SINGULAR VECTOR PERTURBATION UNDER GAUSSIAN NOISE

Rongrong Wang

Department of Mathematics, University of Maryland, College Park, MD, 20742, USA

Abstract

We perform a non-asymptotic analysis on the singular vector distribution under Gaussian noise. In particular, we provide sufficient conditions of a matrix for its first few singular vectors to have near normal distribution. Our result can be used to facilitate the error analysis in PCA.

1 INTRODUCTION

The singular value decomposition (SVD) lies in the heart of various dimension reduction (DR) techniques, such as PCA, LE, LLE etc. In real world problems, noise from unknown sources may largely change the dimension reduction result by changing either the principal directions, or the data distribution along those directions. Aside from some unusual cases that involve nonlinear bias in the measurements, and because of the central limit theorem, it is always assumed that the noise vector in the raw data obeys the i.i.d. Gaussian distribution, which simplifies both data processing and error analysis. However, when doing dimension reductions, the Gaussian distribution is not preserved. This is because the low dimensional data is contained in the first few singular vectors of some predefined kernel, and elements of any singular vector are bounded by one, which prevent them from being Gaussian variables. In [2], it is pointed out in a statistical and asymptotic way that for a fixed data matrix, as the noise level goes to 0, the distribution of singular vectors would eventually become very close to a multivariate normal distribution. In this paper, we go further in this direction and provide non-asymptotic bound on the distance between the perturbed singular vectors and a multi-variate normal random variable. In applications, our result could be used in two ways:

• To predict for a fixed data matrix and noise level, whether or not can one reliably (within a given confidence interval) assume that the noise after dimension reduction is Gaussian.

• For a given noise level, to determine how many samples are necessary for the noise in the embedded space to be close (e.g., the distance is less than some given threshold) to Gaussian.
We state the mathematical description of our problem as follows.

**Problem:** Suppose \( Y \) is an \( n \times n \) noiseless data matrix, and \( \tilde{Y} = Y + \epsilon W \) is the observed noisy data. Their associated SVDs are:

\[
Y = (U_1, U_2) \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} (V_1, V_2)^T,
\]

and

\[
\tilde{Y} = (\tilde{U}_1, \tilde{U}_2) \begin{pmatrix} \tilde{\Sigma}_1 & 0 \\ 0 & \tilde{\Sigma}_2 \end{pmatrix} (\tilde{V}_1, \tilde{V}_2)^T,
\]

where the entries of \( W \) are i.i.d. \( N(0,1) \) and \( \epsilon > 0 \) is an absolute constant. Due to the randomness of \( W \), all the quantities in (2) are random variables whose measures are induced by those of \( W \). Under this setting, what is the distribution of \( \tilde{U}_1 - U_1 \)?

In this paper, without loss of generality, we focus only on the low rank and square \( Y \)s (i.e., \( \Sigma_2 = 0 \)). All our techniques can be carried through in high rank and rectangular cases.

The perturbation problem of singular vectors has brought great attention in the field of statistics and numerical analysis. Davis and Kahan \[3\] studied the deterministic perturbation on Hermitain matrices and provides an upper bound for the rotation of singular vector subspaces caused by the perturbation. They showed that the span of a group of singular vectors as a subspace is changed by an amount which is proportional to the noise power and the reciprocal of eigenvalue gap. (the change is charaterized in canonical angle between subspaces, denoted by \( \angle(U_1, \tilde{U}_1) \)). Wedin \[12\] extended this result to non-Hermitain cases. Dopico \[5\] proved that the left and right singular vectors is perturbed towards the same direction. Van Vu \[11\] considered random perturbations (i.i.d. Bernoulli distribution) and provided a tighter upper bound for the canonical angles \( \angle(U_1, \tilde{U}_1) \), which hold with large probability. The necessity of using the canonical angles to characterize the change is well illustrated in the following example in \[11\]. Consider the following two matrices,

\[
Y = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \tilde{Y} = \begin{pmatrix} 1 & \epsilon \\ \epsilon & 1 \end{pmatrix}.
\]

Both matrices are diagonalizable, the eigenvectors of \( Y \) are \((1,0)\) and \((0,1)\), while those of \( \tilde{Y} \) are \((1/\sqrt{2}, 1/\sqrt{2})\) and \((1/\sqrt{2}, -1/\sqrt{2})\), for any \( \epsilon \). Hence, the difference of these two bases does not go to zero with \( \epsilon \).

