Compressive-Projection Principal Component Analysis and the First Eigenvector

James E. Fowler

Department of Electrical and Computer Engineering
Geosystems Research Institute (GRI)
Mississippi State University, Starkville, Mississippi

Abstract

An analysis is presented that extends existing Rayleigh-Ritz theory to the special case of highly eccentric distributions. Specifically, a bound on the angle between the first Ritz vector and the orthonormal projection of the first eigenvector is developed for the case of a random projection onto a lower-dimensional subspace. It is shown that this bound is expected to be small if the eigenvalues are widely separated, i.e., if the data distribution is highly eccentric. This analysis verifies the validity of a fundamental approximation behind compressive projection principal component analysis, a technique proposed previously to recover from random projections not only the coefficients associated with principal component analysis but also an approximation to the principal-component transform basis itself.

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; 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 fact, if dimensionality reduction could be directly integrated into the signal sensing and acquisition process, then one could avoid not only the computational burden of explicit dimensionality reduction, but also the production of onerous quantities of data in the first place.

In [1, 2], we presented *compressive-projection principal component analysis* (CPPCA), a technique which effectively shifts the computational burden of PCA from a resource-constrained encoder to the decoder which presumably resides on a significantly more powerful "base-station" system. On the encoder side, CPPCA is driven by projections at the 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

In *Proceedings of the IEEE Data Compression Conference*, J. A. Storer and M. W. Marcellin, Eds., Snowbird, UT, March 2009, pp. 223-232

transform, but also an approximation to the PCA transform basis itself. In its reliance on encoder-side random projections, CPPCA permits integrated acquisition and dimensionality reduction directly within the hardware of the sensor. This process of random projection effectively eliminates explicit computation of dimensionality reduction at the encoder; in addition, since data is sensed directly in a reduced dimensionality, the dataset never exists in its full signal resolution at any point within the sensor, dramatically reducing memory requirements. The computational and memory burdens are instead shifted to the CPPCA decoder which consists of a novel eigenvector-reconstruction process based on a projections-onto-convex-sets (POCS) optimization driven by Ritz vectors extracted from the projected subspaces. 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.

At the core of the CPPCA decoder is an approximation that uses Ritz vectors as close representations of orthonormal projections of eigenvectors. In [1], this approximation is presented as a heuristic whose validity is supported by experimental evidence. In this paper, we strengthen the argument by providing an analytical derivation of an upper bound on the error of this approximation. Below, we first overview CPPCA before presenting the details of our analysis.

An Overview of CPPCA

At the core of the CPPCA technique is a decoder-side process that produces an approximation to the PCA transform basis. Specifically, consider a dataset of M zero-mean vectors $\mathbf{X} = \begin{bmatrix}\mathbf{x}_1 & \cdots & \mathbf{x}_M\end{bmatrix}$, where each $\mathbf{x}_m \in \mathbb{R}^N$. The covariance matrix of $\mathbf{X}$ is $\mathbf{\Sigma} = \mathbf{X}\mathbf{X}^T/M$, and the PCA transform matrix is the $N \times N$ matrix $\mathbf{W}$ of eigenvectors that 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. However, central to the CPPCA paradigm is that production of the PCA transform matrix occurs at the decoder rather than at the encoder as in the traditional use of PCA; that is, the CPPCA decoder cannot implement eigendecomposition (1) directly as it does not know either $\mathbf{X}$ or $\mathbf{\Sigma}$. Instead, the decoder knows only K-dimensional projections of $\mathbf{X}$. Specifically, suppose we have K orthonormal vectors $\mathbf{p}_k$ that form the basis of K-dimensional subspace $\mathcal{P}$ such that $\mathbf{P} = \begin{bmatrix}\mathbf{p}_1 & \cdots & \mathbf{p}_K\end{bmatrix}$ provides an orthogonal projection onto $\mathcal{P}$. The CPPCA encoder produces $\widetilde{\mathbf{Y}} = \mathbf{P}^T\mathbf{X}$, and it is from projections $\widetilde{\mathbf{Y}}$ that the CPPCA decoder approximates $\mathbf{W}$. The projected vectors have covariance

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

which the CPPCA decoder calculates having received $\widetilde{\mathbf{Y}}$ from the encoder.

