most 1. Thus, we can form $\mathbf{A}'_n = \mathrm{diag}(\ldots, 0, \sqrt{a_n}, 0, \ldots)$, factor $\mathbf{P}^T\mathbf{A}_n\mathbf{P}$ as $\mathbf{P}^T\mathbf{A}_n\mathbf{P} = (\mathbf{P}^T\mathbf{A}'^{\,T}_n)(\mathbf{A}'_n\mathbf{P})$, and then know that $\mathbf{P}^T\mathbf{A}_n\mathbf{P}$ is positive semidefinite ([38], Theorem 6E). As a consequence, $\lambda_k(\mathbf{P}^T\mathbf{A}_n\mathbf{P}) \geq 0$, and only $\lambda_1(\mathbf{P}^T\mathbf{A}_n\mathbf{P})$ is potentially nonzero.

The multiplication $\mathbf{A}_n\mathbf{P}$ simply scales the n^{th} row of $\mathbf{P}$ by a_n. Thus

$$\mathbf{P}^T\mathbf{A}_n\mathbf{P} = [\bar{\mathbf{p}}_1 \quad \cdots \quad \bar{\mathbf{p}}_N][\cdots \quad 0 \quad a_n\bar{\mathbf{p}}_n \quad 0 \quad \cdots]^T \tag{27}$$

and along the diagonal we have

$$\mathrm{diag}(\mathbf{P}^T\mathbf{A}_n\mathbf{P}) = \left(a_n\bar{p}_{n1}^2, a_n\bar{p}_{n2}^2, \ldots, a_n\bar{p}_{nK}^2\right) \tag{28}$$

where $\bar{\mathbf{p}}_n = [\bar{p}_{n1} \cdots \bar{p}_{nK}]^T$. For a rank-1 matrix, the trace is equal to the first eigenvalue; thus

$$\lambda_1(\mathbf{P}^T\mathbf{A}_n\mathbf{P}) = \mathrm{trace}(\mathbf{P}^T\mathbf{A}_n\mathbf{P}) = a_n\|\bar{\mathbf{p}}_n\|_2^2 \leq a_n \tag{29}$$

where the inequality is a consequence of (22). ■

We now employ Proposition 2 to aid in establishing the more general diagonal case.

Lemma 2: Suppose $N \times N$ positive-semidefinite diagonal matrix $\mathbf{A} = \mathrm{diag}(a_1, a_2, \ldots, a_N) = \sum_{n=1}^{N}\mathbf{A}_n$, where $\mathbf{A}_n$ is as defined in Lemma 1 in the case $a_n \neq 0$ and $\mathbf{A}_n = \mathbf{0}$ otherwise. For $N \times K$ matrix $\mathbf{P}$ with orthonormal columns

$$\lambda_1(\mathbf{P}^T\mathbf{AP}) \leq \mathrm{trace}(\mathbf{A}). \tag{30}$$

Proof: The first eigenvalue of $\mathbf{P}^T\mathbf{AP}$ satisfies

$$\begin{aligned}
\lambda_1(\mathbf{P}^T\mathbf{AP}) &= \lambda_1\left(\mathbf{P}^T\mathbf{A}_1\mathbf{P} + \sum_{n=2}^{N}\mathbf{P}^T\mathbf{A}_n\mathbf{P}\right) \\
&\leq \lambda_1(\mathbf{P}^T\mathbf{A}_1\mathbf{P}) + \lambda_1\left(\sum_{n=2}^{N}\mathbf{P}^T\mathbf{A}_n\mathbf{P}\right) \\
&\leq a_1 + \lambda_1\left(\sum_{n=2}^{N}\mathbf{P}^T\mathbf{A}_n\mathbf{P}\right)
\end{aligned} \tag{31}$$

where the first inequality is due to the right side of (25) in Proposition 2 and the second is due to Lemma 1. Applying this result recursively to the rightmost term yields

$$\begin{aligned}
\lambda_1(\mathbf{P}^T\mathbf{AP}) &\leq a_1 + a_2 + \lambda_1\left(\sum_{n=3}^{N}\mathbf{P}^T\mathbf{A}_n\mathbf{P}\right) \\
&\leq \cdots \leq \sum_{n=1}^{N} a_n = \mathrm{trace}(\mathbf{A}).
\end{aligned} \tag{32}$$