Beside the canonical angle, some authors used another quantity: \( \min_M \| \tilde{U}_1 - U_1 M \|_F \) to describe the changes of singular subspaces. The latter has been proved to be an upper bound of the former.

In this paper, we use the pointwise matrix norm \( \| \cdot \|_{\text{max}} \) to bound the difference between two matrices. Our technique is valid and simpler for the Frobenius norm case so we omit it. However, if one tries to use the bound we provide on pointwise norm to derive a bound on Frobenius norm using the equivalence of norms, he will get a looser bound than directly running over our technique on Frobenius norm, and vice versa.
The rest of the paper is organized as follows. In Section 2, we introduce notations and some existing theorems to be used. In Section 3, we state several lemmas and our main theorem. Section 4 contains the proof of the main theorem. In Section 5, we apply our result to an audio signal classification problem.

2 PREVIOUS RESULTS

Throughout this paper, we consider only square data matrices with nontrivial dimensionality \( n \geq 2 \) and rank \( k < n \). The variables \( Y, \tilde{Y}, \Sigma_i, \tilde{\Sigma}_i, U_i, \tilde{U}_i, V_i, \tilde{V}_i, i = 1, 2 \), remain the same as defined in (1) and (2). We assume \( \Sigma_2 = 0 \). In addition, we assume that all the diagonal elements of the \( k \times k \) matrix \( \Sigma_1 \) are bounded away from zero.

For a given matrix \( A \), we use \( QA \) and \( RA \) to denote the QR decomposition of \( A \). The set of singular values of \( A \) is denoted by \( \sigma(A) \); the \( k \)th largest one is denoted by \( \sigma_k(A) \); and the minimum singular value is by \( \sigma_{\text{min}}(A) \). We use \( W \) to denote the normalized Gaussian matrix whose entries are \( N(0, 1/n) \). For any matrix, \( \| \cdot \|_F \) denotes the Frobenius norm, and \( \| \cdot \|_2 \) the spectral norm. In addition, we use \( \| \cdot \|_{\text{max}} \) to denote the component-wise maximum norm of a matrix, i.e., \( \| A \|_{\text{max}} = \max_{i,j} |A_{i,j}| \).

The following theorem is due to Dopico [5], who provided the worst-case bound on the Frobenius norm of the deviation of singular vector subspace under perturbation. He proved that this bound is proportional to the Frobenius norm of the perturbation as well as the reciprocal of the eigenvalue gap.

**Theorem 1** ([5]). Let \( Y \) and \( \tilde{Y} \) and their SVDs be as defined in (1) and (2). Define

\[
\delta = \min \left\{ \min_{\mu \in \sigma(\Sigma_1), \tilde{\mu} \in \sigma(\tilde{\Sigma}_1)} |\tilde{\mu} - \mu|, \sigma_{\text{min}}(\Sigma_1) + \sigma_{\text{min}}(\tilde{\Sigma}_1) \right\}.
\]

If \( \delta > 0 \), then

\[
\min_{M \text{ unitary}} \sqrt{\|U_1 M - \tilde{U}_1\|_F^2 + \|V_1 M - \tilde{V}_1\|_F^2} \leq \frac{\sqrt{\|R\|_F^2 + \|S\|_F^2}}{\delta},
\]

where \( R = (Y - \tilde{Y})\tilde{V}_1, S = (Y - \tilde{Y})^T\tilde{U}_1 \). Moreover, the left hand side of (3) is minimized for \( M = Z_{1SZ_{2}}^T \) where \( Z_{1SZ_{2}}^T \) is any SVD of \( U_1^T \tilde{U}_1 + V_1^T \tilde{V}_1 \), and in this case, the equality can be attained.