Rayleigh-Ritz theory [3] describes the relation between the eigenvectors of $\mathbf{\Sigma}$ and those of $\widetilde{\mathbf{\Sigma}}$ as given by (2). Covariance matrix $\mathbf{\Sigma}$ has spectrum $\lambda(\mathbf{\Sigma}) = \{\lambda_1(\mathbf{\Sigma}), \ldots, \lambda_N(\mathbf{\Sigma})\}$, where the eigenvalues satisfy $\lambda_1(\mathbf{\Sigma}) \geq \cdots \geq \lambda_N(\mathbf{\Sigma})$, and the corresponding unit eigenvectors are $\mathbf{w}_n$. 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 $\widetilde{\mathbf{U}} =$

In *Proceedings of the IEEE Data Compression Conference*, J. A. Storer and M. W. Marcellin, Eds., Snowbird, UT, March 2009, pp. 223-232

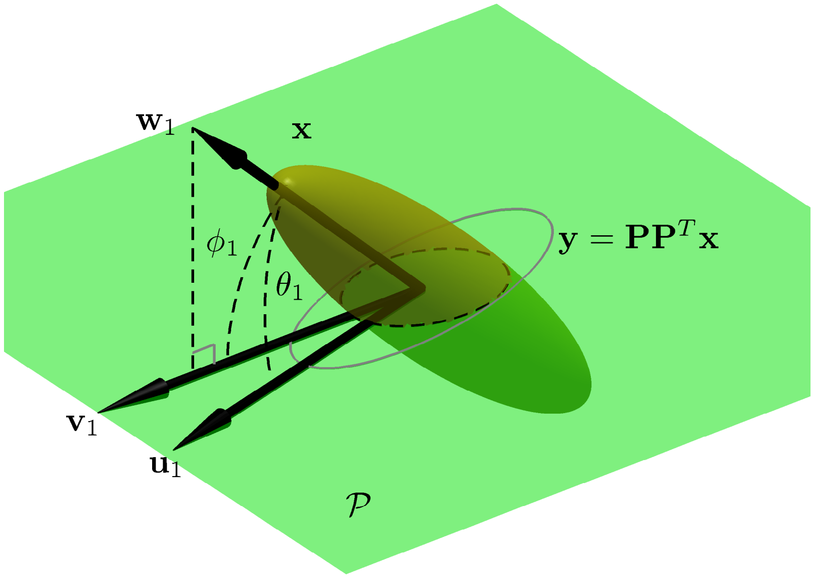

Figure 1: Data distribution of $\mathbf{x}$ in $\mathbb{R}^3$ is projected onto 2D subspace $\mathcal{P}$ as $\mathbf{y}$; the first Ritz vector, $\mathbf{u}_1$, lies close to the normalized projection, $\mathbf{v}_1$, onto $\mathcal{P}$ of the first eigenvector, $\mathbf{w}_1$, of $\mathbf{x}$ (from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

$\big[\widetilde{\mathbf{u}}_1 \quad \cdots \quad \widetilde{\mathbf{u}}_K\big]$, $\widetilde{\mathbf{\Lambda}} = \operatorname{diag}\big(\lambda_1\big(\widetilde{\mathbf{\Sigma}}\big), \ldots, \lambda_K\big(\widetilde{\mathbf{\Sigma}}\big)\big)$, $\|\widetilde{\mathbf{u}}_k\|_2 = 1$, and $\lambda_1\big(\widetilde{\mathbf{\Sigma}}\big) \geq \cdots \geq \lambda_K\big(\widetilde{\mathbf{\Sigma}}\big)$. The K eigenvalues $\lambda_k\big(\widetilde{\mathbf{\Sigma}}\big)$ 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 $\widetilde{\mathbf{\Sigma}}$. 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.

Existing Rayleigh-Ritz theory is rather limited in that it tells us very little about the Ritz vectors for $K < N$. We know only that the Ritz vectors do not typically align with the orthogonal projections of any of the eigenvectors [3]; i.e., $\mathbf{u}_k \neq \mathbf{v}_n$ in general. However, CPPCA is built on the central idea that, if subspace $\mathcal{P}$ is chosen randomly, and the distribution of the vectors in $\mathbf{X}$ is highly eccentric in that eigenvalue $\lambda_k\big(\mathbf{\Sigma}\big)$ 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 $\lambda_k\big(\widetilde{\mathbf{\Sigma}}\big)$. Under the assumption that $\mathbf{u}_k \approx \mathbf{v}_k$, a POCS-based algorithm was devised in [1] to approximate the first L eigenvectors $\mathbf{w}_n$ from $\widetilde{\mathbf{Y}}$; the reader is referred to [1] for the specific details of this process. Suffice it to say, however, that the entire feasibility of the CPPCA technique rests on the approximation $\mathbf{u}_k \approx \mathbf{v}_k$. Empirical evidence was presented in [1] that argued for the validity of this approximation. Below, however, we analyze this approximation for the specific case of the first eigenvector; i.e., we derive an upper bound on the separation between $\mathbf{u}_1$ and $\mathbf{v}_1$. This analysis, which is presented in the next section, effectively extends existing Rayleigh-Ritz theory for the first eigenvector in the special case of random projection of an eccentric distribution, thereby verifying the validity of a fundamental aspect of the CPPCA technique.