■

APPENDIX II
PROOF OF THEOREM 1

For the analysis here and in the next appendix, we assume a diagonal covariance matrix of the form $\boldsymbol{\Sigma} = \mathrm{diag}(\lambda_1(\boldsymbol{\Sigma}), \ldots, \lambda_N(\boldsymbol{\Sigma}))$ where $\lambda_1(\boldsymbol{\Sigma}) \geq \cdots \geq \lambda_N(\boldsymbol{\Sigma})$ are the eigenvalues with corresponding eigenvectors being columns of the identity matrix. We argue that we can do so without loss of generality since the alignment of the eigenvectors with the coordinate axes of $\mathbb{R}^N$ is merely a matter of the selection of the coordinate system for $\mathbb{R}^N$, and a unitary rotation will suffice to achieve this alignment in the more general case of a nondiagonal $\boldsymbol{\Sigma}$. The quantities of interest in the following proofs—vector lengths and angles between vectors—are invariant to such unitary rotations.

We have diagonal, single-spike covariance $\boldsymbol{\Sigma}$ with first eigenvector

$$\mathbf{w}_1 = [1 \quad 0 \quad \cdots \quad 0]^T \tag{33}$$

and $\lambda_1(\boldsymbol{\Sigma}) > \lambda_2(\boldsymbol{\Sigma}) = \lambda_3(\boldsymbol{\Sigma}) = \ldots = \lambda_N(\boldsymbol{\Sigma}) > 0$. Define matrix $\boldsymbol{\Delta}$ as

$$\boldsymbol{\Delta} = \boldsymbol{\Sigma} - \lambda_N(\boldsymbol{\Sigma})\mathbf{I} = \mathrm{diag}(\delta_1, 0, \ldots, 0) \tag{34}$$

and form an $N \times N$ orthonormal matrix from $\mathbf{P}$ by concatenating additional orthonormal columns; i.e., let $\mathbf{G} = [\mathbf{P} \ \mathbf{R}]$ where $\mathbf{R} = \mathrm{null}(\mathbf{P}^T)$ as in Proposition 1. Define also $\mathbf{H}$ as in (21) and note that $\mathbf{P}^T\mathbf{w}_1 = \bar{\mathbf{p}}_1 \neq \mathbf{0}$ implies $\|\bar{\mathbf{p}}_1\|_2^2 > 0$, and

$$h_{11} = \|\bar{\mathbf{r}}_1\|_2^2 = 1 - \|\bar{\mathbf{p}}_1\|_2^2 < 1 \tag{35}$$

from (21) and (22).

The similarity transform $\mathbf{G}^{-1}\mathbf{\Sigma}\mathbf{G}$ yields

$$\mathbf{G}^{-1}\mathbf{\Sigma}\mathbf{G} = \mathbf{G}^T\mathbf{\Sigma}\mathbf{G} = \begin{bmatrix} \mathbf{P}^T \\ \mathbf{R}^T \end{bmatrix} \mathbf{\Sigma}[\mathbf{P} \quad \mathbf{R}] = \begin{bmatrix} \widetilde{\mathbf{\Sigma}} & \mathbf{P}^T\mathbf{\Sigma}\mathbf{R} \\ \mathbf{R}^T\mathbf{\Sigma}\mathbf{P} & \mathbf{R}^T\mathbf{\Sigma}\mathbf{R} \end{bmatrix}. \quad (36)$$