In the proof of our main theorem (Theorem 4), we will frequently encounter the spectral norm of Gaussian matrices, which is known to be bounded in the following theorem.

**Theorem 2** ([10]). Let \( W \) be an \( n \times k \) matrix with i.i.d. normal entries with mean \( 0 \) and variance \( 1/n \). Then, its largest and smallest singular values obey:

\[
P(\lambda_k(W) > 1 + \sqrt{k/n} + \gamma) \leq e^{-n \gamma^2 / 2},
\]

In order to estimate the gap of eigenvalues in (3), we will make use of the following result.
Theorem 3. \((\ref{thm3})\) Let \(A\) be an \(n \times k\) matrix and \(A+E\) be a perturbation of \(A\). Let \(\lambda_i\) and \(\tilde{\lambda}_i\) be the \(i\)th largest eigenvalue of \(A\) and \(A+E\), respectively. Then, for \(i = 1, \ldots, \min\{n,k\}\),

\[
|\lambda_i - \tilde{\lambda}_i| \leq \|E\|_2.
\]

3 MAIN THEOREM

We now ready to state our main theorem.

Theorem 4. Let \(Y\) and \(\tilde{Y}\) and their SVDs be defined as in (1) and (2). Assume \(\epsilon, \frac{1}{n}\) are small enough such that the following defined quantities satisfies \(E_2 > 1/(2\sqrt{k})\) and \(\delta_1 > 0\) (see below). Then with probability (with respect to the random Gaussian noise \(W\)) exceeding \(1 - 3e^{-(n-k)\beta + \ln(k(n+k))} - 4e^{-n\gamma^2/2}\), there exists an unitary \(k \times k\) matrix \(M\), such that for \(0 < \beta < 1/2\), we have

\[
\|(\tilde{U}_1 - U_1M) - \epsilon N\|_{\max} \leq \frac{\sqrt{2k}E_1}{\delta_1(1 - E_1)} + \sqrt{2k}(\|U_1\|_{\max} + E_4)(\sqrt{2} + 1) + \frac{\sqrt{k}E_2}{1 - \sqrt{k}E_2} + E_3, \quad (4)
\]

where \(N\) is a Gaussian matrix defined by \(N = U_2U_2^TWV_1\Sigma_1^{-1}\), and where

\[
E_1(\epsilon, \Sigma, k, n, \gamma) = \epsilon^3\|\Sigma^{-1}\|_2^2(\alpha_2^2\alpha_3 + 2\alpha_1\alpha_2 + 2\alpha_1\Sigma^{-1}) + \epsilon^4\|\Sigma^{-1}\|_2^2(\alpha_1^2\alpha_2^2 + 2\alpha_1^2\alpha_2\alpha_3) + \epsilon^5\|\Sigma^{-1}\|_2^2(\alpha_1^2\alpha_2^2\alpha_3),
\]

\[
E_2(\epsilon, \Sigma, k, n, \gamma) = \epsilon^2\|\Sigma^{-1}\|_2^2\alpha_1^2 + 2\epsilon^3\|\Sigma^{-1}\|_2^2\alpha_1^2\alpha_2 + \epsilon^4\|\Sigma^{-1}\|_2^2\alpha_1^2\alpha_3 + \epsilon^5\|\Sigma^{-1}\|_2^2\alpha_1^2\alpha_2^2,
\]

\[
E_3(\epsilon, \Sigma, k, n, \gamma) = \epsilon(1 + k)(\frac{2}{n^{1/2}})(n - k)^{-\frac{1}{2} + \beta}\|\Sigma^{-1}\|_2^2,
\]