Analysis of the First Eigenvector

The Single-Spike Covariance

For our main results, we conduct our analysis under the assumption of a diagonal covariance matrix of the form $\mathbf{\Sigma} = \operatorname{diag}\big(\lambda_1(\mathbf{\Sigma}), \ldots, \lambda_N(\mathbf{\Sigma})\big)$ where $\lambda_1(\mathbf{\Sigma}) \geq \cdots \geq \lambda_N(\mathbf{\Sigma})$ 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 non-diagonal $\mathbf{\Sigma}$. The quantities of interest in the following theorems—vector lengths and angles between vectors—are invariant to such unitary rotations.

To begin, let $\mathbf{P}$ be an $N \times K$ orthonormal projection matrix such that $\mathbf{P}^T\mathbf{P} = \mathbf{I}$. Denote the columns of $\mathbf{P}$ as $\mathbf{p}_k$ and the rows as $\bar{\mathbf{p}}_n^T$ such that $\mathbf{P} = \begin{bmatrix}\mathbf{p}_1 & \cdots & \mathbf{p}_K\end{bmatrix} = \begin{bmatrix}\bar{\mathbf{p}}_1 & \cdots & \bar{\mathbf{p}}_N\end{bmatrix}^T$. Let $N \times (N-K)$ matrix $\mathbf{R} = \operatorname{null}(\mathbf{P}^T)$; 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 $\mathbf{G} = \begin{bmatrix}\mathbf{P} & \mathbf{R}\end{bmatrix}$. Let the columns and rows of $\mathbf{G}$ be $\mathbf{g}_n$ and $\bar{\mathbf{g}}_n$, respectively, and note that

$$\left\|\mathbf{g}_n\right\|_2^2 = \left\|\bar{\mathbf{g}}_n\right\|_2^2 = \left\|\bar{\mathbf{p}}_n\right\|_2^2 + \left\|\bar{\mathbf{r}}_n\right\|_2^2 = 1. \tag{5}$$

Define $N \times N$ matrix $\mathbf{H} = \mathbf{R}\mathbf{R}^T$, and note that the diagonal elements of $\mathbf{H}$ are $h_{nn} = \bar{\mathbf{r}}_n^T\bar{\mathbf{r}}_n = \left\|\bar{\mathbf{r}}_n\right\|_2^2$.

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 $\lambda_1(\mathbf{\Sigma}) > \lambda_N(\mathbf{\Sigma}) = \cdots = \lambda_N(\mathbf{\Sigma}) > 0$. 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 $\widetilde{\mathbf{u}}_1 = \mathbf{P}^T\mathbf{w}_1 / \left\|\mathbf{P}^T\mathbf{w}_1\right\|_2$. The corresponding first eigenvalue is*

$$\lambda_1(\widetilde{\mathbf{\Sigma}}) = \delta_1\left\|\mathbf{P}^T\mathbf{w}_1\right\|_2^2 + \lambda_N(\mathbf{\Sigma}), \tag{6}$$

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

Proof. As we argue above, without loss of generality, consider the case of a diagonal $\mathbf{\Sigma}$ with first eigenvector

$$\mathbf{w}_1 = \begin{bmatrix}1 & 0 & \cdots & 0\end{bmatrix}^T. \tag{7}$$

Define matrix $\mathbf{\Delta}$ as

$$\mathbf{\Delta} = \mathbf{\Sigma} - \lambda_N(\mathbf{\Sigma})\mathbf{I} = \operatorname{diag}(\delta_1, 0, \ldots, 0), \tag{8}$$

and form an $N \times N$ orthonormal matrix from $\mathbf{P}$ by concatenating additional orthonormal columns; i.e., let $\mathbf{G} = \begin{bmatrix}\mathbf{P} & \mathbf{R}\end{bmatrix}$ where $\mathbf{R} = \operatorname{null}(\mathbf{P}^T)$. Define also $\mathbf{H} = \mathbf{R}\mathbf{R}^T$, and note that $\mathbf{P}^T\mathbf{w}_1 = \bar{\mathbf{p}}_1 \neq \mathbf{0}$ implies $\left\|\bar{\mathbf{p}}_1\right\|_2^2 > 0$, and

$$h_{11} = \left\|\bar{\mathbf{r}}_1\right\|_2^2 = 1 - \left\|\bar{\mathbf{p}}_1\right\|_2^2 < 1 \tag{9}$$

In *Proceedings of the IEEE Data Compression Conference*, J. A. Storer and M. W. Marcellin, Eds., Snowbird, UT, March 2009, pp. 223-232

from (5).