Since $(\lambda_1(\mathbf{\Sigma}), \mathbf{w}_1)$ is an eigenpair of $\mathbf{\Sigma}$, $(\lambda_1(\mathbf{\Sigma}), \mathbf{G}^T\mathbf{w}_1)$ is an eigenpair of $\mathbf{G}^T\mathbf{\Sigma}\mathbf{G}$ due to the inherent nature of a similarity transform (Theorem 5P of [38]); thus, $\mathbf{G}^T\mathbf{\Sigma}\mathbf{G}\mathbf{G}^T\mathbf{w}_1 = \lambda_1(\mathbf{\Sigma})\mathbf{G}^T\mathbf{w}_1$, or

$$\begin{bmatrix} \widetilde{\mathbf{\Sigma}} & \mathbf{P}^T\mathbf{\Sigma}\mathbf{R} \\ \mathbf{R}^T\mathbf{\Sigma}\mathbf{P} & \mathbf{R}^T\mathbf{\Sigma}\mathbf{R} \end{bmatrix} \begin{bmatrix} \mathbf{P}^T \\ \mathbf{R}^T \end{bmatrix} \mathbf{w}_1 = \lambda_1(\mathbf{\Sigma}) \begin{bmatrix} \mathbf{P}^T \\ \mathbf{R}^T \end{bmatrix} \mathbf{w}_1. \quad (37)$$

Considering only the first row of (37), and noting that $\mathbf{P}^T\mathbf{H} = \mathbf{0}$, we have

$$\begin{aligned} \lambda_1(\mathbf{\Sigma})\mathbf{P}^T\mathbf{w}_1 &= (\widetilde{\mathbf{\Sigma}}\mathbf{P}^T + \mathbf{P}^T\mathbf{\Sigma}\mathbf{H})\mathbf{w}_1 \\ &= \left(\widetilde{\mathbf{\Sigma}}\mathbf{P}^T + \mathbf{P}^T\left(\mathbf{\Delta} + \lambda_N(\mathbf{\Sigma})\mathbf{I}\right)\mathbf{H}\right)\mathbf{w}_1 \\ &= (\widetilde{\mathbf{\Sigma}}\mathbf{P}^T + \mathbf{P}^T\mathbf{\Delta}\mathbf{H})\mathbf{w}_1 \end{aligned} \quad (38)$$

from the definition of $\mathbf{\Delta}$ in (34). From (33) and (34), we see that $\mathbf{\Delta}\mathbf{H}\mathbf{w}_1 = \delta_1 h_{11}\mathbf{w}_1$, where h_{11} is the first element on the diagonal of $\mathbf{H}$. Thus, (38) becomes

$$\lambda_1(\mathbf{\Sigma})\mathbf{P}^T\mathbf{w}_1 = \widetilde{\mathbf{\Sigma}}\mathbf{P}^T\mathbf{w}_1 + \delta_1 h_{11}\mathbf{P}^T\mathbf{w}_1 \quad (39)$$

or

$$\widetilde{\mathbf{\Sigma}}\mathbf{P}^T\mathbf{w}_1 = \left(\lambda_1(\mathbf{\Sigma}) - \delta_1 h_{11}\right)\mathbf{P}^T\mathbf{w}_1. \quad (40)$$

Thus, we see that $\mathbf{P}^T\mathbf{w}_1$ is an eigenvector of $\widetilde{\mathbf{\Sigma}}$.

We now must establish that $\lambda_1(\mathbf{\Sigma}) - \delta_1 h_{11}$ is in fact the largest eigenvalue of $\widetilde{\mathbf{\Sigma}}$ such that $\mathbf{P}^T\mathbf{w}_1/\|\mathbf{P}^T\mathbf{w}_1\|_2$ is actually its first eigenvector. We note that

$$\begin{aligned} \widetilde{\mathbf{\Sigma}} &= \mathbf{P}^T\mathbf{\Sigma}\mathbf{P} = \mathbf{P}^T\left(\mathbf{\Delta} + \lambda_N(\mathbf{\Sigma})\mathbf{I}\right)\mathbf{P} \\ &= \mathbf{P}^T\mathbf{\Delta}\mathbf{P} + \lambda_N(\mathbf{\Sigma})\mathbf{I} \end{aligned} \quad (41)$$

since $\mathbf{P}^T\mathbf{P} = \mathbf{I}$. Due to the fact that all eigenvalues of $\lambda_N(\mathbf{\Sigma})\mathbf{I}$ are $\lambda_N(\mathbf{\Sigma})$, the inequalities in Proposition 2 become equalities, and we have