\[
E_4(\epsilon, \Sigma, k, n, \gamma) = \epsilon\alpha_1\|\Sigma^{-1}\|_2^2 + \epsilon^2\|\Sigma^{-1}\|_2^2\alpha_1\alpha_3,
\]

\[
\delta_1 = \sigma_{\min}(\Sigma_1) - (2 + \gamma)\epsilon.
\]

are defined later in formulas (19), (20), (25), (29), (17), and where \(\alpha_1 = 1 + \gamma + \sqrt{k/n}\), \(\alpha_2 = 2 + 2\gamma + 2\sqrt{k/n}\), and \(\alpha_3 = 2 + \gamma\).

Note that the order of \(E_1, \ldots, E_4\) in terms of \(\epsilon\) and \(n\) are: \(E_1 = O(\epsilon^3), E_2 = O(\epsilon^2), E_3 = O(\epsilon^2n^{-\frac{1}{2} + \beta}), E_4 = O(\epsilon)\) and \(\delta = O(1)\).

Corollary 1. If the eigenvalues of \(Y\) are all different, then (4) becomes

\[
\|(\tilde{U}_1 - U_1) - \epsilon N\|_{\max} \leq \frac{\sqrt{2k}E_1}{\delta_2(1 - E_2)} + \sqrt{2k}(\|U_1\|_{\max} + E_4)(\sqrt{2} + 1) + \frac{\sqrt{k}E_2}{1 - \sqrt{k}E_2} + E_3,
\]

where \(\delta_2 = \min\{\min\{\sigma_i(\Sigma_1) - \sigma_j(\Sigma_1)\}, i, j = 1, \ldots, k, i \neq j\}, \sigma_{\min}(\Sigma_1)\} - 2(2 + \gamma)\epsilon.

Remark 1: We observe that the perturbation \(\tilde{U}_1 - U_1\) can be approximated by a Gaussian term \(\epsilon N\) only when it is the leading term in the error. The average magnitude of this term has asymptotic order of \(O(\epsilon/\sqrt{n})\) as \(\epsilon\) and \(\frac{1}{n}\) going to 0, while the order of the leading terms on the right hand side is either \(\epsilon^2\|U_1\|_{\max}\) or \(\epsilon^2n^{-\frac{1}{2} + \beta}\). Therefore, to
ensure the Gaussian term is higher in order, we need the following condition on the \((\epsilon, n)\) pair:

\[
\epsilon = o \left( \min \left\{ n^{-\beta}, \frac{1}{\|U_i\|_{\max} n^2} \right\} \right). \tag{5}
\]

Before going into the proof, we first establish three useful lemmas. The first one is an elementary observation in linear algebra, and the last one is a direct application of Theorem 2, so we omit their proofs.

**Lemma 1.** If \(A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}\), then \(\|A\|_2 \leq \sqrt{2} \max \left\{ \sqrt{\|A_{11}\|^2 + \|A_{21}\|^2}, \sqrt{\|A_{12}\|^2 + \|A_{22}\|^2} \right\}\).

**Lemma 2.** Let \(F\) be the distribution of the product of two independent \(N(0, 1)\) normal random variables. Let \(X_1, X_2, \ldots, X_n\) be i.i.d. random variables drawn from \(F\). If \(n \geq 2\), we have

\[P(\bar{X} > n^{-\frac{1}{2}} + \beta) \leq 2e^{-n^\beta}, \quad \text{for any } \beta \in \mathbb{R},\]

where \(\bar{X} = \left( \sum_{i=1}^{n} X_i \right) / n\).