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

$$\mathbf{G}^{-1}\mathbf{\Sigma}\mathbf{G} = \mathbf{G}^T\mathbf{\Sigma}\mathbf{G} = \begin{bmatrix}\mathbf{P}^T \\ \mathbf{R}^T\end{bmatrix} \mathbf{\Sigma} \begin{bmatrix}\mathbf{P} & \mathbf{R}\end{bmatrix} = \begin{bmatrix}\widetilde{\mathbf{\Sigma}} & \mathbf{P}^T\mathbf{\Sigma}\mathbf{R} \\ \mathbf{R}^T\mathbf{\Sigma}\mathbf{P} & \mathbf{R}^T\mathbf{\Sigma}\mathbf{R}\end{bmatrix}. \tag{10}$$

Since $\big(\lambda_1(\mathbf{\Sigma}), \mathbf{w}_1\big)$ is an eigenpair of $\mathbf{\Sigma}$, $\big(\lambda_1(\mathbf{\Sigma}), \mathbf{G}^T\mathbf{w}_1\big)$ is an eigenpair of $\mathbf{G}^T\mathbf{\Sigma}\mathbf{G}$ due to the inherent nature of a similarity transform (Theorem 5P of [4]); 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

$$\begin{bmatrix}\widetilde{\mathbf{\Sigma}} & \mathbf{P}^T\mathbf{\Sigma}\mathbf{R} \\ \mathbf{R}^T\mathbf{\Sigma}\mathbf{P} & \mathbf{R}^T\mathbf{\Sigma}\mathbf{R}\end{bmatrix} \begin{bmatrix}\mathbf{P}^T \\ \mathbf{R}^T\end{bmatrix} \mathbf{w}_1 = \lambda_1(\mathbf{\Sigma}) \begin{bmatrix}\mathbf{P}^T \\ \mathbf{R}^T\end{bmatrix} \mathbf{w}_1. \tag{11}$$

Considering only the first row of (11), and noting that $\mathbf{P}^T\mathbf{H} = \mathbf{0}$, we have

$$\begin{aligned}\lambda_1(\mathbf{\Sigma})\mathbf{P}^T\mathbf{w}_1 &= \big(\widetilde{\mathbf{\Sigma}}\mathbf{P}^T + \mathbf{P}^T\mathbf{\Sigma}\mathbf{H}\big)\mathbf{w}_1 = \Big(\widetilde{\mathbf{\Sigma}}\mathbf{P}^T + \mathbf{P}^T\big(\mathbf{\Delta} + \lambda_N(\mathbf{\Sigma})\mathbf{I}\big)\mathbf{H}\Big)\mathbf{w}_1 \\ &= \big(\widetilde{\mathbf{\Sigma}}\mathbf{P}^T + \mathbf{P}^T\mathbf{\Delta}\mathbf{H}\big)\mathbf{w}_1 \end{aligned} \tag{12}$$

from the definition of $\mathbf{\Delta}$ in (8). From (7) and (8), 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, (12) becomes

$$\lambda_1(\mathbf{\Sigma})\mathbf{P}^T\mathbf{w}_1 = \widetilde{\mathbf{\Sigma}}\mathbf{P}^T\mathbf{w}_1 + \delta_1 h_{11}\mathbf{P}^T\mathbf{w}_1, \tag{13}$$

or

$$\widetilde{\mathbf{\Sigma}}\mathbf{P}^T\mathbf{w}_1 = \Big(\lambda_1(\mathbf{\Sigma}) - \delta_1 h_{11}\Big)\mathbf{P}^T\mathbf{w}_1. \tag{14}$$

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 $\mathbf{P}^T\mathbf{w}_1/\big\|\mathbf{P}^T\mathbf{w}_1\big\|_2$ is actually its first eigenvector. We note that

$$\widetilde{\mathbf{\Sigma}} = \mathbf{P}^T\mathbf{\Sigma}\mathbf{P} = \mathbf{P}^T\Big(\mathbf{\Delta} + \lambda_N(\mathbf{\Sigma})\mathbf{I}\Big)\mathbf{P} = \mathbf{P}^T\mathbf{\Delta}\mathbf{P} + \lambda_N(\mathbf{\Sigma})\mathbf{I} \tag{15}$$