$$\begin{aligned} \lambda_k(\widetilde{\mathbf{\Sigma}}) &= \lambda_k\left(\mathbf{P}^T\mathbf{\Delta}\mathbf{P} + \lambda_N(\mathbf{\Sigma})\mathbf{I}\right) \\ &= \lambda_k(\mathbf{P}^T\mathbf{\Delta}\mathbf{P}) + \lambda_N(\mathbf{\Sigma}) \\ &= \begin{cases} \lambda_1(\mathbf{P}^T\mathbf{\Delta}\mathbf{P}) + \lambda_N(\mathbf{\Sigma}), & k = 1, \\ \lambda_N(\mathbf{\Sigma}), & 2 \le k \le K \end{cases} \end{aligned} \quad (42)$$

where we note that $\mathbf{\Delta}$ is a rank-1 diagonal matrix and invoke Lemma 1 to reveal that $\lambda_k(\mathbf{P}^T\mathbf{\Delta}\mathbf{P}) = 0$ for $k > 1$. From (35), $0 \le \delta_1 h_{11} < \delta_1$, and the eigenvalue in question, $\lambda_1(\mathbf{\Sigma}) - \delta_1 h_{11}$, satisfies

$$\lambda_1(\mathbf{\Sigma}) - \delta_1 h_{11} > \lambda_1(\mathbf{\Sigma}) - \delta_1 = \lambda_N(\mathbf{\Sigma}). \quad (43)$$

Thus, $\lambda_1(\mathbf{\Sigma}) - \delta_1 h_{11}$ must be $\lambda_1(\widetilde{\mathbf{\Sigma}})$, the first eigenvalue, and (8) is established following simple algebra.

Appendix III
Proof of Theorem 2

Without loss of generality, consider a diagonal matrix, $\mathbf{\Sigma} = \operatorname{diag}(\lambda_1(\mathbf{\Sigma}), \ldots, \lambda_N(\mathbf{\Sigma})) = \mathbf{\Sigma}' + \mathbf{\Delta}$, where we define $\mathbf{\Sigma}' = \operatorname{diag}(\lambda_1(\mathbf{\Sigma}), \lambda_N(\mathbf{\Sigma}), \ldots, \lambda_N(\mathbf{\Sigma}))$ and $\mathbf{\Delta} = \operatorname{diag}(0, \delta_2, \delta_3, \ldots, \delta_{N-1}, 0)$. We have

$$\widetilde{\mathbf{\Sigma}} = \mathbf{P}^T\mathbf{\Sigma}\mathbf{P} = \mathbf{P}^T\mathbf{\Sigma}'\mathbf{P} + \mathbf{P}^T\mathbf{\Delta}\mathbf{P}. \quad (44)$$

From Theorem 1, we have that $\mathbf{P}^T\mathbf{w}_1$ is the first eigenvector of $\widetilde{\mathbf{\Sigma}}' = \mathbf{P}^T\mathbf{\Sigma}'\mathbf{P}$. From Proposition 3, we have then

$$\sin\angle(\tilde{\mathbf{u}}_1, \mathbf{P}^T\mathbf{w}_1) \le \frac{4\lambda_1(\mathbf{P}^T\mathbf{\Delta}\mathbf{P})}{\lambda_1(\mathbf{P}^T\mathbf{\Sigma}'\mathbf{P}) - \lambda_2(\mathbf{P}^T\mathbf{\Sigma}'\mathbf{P})}. \quad (45)$$

However, we have from Lemma 2 that

$$\lambda_1(\mathbf{P}^T\mathbf{\Delta}\mathbf{P}) \le \operatorname{trace}(\mathbf{\Delta}) = \delta. \quad (46)$$

From (42), we have $\lambda_2(\mathbf{P}^T\mathbf{\Sigma}'\mathbf{P}) = \lambda_N(\mathbf{\Sigma}')$, while Theorem 1 determines the first eigenvalue to be $\lambda_1(\mathbf{P}^T\mathbf{\Sigma}'\mathbf{P}) = \delta_1\|\mathbf{P}^T\mathbf{w}_1\|_2^2 + \lambda_N(\mathbf{\Sigma}')$. Thus, the denominator of (45) becomes

$$\lambda_1(\mathbf{P}^T\mathbf{\Sigma}'\mathbf{P}) - \lambda_2(\mathbf{P}^T\mathbf{\Sigma}'\mathbf{P}) = \delta_1\|\mathbf{P}^T\mathbf{w}_1\|_2^2. \quad (47)$$