**Proof.** If \(X_i \sim F\), one can verify that for all \(0 \leq \theta < 1\),

\[E e^{\theta X_i} = \sqrt{\frac{1}{1 - \theta^2}}.\]

We apply Markov’s inequality (see eg. [8]) to have:

\[P(\bar{X} \geq n^{-\frac{1}{2}} + \beta) = P(e^{\bar{X}} \geq e^{n^{-\frac{1}{2}} + \beta}) \leq \frac{E e^{t\bar{X}}}{e^{nt^{-\frac{1}{2}} + \beta}}, \quad \text{for any } t \in \mathbb{R}.\]

Letting \(t = n^{1/2}\) in the above formula, we get

\[P(\bar{X} \geq n^{-\frac{1}{2}} + \beta) \leq \frac{E e^{nt^{1/2} X}}{e^{n^\beta}} = e^{-n^\beta} \prod_{i=1}^{n} E(n^{-1/2} X_i)\]

\[= e^{-n^\beta} (1 - \frac{1}{n})^{-n/2} \leq e^{-n^\beta} (1 - \frac{1}{2})^{-2/2} = 2e^{-n^\beta}.\]

\[
\Box
\]

**Lemma 3.** Let \(T\) be an \(n \times n\) Gaussian matrix whose elements are i.i.d. \(N(0, \frac{1}{n})\). Let \(T\) be written as

\[T = \begin{pmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{pmatrix},\]

with \(T_{11}\) a \(k \times k\) matrix. Then, with probability exceeding \(1 - 4e^{-(n-k)\gamma^2/2}\),

\[\|T_{11}\|_2 \leq \gamma + 2\sqrt{\frac{k}{n}}, \quad \|T_{21}\|_2, \|T_{12}\|_2 \leq 1 + \gamma + \sqrt{\frac{k}{n}}, \quad \|T\|_2, \|T_{22}\|_2 \leq 2 + \gamma.\]
4 PROOF OF THE MAIN THEOREM

Proof. The idea of the proof is the following: Our goal is to characterize the perturbed singular vector space $\hat{U}_1$, which is a left orthogonal matrix and satisfies $\hat{U}_1^T\hat{Y}\hat{V}_1 = \hat{\Sigma}_1$, meaning that it, together with $\hat{V}_1$, diagonalizes $\hat{Y}$. Thus, if we can construct two other orthogonal matrices which also diagonalize $\hat{Y}$ with small error, then, by Theorem 1, they are expected to be good approximations to $\hat{U}_1$ and $\hat{V}_1$.

We start the proof similarly to that in [2]. Define the random matrix $C$ as follows,

$$ U^T\hat{Y}V = \Sigma + \epsilon U^T W V = \Sigma + \epsilon C, \quad (6) $$

where $\Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix}$ and $C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} = \begin{pmatrix} U_1^T W V_1 & U_2^T W V_2 \\ U_2^T W V_1 & U_2^T W V_2 \end{pmatrix}$. Due to the invariant property of Gaussian matrix, $C$ has the same distribution as $W$. The asymptotic order of $\epsilon$ in the error term $\epsilon C$ is not satisfactory, we thus want to further diagonalize $C$ by using the following two matrices,

$$ P = I + \epsilon \begin{pmatrix} 0 & -\Sigma_1^{-1} C_{21}^T \\ \Sigma_1^{-1} C_{21} & 0 \end{pmatrix} + \epsilon^2 \begin{pmatrix} 0 & B^T \\ -B & 0 \end{pmatrix}, $$

with $B = -C_{22} C_{12}^T \Sigma_1^{-2} + C_{21} \Sigma_1^{-1} C_{11} \Sigma_1^{-1}$, and

$$ O = I + \epsilon \begin{pmatrix} 0 & -\Sigma_1^{-1} C_{12} \\ \Sigma_1^{-1} C_{12} & 0 \end{pmatrix} + \epsilon^2 \begin{pmatrix} 0 & D \\ -D^T & 0 \end{pmatrix}, $$

with $D = -\Sigma_1^{-2} C_{12} C_{22} + \Sigma_1^{-1} C_{11} \Sigma_1^{-1} C_{12}$. Note that in doing so, we start to differ from [2] by defining $P$ and $O$ explicitly. The second order terms in $P$ and $O$ are crucial for obtaining (4).