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 Theorem 4 in the appendix become equalities, and we have

$$\begin{aligned}\lambda_k(\widetilde{\mathbf{\Sigma}}) &= \lambda_k\big(\mathbf{P}^T\mathbf{\Delta}\mathbf{P} + \lambda_N(\mathbf{\Sigma})\mathbf{I}\big) = \lambda_k\big(\mathbf{P}^T\mathbf{\Delta}\mathbf{P}\big) + \lambda_N(\mathbf{\Sigma}) \\ &= \begin{cases}\lambda_1\big(\mathbf{P}^T\mathbf{\Delta}\mathbf{P}\big) + \lambda_N(\mathbf{\Sigma}), & k = 1, \\ \lambda_N(\mathbf{\Sigma}), & 2 \le k \le K,\end{cases}\end{aligned} \tag{16}$$

where we note that $\mathbf{\Delta}$ is a rank-1 diagonal matrix and invoke Theorem 6 in the appendix to reveal that $\lambda_k\big(\mathbf{P}^T\mathbf{\Delta}\mathbf{P}\big) = 0$ for $k > 1$. From (9), $0 \le \delta_1 h_{11} < \delta_1$, and the eigenvalue in question, $\lambda_1(\mathbf{\Sigma}) - \delta_1 h_{11}$, satisfies

$$\lambda_1(\mathbf{\Sigma}) - \delta_1 h_{11} > \lambda_1(\mathbf{\Sigma}) - \delta_1 = \lambda_N(\mathbf{\Sigma}). \tag{17}$$

Thus, $\lambda_1(\mathbf{\Sigma}) - \delta_1 h_{11}$ must be $\lambda_1(\widetilde{\mathbf{\Sigma}})$, the first eigenvalue, and (6) is established following simple algebra. □

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, $\mathbf{u}_1 = \mathbf{P}^T\widetilde{\mathbf{u}}_1$, 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.

The General Eccentric Covariance

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 $\mathbf{v}_1$ using the perturbation-theory result of Theorem 5 in the appendix. 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 $\lambda_1(\mathbf{\Sigma}) \geq \lambda_2(\mathbf{\Sigma}) \geq \cdots \geq \lambda_N(\mathbf{\Sigma})$, 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*

$$\delta \leq \frac{\delta_1}{5}\left\|\mathbf{P}^T\mathbf{w}_1\right\|_2^2, \tag{18}$$

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

$$\sin\omega_1 \leq \frac{4\delta}{\delta_1\left\|\mathbf{P}^T\mathbf{w}_1\right\|_2^2}, \tag{19}$$

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

Proof. As before and without loss of generality, let us consider a diagonal matrix, $\mathbf{\Sigma} = \operatorname{diag}(\lambda_1(\mathbf{\Sigma}), \ldots, \lambda_N(\mathbf{\Sigma}))$. We then express $\mathbf{\Sigma}$ as $\mathbf{\Sigma} = \mathbf{\Sigma}' + \mathbf{\Delta}$, 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

$$\widetilde{\mathbf{\Sigma}} = \mathbf{P}^T\mathbf{\Sigma}\mathbf{P} = \mathbf{P}^T\mathbf{\Sigma}'\mathbf{P} + \mathbf{P}^T\mathbf{\Delta}\mathbf{P}. \tag{20}$$

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 Theorem 5 in the appendix, we have then

$$\sin\angle(\widetilde{\mathbf{u}}_1, \mathbf{P}^T\mathbf{w}_1) \leq \frac{4\lambda_1(\mathbf{P}^T\mathbf{\Delta}\mathbf{P})}{\lambda_1(\mathbf{P}^T\mathbf{\Sigma}'\mathbf{P}) - \lambda_2(\mathbf{P}^T\mathbf{\Sigma}'\mathbf{P})}. \tag{21}$$

However, we have from Theorem 7 in the appendix that

$$\lambda_1(\mathbf{P}^T\mathbf{\Delta}\mathbf{P}) \leq \operatorname{trace}(\mathbf{\Delta}) = \delta. \tag{22}$$