Combining (45), (46), and (47) yield (10), the desired result. Note that, for Proposition 3 to apply here, we need

$$\lambda_1(\mathbf{P}^T\mathbf{\Delta}\mathbf{P}) \le \frac{1}{5}\left(\lambda_1(\mathbf{P}^T\mathbf{\Sigma}'\mathbf{P}) - \lambda_2(\mathbf{P}^T\mathbf{\Sigma}'\mathbf{P})\right). \quad (48)$$

However, if (9) is true, then, from (46) and (47), so is (48).

Appendix IV
Proof of Theorem 3

To create a random projection matrix $\mathbf{P}$, let us use the following procedure. Populate $\mathbf{G}'$ as an $N \times N$ matrix of independent, identically distributed, zero-mean, unit-variance Gaussian random variables, and partition $\mathbf{G}'$ into $N \times K$ matrix $\mathbf{P}'$ and $N \times (N-K)$ matrix $\mathbf{R}'$ as $\mathbf{G}' = [\mathbf{P}' \, \mathbf{R}']$. Create an orthonormal $\mathbf{G} = [\mathbf{P} \, \mathbf{R}]$ from $\mathbf{G}'$ by orthogonalizing its rows; i.e., normalize the first row of $\mathbf{G}'$ and orthogonalize the remaining rows with respect to the first via a Gram–Schmidt procedure. This procedure will result in $\mathbf{P}$ (and $\mathbf{R}$) having orthonormal columns. To prove Theorem 3, we first prove the following lemma.

Lemma 3: For random orthonormal matrix $\mathbf{P}$ formed via the procedure outlined above, $\|\bar{\mathbf{p}}_1\|_2^2$ has a beta distribution

$$\|\bar{\mathbf{p}}_1\|_2^2 \sim \beta\left(\frac{K}{2}, \frac{N-K}{2}\right) \quad (49)$$

with mean $E[\|\bar{\mathbf{p}}_1\|_2^2] = K/N$, where $\bar{\mathbf{p}}_1^T$ is the first row of $\mathbf{P}$.

Proof: $\|\bar{\mathbf{p}}'_1\|_2^2$ is the sum of the squares of K unit-variance Gaussian random variables; it, thus, has a chi-square distribution with K degrees of freedom [39], $\|\bar{\mathbf{p}}'_1\|_2^2 \sim \chi_K^2$. Likewise, $\|\bar{\mathbf{r}}'_1\|_2^2 \sim \chi_{N-K}^2$. After normalization, $\|\bar{\mathbf{p}}_1\|_2^2$ has a beta distribution with parameters $K/2$ and $(N-K)/2$ since

$$\|\bar{\mathbf{p}}_1\|_2^2 = \frac{\|\bar{\mathbf{p}}'_1\|_2^2}{\|\bar{\mathbf{p}}'_1\|_2^2 + \|\bar{\mathbf{r}}'_1\|_2^2} \sim \beta\left(\frac{K}{2}, \frac{N-K}{2}\right) \quad (50)$$

(see Sec. 25.2 of [40]). As for the mean, we note that it is known that if random variable $X \sim \beta(a, b)$, then $E[X] = a/(a+b)$. ■

Now, to prove Theorem 3, we again assume without loss of generality the case of a diagonal covariance matrix such that $\mathbf{P}^T\mathbf{w}_1 = \bar{\mathbf{p}}_1$. It is known that, if random variable $X \sim \beta(a, b)$, then $E[1/X] = (a + b - 1)/(a - 1)$ [40]. Thus, from (49), $E[\|\bar{\mathbf{p}}_1\|_2^{-2}] = (N - 2)/(K - 2)$ which yields (11).