From (16), we have $\lambda_2(\mathbf{P}^T\mathbf{\Sigma}'\mathbf{P}) = \lambda_N(\mathbf{\Sigma}')$, while Theorem 1 determines the first eigenvalue to be $\lambda_1(\mathbf{P}^T\mathbf{\Sigma}'\mathbf{P}) = \delta_1\left\|\mathbf{P}^T\mathbf{w}_1\right\|_2^2 + \lambda_N(\mathbf{\Sigma}')$. Thus, the denominator of (21) becomes

$$\lambda_1(\mathbf{P}^T\mathbf{\Sigma}'\mathbf{P}) - \lambda_2(\mathbf{P}^T\mathbf{\Sigma}'\mathbf{P}) = \delta_1\left\|\mathbf{P}^T\mathbf{w}_1\right\|_2^2. \tag{23}$$

Combining (21), (22), and (23) yield (19), the desired result. Note that, for Theorem 5 to apply here, we need

$$\lambda_1\big(\mathbf{P}^T\boldsymbol{\Delta}\mathbf{P}\big) \leq \frac{1}{5}\Big(\lambda_1\big(\mathbf{P}^T\boldsymbol{\Sigma}'\mathbf{P}\big) - \lambda_2\big(\mathbf{P}^T\boldsymbol{\Sigma}'\mathbf{P}\big)\Big). \tag{24}$$

However, if (18) is true, then, from (22) and (23), so is (24). □

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 distribution, 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 $\mathbf{G}' = \begin{bmatrix}\mathbf{P}' & \mathbf{R}'\end{bmatrix}$. Create an orthonormal $\mathbf{G} = \begin{bmatrix}\mathbf{P} & \mathbf{R}\end{bmatrix}$ 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.

Theorem 3. *For random orthonormal matrix $\mathbf{P}$ formed via the procedure outlined above, $\left\|\mathbf{P}^T\mathbf{w}_1\right\|_2^2$ has a beta distribution; thus, for a fixed $\boldsymbol{\Sigma}$ and therefore fixed δ and δ_1 as defined in Theorem 2, the expected value of the upper bound in* (19) *is*

$$E\left[\frac{4\delta}{\delta_1\left\|\mathbf{P}^T\mathbf{w}_1\right\|_2^2}\right] = \frac{4\delta\,(N-2)}{\delta_1\,(K-2)}. \tag{25}$$

Proof. $\left\|\bar{\mathbf{p}}_1'\right\|_2^2$ 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 [5]; i.e., $\left\|\bar{\mathbf{p}}_1'\right\|_2^2 \sim \chi_K^2$. Likewise, $\left\|\bar{\mathbf{r}}_1'\right\|_2^2 \sim \chi_{N-K}^2$. After normalization, $\left\|\bar{\mathbf{p}}_1\right\|_2^2$ has a beta distribution with parameters $K/2$ and $(N-K)/2$ since

$$\left\|\bar{\mathbf{p}}_1\right\|_2^2 = \frac{\left\|\bar{\mathbf{p}}_1'\right\|_2^2}{\left\|\bar{\mathbf{p}}_1'\right\|_2^2 + \left\|\bar{\mathbf{r}}_1'\right\|_2^2} \sim \beta\left(\frac{K}{2}, \frac{N-K}{2}\right) \tag{26}$$

(see Sec. 25.2 of [6]). Again, without loss of generality, assume 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 $E[1/X] = (a+b-1)/(a-1)$ [6]. Thus, $E\left[\left\|\bar{\mathbf{p}}_1\right\|_2^{-2}\right] = (N-2)/(K-2)$ which yields (25). □

We note that, if $\lambda_1\big(\boldsymbol{\Sigma}\big) \gg \lambda_2\big(\boldsymbol{\Sigma}\big) \geq \lambda_N\big(\boldsymbol{\Sigma}\big)$, then δ_1 will be large. Assuming further that $\delta_1 \gg \delta$, (25) 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 $\widetilde{\boldsymbol{\Sigma}} = \mathbf{P}^T\boldsymbol{\Sigma}\mathbf{P}$ 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, $\mathbf{v}_1$.

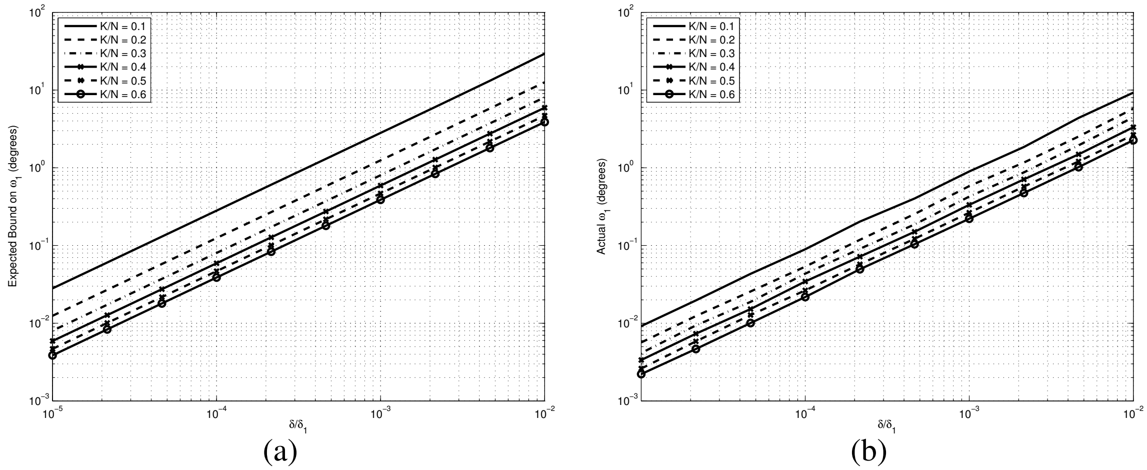

Figure 2: (a) The expected bound on ω_1 as given by (25) of Theorem 3 (for $N = 100$). (b) Experimental evaluation of ω_1 for $\mathbf{\Sigma} = \mathbf{W}\mathbf{\Lambda}\mathbf{W}^T$ with $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, 1, 1, \ldots, 1)$ and $N = 100$.

Experimental Evaluation

To experimentally evaluate the ramifications of Theorems 2 and 3, consider Fig. 2. In Fig. 2(a), we plot the bound of (25) for varying values of K/N and δ/δ_1. Fig. 2(a) 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 $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, 1, 1, \ldots, 1)$, 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 projection matrices $\mathbf{P}$ using the procedure outlined above 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. 2(b). We observe that the behavior of the curves in Fig. 2(b) is fairly similar to that of the curves in Fig. 2(a), 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 Theorem 7 in the appendix is not particularly tight, resulting in some looseness to the bound in Theorem 2. In any event, though, Figs. 2(a) and (b) 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.

Conclusions

The CPPCA technique originally proposed in [1] couples random projections at the encoder with a Rayleigh-Ritz process for approximating eigenvectors at the decoder. In this paper, we presented an analysis that establishes the validity of the Ritz-vector approximation at the heart of the CPPCA eigenvector-reconstruction process. Our analysis considers the relation between Ritz vectors and normalized projections of eigenvectors, providing a bound on the angle between the first Ritz vector and the first normalized projection. It was observed that this bound is expected to be small under the conditions of widely sep-

In *Proceedings of the IEEE Data Compression Conference*, J. A. Storer and M. W. Marcellin, Eds., Snowbird, UT, March 2009, pp. 223-232

arated eigenvalues (i.e., a highly eccentric data distribution) as well as randomly selected subspace projections. This analysis provides a solid theoretical foundation for the CPPCA algorithm proposed in [1] which features a POCS-based optimization driven by Ritz vectors that approximates the eigenvectors constituting the PCA transform. This in turn permits the CPPCA decoder, given only the random projections created by the encoder, to recover not only the coefficients associated with the PCA transform, but also an approximation to the PCA transform basis itself. We note that, while we consider only the first eigenvector here, [2] further extends the analysis to the subsequent eigenvectors using a deflation argument (see Chap. 5 of [3]).

Appendix

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 $\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \cdots \lambda_N(\mathbf{A})$, where $\lambda_n(\mathbf{A})$ is the n^{th} largest eigenvalue of $\mathbf{A}$. The following results from [7] characterize how the eigendecomposition of $\mathbf{A}$ is affected by a symmetric perturbation $\mathbf{E}$.

Theorem 4. *If $\mathbf{A}$ and $\mathbf{E}$ are symmetric $N \times N$ matrices,*

$$\lambda_n(\mathbf{A}) + \lambda_N(\mathbf{E}) \leq \lambda_n(\mathbf{A} + \mathbf{E}) \leq \lambda_n(\mathbf{A}) + \lambda_1(\mathbf{E}), \tag{27}$$

for $1 \leq n \leq N$.