Appendix V
Justification of Approximation 1

Given $\mathbf{\Sigma}$ with first eigenpair $(\lambda_1(\mathbf{\Sigma}), \mathbf{w}_1)$, the usual approach for determining pairs $(\lambda_n(\mathbf{\Sigma}), \mathbf{w}_n)$, for $n \geq 2$ is to successively apply some form of *deflation* (see [6, Chap. 5]). For example, the first eigenpair $(\lambda_1(\mathbf{\Sigma}'), \mathbf{w}_1')$ of

$$\mathbf{\Sigma}' = \mathbf{\Sigma} - \lambda_1(\mathbf{\Sigma})\mathbf{w}_1\mathbf{w}_1^T \tag{51}$$

is $(\lambda_2(\mathbf{\Sigma}), \mathbf{w}_2)$. Applying $\mathbf{P}$ to this deflated matrix $\mathbf{\Sigma}'$ yields

$$\begin{aligned}\widetilde{\mathbf{\Sigma}}' &= \mathbf{P}^T\mathbf{\Sigma}'\mathbf{P} \\ &= \mathbf{P}^T\mathbf{\Sigma}\mathbf{P} - \lambda_1(\mathbf{\Sigma})\mathbf{P}^T\mathbf{w}_1\mathbf{w}_1^T\mathbf{P} \\ &\approx \widetilde{\mathbf{\Sigma}} - \lambda_1(\mathbf{\Sigma})\|\mathbf{P}^T\mathbf{w}_1\|_2^2\widetilde{\mathbf{u}}_1\widetilde{\mathbf{u}}_1^T \end{aligned} \tag{52}$$

where the approximation stems from (12). However, (13) yields

$$\begin{aligned}\lambda_1(\mathbf{\Sigma})\|\mathbf{P}^T\mathbf{w}_1\|_2^2 &\approx \lambda_1(\widetilde{\mathbf{\Sigma}}) - \left(1 - \|\mathbf{P}^T\mathbf{w}\|_2^2\right)\lambda_N(\mathbf{\Sigma}) \\ &\approx \lambda_1(\widetilde{\mathbf{\Sigma}})\end{aligned} \tag{53}$$

where the second approximation assumes[5] that $\lambda_1(\widetilde{\mathbf{\Sigma}}) \gg \lambda_N(\mathbf{\Sigma})$. From (52), we have then

$$\widetilde{\mathbf{\Sigma}}' \approx \widetilde{\mathbf{\Sigma}} - \lambda_1(\widetilde{\mathbf{\Sigma}})\widetilde{\mathbf{u}}_1\widetilde{\mathbf{u}}_1^T \tag{54}$$

the right side of which is the deflation of $\widetilde{\mathbf{\Sigma}}'$. Theorem 2 bounds the angle between $\mathbf{P}^T\mathbf{w}_1'$ and $\widetilde{\mathbf{u}}_1'$, the first eigenvector of $\widetilde{\mathbf{\Sigma}}'$. But, since $\mathbf{w}_1' = \mathbf{w}_2$ and $\widetilde{\mathbf{u}}_1' = \widetilde{\mathbf{u}}_2$, (14) holds if $\angle(\mathbf{P}^T\mathbf{w}_1', \widetilde{\mathbf{u}}_1') = \angle(\mathbf{P}^T\mathbf{w}_2, \widetilde{\mathbf{u}}_2)$ is small, which Theorem 2 argues will be the case if $\lambda_2(\mathbf{\Sigma}) \gg \lambda_3(\mathbf{\Sigma})$.

References

[1] Q. Du and J. E. Fowler, “Hyperspectral image compression using JPEG2000 and principal component analysis,” *IEEE Geosci. Remote Sens. Lett.*, vol. 4, no. 2, pp. 201–205, Apr. 2007.

[2] B. Penna, T. Tillo, E. Magli, and G. Olmo, “A new low complexity KLT for lossy hyperspectral data compression,” in *Proc. Int. Geoscience and Remote Sensing Symp.*, Denver, CO, Aug. 2006, vol. 7, pp. 3525–3528.

[3] J. C. Harsanyi and C.-I. Chang, “Hyperspectral image classification and dimensionality reduction: An orthogonal subspace projection approach,” *IEEE Trans. Geosci. Remote Sens.*, vol. 32, no. 4, pp. 779–785, Jul. 1994.

[4] J. B. Lee, S. Woodyatt, and M. Berman, “Enhancement of high spectral resolution remote-sensing data by a noise-adjusted principal components transform,” *IEEE Trans. Geosci. Remote Sens.*, vol. 28, no. 3, pp. 295–304, May 1990.

[5] X. Jia and J. A. Richards, “Segmented principal components transformation for efficient hyperspectral remote-sensing image display and classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 1, pp. 538–542, Jan. 1999.

[6] B. N. Parlett*, The Symmetric Eigenvalue Problem*. Philadelphia, PA: SIAM, 1998.

[7] E. Candès, J. Romberg, and T. Tao, “Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information,” *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[8] E. Candès and T. Tao, “Near-optimal signal recovery from random projections: Universal encoding strategies?,” *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, Dec. 2006.

[9] D. L. Donoho, “Compressed sensing,” *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[10] E. J. Candès and M. B. Wakin, “An introduction to compressive sampling,” *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[11] V. K. Goyal, A. K. Fletcher, and S. Rangan, “Compressive sampling and lossy compression,” *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 48–56, Mar. 2008.

[12] J. E. Fowler, “Compressive-projection principal component analysis for the compression of hyperspectral signatures,” in *Proc. IEEE Data Compression Conf.*, J. A. Storer and M. W. Marcellin, Eds., Snowbird, UT, Mar. 2008, pp. 83–92.

[13] J. E. Fowler, “Compressive-projection principal component analysis and the first eigenvector,” in *Proc. IEEE Data Compression Conf.*, J. A. Storer and M. W. Marcellin, Eds., Snowbird, UT, Mar. 2009, pp. 223–232.

[14] Z. Jia and G. W. Stewart, “An analysis of the Rayleigh–Ritz method for approximating eigenspaces,” *Math. Comput.*, vol. 70, no. 234, pp. 637–647, Apr. 2001.

[15] I. T. Jolliffe*, Principal Component Analysis*. New York: Springer-Verlag, 1986.

[16] J. E. Jackson*, A User’s Guide to Principal Component Analysis*. New York: Wiley, 1991.

[17] R. B. Cattell, “The scree test for the number of factors,” *Multivariate Behav. Res.*, vol. 1, no. 2, pp. 245–276, Apr. 1966.

[18] J. M. Craddock and C. R. Flood, “Eigenvectors for representing the 500 mb geopotential surface over the northern hemisphere,” *Quart. J. Roy. Meteorol. Soc.*, vol. 95, no. 405, pp. 576–593, Jul. 1969.

[19] S. A. Farmer, “An investigation into the results of principal component analysis of data derived from random numbers,” *The Statistician*, vol. 20, no. 4, pp. 63–72, Dec. 1971.

[20] P. L. Combettes, “The foundations of set theoretic estimation,” *Proc. IEEE*, vol. 81, no. 2, pp. 182–208, Feb. 1993.

[21] Q. Du and J. E. Fowler, “Low-complexity principal component analysis for hyperspectral image compression,” *Int. J. High Performance Comput. Appl.*, vol. 22, no. 4, pp. 438–448, Nov. 2008.

[22] A. Gersho and R. M. Gray*, Vector Quantization and Signal Compression*. Norwell, MA: Kluwer, 1992.

[23] G. H. Golub and C. F. Van Loan*, Matrix Computations*, 3rd ed. Baltimore, MD: Johns Hopkins Univ. Press, 1996.

[24] S. Ji, D. Dunson, and L. Carin, “Multitask compressive sensing,” *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 92–106, Jan. 2009.

[25] D. Baron, M. B. Wakin, M. F. Duarte, S. Savrvotham, and R. G. Baranuik, Distributed Compressed Sensing 2005, submitted.

[26] M. Fornasier and H. Rauhut, “Recovery algorithms for vector-valued data with joint sparsity constraints,” *SIAM J. Numer. Anal.*, vol. 46, no. 2, pp. 577–613, 2008.

[27] M. Mishali and Y. C. Eldar, “Reduce and boost: Recovering arbitrary sets of jointly sparse vectors,” *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 4692–4702, Oct. 2008.

[28] D. P. Wipf and B. D. Rao, “An empirical Bayesian strategy for solving the simultaneous sparse approximation problem,” *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 3704–3716, Jul. 2007.