Proof. This is Theorem 8.1.5 of [7]. □

Theorem 5. *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*

$$\sin \angle(\mathbf{w}_1, \mathbf{w}'_1) \leq \frac{4}{\delta}\lambda_1(\mathbf{E}), \tag{28}$$

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 App. A.1 of [8] which in turn derives from Theorem 8.1.12 of [7]. □

Theorem 6. *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, $\mathbf{A}_n = \operatorname{diag}(\ldots, 0, a_n, 0, \ldots)$. 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. We note that $\operatorname{rank}(\mathbf{P}^T\mathbf{A}_n\mathbf{P}) \leq \min\{\operatorname{rank}(\mathbf{P}^T), \operatorname{rank}(\mathbf{A}_n), \operatorname{rank}(\mathbf{P})\}$. Because $\operatorname{rank}(\mathbf{A}_n) = 1$, we have then $\operatorname{rank}(\mathbf{P}^T\mathbf{A}_n\mathbf{P}) \leq 1$. Let $\mathbf{A}'_n = \operatorname{diag}(\ldots, 0, \sqrt{a_n}, 0, \ldots)$, factor $\mathbf{P}^T\mathbf{A}_n\mathbf{P}$ as $\mathbf{P}^T\mathbf{A}_n\mathbf{P} = (\mathbf{P}^T{\mathbf{A}'_n}^T)(\mathbf{A}'_n\mathbf{P})$, and then know that $\mathbf{P}^T\mathbf{A}_n\mathbf{P}$ is positive semidefinite ([4], 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, $\mathbf{P}^T\mathbf{A}_n\mathbf{P} = \begin{bmatrix}\bar{\mathbf{p}}_1 & \cdots & \bar{\mathbf{p}}_N\end{bmatrix}\begin{bmatrix}\cdots & 0 & a_n\bar{\mathbf{p}}_n & 0 & \cdots\end{bmatrix}^T$, and along the diagonal we have $\operatorname{diag}\left(\mathbf{P}^T\mathbf{A}_n\mathbf{P}\right) = \left(a_n\bar{p}_{n1}^2,\, a_n\bar{p}_{n2}^2,\, \ldots,\, a_n\bar{p}_{nK}^2\right)$, where $\bar{\mathbf{p}}_n = \begin{bmatrix}\bar{p}_{n1} & \cdots & \bar{p}_{nK}\end{bmatrix}^T$. For a rank-1 matrix, the trace is equal to the first eigenvalue; thus,

$$\lambda_1\left(\mathbf{P}^T\mathbf{A}_n\mathbf{P}\right) = \operatorname{trace}\left(\mathbf{P}^T\mathbf{A}_n\mathbf{P}\right) = a_n\left\|\bar{\mathbf{p}}_n\right\|_2^2 \le a_n, \tag{29}$$

where the inequality is a consequence of (5). □

Theorem 7. *Consider $N \times N$ positive-semidefinite diagonal matrix $\mathbf{A}$. Express $\mathbf{A}$ as $\mathbf{A} = \operatorname{diag}\left(a_1, a_2, \ldots, a_N\right) = \sum_{n=1}^{N}\mathbf{A}_n$, where $\mathbf{A}_n$ is as defined in Theorem 6 in the case $a_n \ne 0$ and $\mathbf{A}_n = \mathbf{0}$ otherwise. For $N \times K$ orthonormal matrix $\mathbf{P}$,*

$$\lambda_1\left(\mathbf{P}^T\mathbf{A}\mathbf{P}\right) \le \operatorname{trace}\left(\mathbf{A}\right). \tag{30}$$

Proof. The first eigenvalue of $\mathbf{P}^T\mathbf{A}\mathbf{P}$ satisfies

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

where the first inequality is due to the right side of (27) in Theorem 4 and the second is due to Theorem 6. Applying this result recursively to the rightmost term yields $\lambda_1\left(\mathbf{P}^T\mathbf{A}\mathbf{P}\right) \le a_1 + a_2 + \lambda_1\left(\sum_{n=3}^{N}\mathbf{P}^T\mathbf{A}_n\mathbf{P}\right) \le \cdots \le \sum_{n=1}^{N}a_n = \operatorname{trace}\left(\mathbf{A}\right)$. □

References

[1] J. E. Fowler, "Compressive-projection principal component analysis for the compression of hyperspectral signatures," in *Proceedings of the IEEE Data Compression Conference*, J. A. Storer and M. W. Marcellin, Eds., Snowbird, UT, March 2008, pp. 83–92.

[2] ——, "Compressive-projection principal component analysis," *IEEE Transactions on Image Processing*, submitted.

[3] B. N. Parlett, *The Symmetric Eigenvalue Problem*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 1998.

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

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

[6] ——, *Continuous Univariate Distributions*, 2nd ed. New York: John Wiley & Sons, Inc., 1995, vol. 2.

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

[8] I. Johnstone and A. Y. Lu, "Sparse principal component analysis," *Journal of the American Statistical Association*, to appear.