[29] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, “Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit,” *Signal Process.*, vol. 86, no. 3, pp. 572–588, March 2006.

[30] J. A. Tropp, “Algorithms for simultaneous sparse approximation. Part II: Convex relaxation,” *Signal Process.*, vol. 86, no. 3, pp. 589–602, Mar. 2006.

[31] W. B. Johnson and J. Lindenstrauss, “Extensions of Lipschitz mappings into a Hilbert space,” *Contemporary Math.*, vol. 26, pp. 189–206, 1984.

[32] S. Dasgupta and A. Gupta, “An elementary proof of a theorem of Johnson and Lindenstrauss,” *Random Structures and Algorithms*, vol. 22, no. 1, pp. 60–65, Jan. 2003.

[33] D. Achlioptas, “Database-friendly random projections: Johnson–Lindenstrauss with binary coins,” *J. Comput. Syst. Sci.*, vol. 66, no. 4, pp. 671–687, June 2003.

[5] If $\lambda_1(\mathbf{\Sigma}) \gg \lambda_2(\mathbf{\Sigma})$ as assumed initially, then δ_1 in (8) is large and $\lambda_1(\widetilde{\mathbf{\Sigma}}) \gg \lambda_N(\mathbf{\Sigma})$ provided $\delta_1\|\mathbf{P}^T\mathbf{w}_1\|_2^2$ is also large. This will typically be the case, since Lemma 3 indicates $E[\|\mathbf{P}^T\mathbf{w}_1\|_2^2] = K/N$, and we usually have K/N falling in the approximate range (0.2, 0.5) in practice.

[34] R. M. Willett, M. E. Gehm, and D. J. Brady, “Multiscale reconstruction for computational spectral imaging,” in *Computational Imaging V*, C. A. Bouman, E. L. Miller, and I. Pollak, Eds. San Jose, CA: SPIE, Jan. 2007, vol. 6498, p. 64980L.

[35] N. P. Pitsianis, D. J. Brady, A. Portnoy, X. Sun, T. Suleski, M. A. Fiddy, M. R. Feldman, and R. D. TeKolste, “Compressive imaging sensors,” in *Intelligent Integrated Microsystems*, R. A. Athale and J. C. Zolper, Eds. Kissimmee, FL: SPIE, Apr. 2006, vol. 6232, p. 62320A.

[36] D. J. Brady, “Micro-optics and megapixels,” *Opt. Photon. News*, vol. 17, no. 11, pp. 24–29, Nov. 2006.

[37] I. Johnstone and A. Y. Lu, “Sparse principal component analysis,” *J. Amer. Statist. Assoc.*, to appear.

[38] G. Strang, *Linear Algebra and Its Applications*, 3rd ed. San Diego, CA: Harcourt Brace, 1988.

[39] N. K. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, 2nd ed. New York: Wiley, 1994, vol. 1.

[40] N. K. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, 2nd ed. New York: Wiley, 1995, vol. 2.

James E. Fowler (S’91–M’96–SM’02) received the B.S. degree in computer and information science engineering and the M.S. and Ph.D. degrees in electrical engineering in 1990, 1992, and 1996, respectively, all from The Ohio State University, Columbus.

In 1995, he was an intern researcher at AT&T Labs in Holmdel, NJ, and, in 1997, he held an NSF-sponsored postdoctoral assignment at the Universié de Nice-Sophia Antipolis, France. In 2004, he was a Visiting Professor in the Département TSI at École Nationale Supérieure des Télécommunications (ENST), Paris, France. He is currently a Professor in the Department of Electrical and Computer Engineering, Mississippi State University, Starkville, and is also a researcher in the Geosystems Research Institute (GRI), Mississippi State.

Dr. Fowler is an Associate Editor of the IEEE Transactions on Multimedia, the IEEE Signal Processing Letters, and the *EURASIP Journal of Image & Video Processing*. He served as a Guest Editor for the special issue on “Wavelets in Source Coding, Communications, and Networks” which appeared in the *EURASIP Journal of Image & Video Processing* in January 2007. He is a member of the Image, Video, and Multidimensional Signal Processing Technical Committee of the IEEE Signal Processing Society, as well as a member of the program committee for the Data Compression Conference.