Keeping in mind that $P$ and $O$ are not yet unitary, we multiply the left hand side of (6) by $P^T$ on the left, and by $O$ on the right, to obtain,

$$ P^T U^T \hat{Y} V O = \left( I + \epsilon \begin{pmatrix} 0 & -\Sigma_1^{-1} C_{21}^T \\ \Sigma_1^{-1} C_{21} & 0 \end{pmatrix} + \epsilon^2 \begin{pmatrix} 0 & B^T \\ -B & 0 \end{pmatrix} \right)^T X \left( \Sigma + \epsilon \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} \right) \left( I + \epsilon \begin{pmatrix} 0 & -\Sigma_1^{-1} C_{12} \\ \Sigma_1^{-1} C_{12} & 0 \end{pmatrix} + \epsilon^2 \begin{pmatrix} 0 & D \\ -D^T & 0 \end{pmatrix} \right) $$

$$ = \left( \Sigma_1 + \epsilon C_{11} + \epsilon^2 (\Sigma_1^{-1} C_{21}^T C_{21} + C_{12} C_{12}^T \Sigma_1^{-1}) \right) \left( 0 \right) + \epsilon \left( \epsilon C_{22} - 2 \epsilon^2 C_{21} \Sigma_1^{-1} C_{12} \right) + \text{Err}, \quad (7) $$

where the matrix Err includes all the terms whose order on $\epsilon$ is greater than or equal to 3, so Err is $O(\epsilon^3)$. Compare (7) with (6), we see the above operation has changed the order of error term from $O(\epsilon)$ to $O(\epsilon^3)$. If we denote the eigenvector subspaces corresponding to the two diagonal blocks by $UP = ((UP)_1, (UP)_2)$, $VO = ((VO)_1, (VO)_2)$, then they have the following expressions,

$$ (UP)_1 = U_1 + \epsilon U_2 C_{21} \Sigma_1^{-1} - \epsilon^2 U_2 B, \quad (8) $$

$$ (VO)_1 = V_1 + \epsilon V_2 C_{12} \Sigma_1^{-1} - \epsilon^2 V_2 D^T. \quad (9) $$
Recall that $Q_A$ and $R_A$ denote the QR decomposition of a matrix $A$. Now, we want to orthogonalize $UP$ and $VO$. For that purpose, for both sides of (7), we multiply them by
\[
\begin{pmatrix}
R_{(UP)1}^{-T} & 0 \\
0 & R_{(UP)2}^{-T}
\end{pmatrix}
\] on the left, and
\[
\begin{pmatrix}
R_{(VO)1}^{-1} & 0 \\
0 & R_{(VO)2}^{-1}
\end{pmatrix}
\] on the right. This way we obtain:
\[
\begin{pmatrix}
R_{(UP)1}^{-T} & 0 \\
0 & R_{(UP)2}^{-T}
\end{pmatrix}
P^T U^T \tilde{Y} VO \begin{pmatrix}
R_{(VO)1}^{-1} & 0 \\
0 & R_{(VO)2}^{-1}
\end{pmatrix} = \begin{pmatrix} L_1 & 0 \\ 0 & L_2 \end{pmatrix} + Err \tag{10}
\]
where
\[
\begin{align*}
L_1 &= R_{(UP)1}^{-T} (\Sigma_1 + \epsilon C_{11} + \epsilon^2 (\Sigma_1^{-1} C_{21}^T C_{21} + C_{12} C_{12}^T \Sigma_1^{-1})) R_{(VO)1}^{-1}, \\
L_2 &= R_{(UP)2}^{-T} (\epsilon C_{22} - \epsilon^2 C_{21} \Sigma_1^{-1} C_{12}) R_{(VO)2}^{-1}.
\end{align*}
\]
With a little abuse of notation, we continue to use $Err$ to denote the error term in (10), though it was already changed by the left and right multiplication. We denote by $M_i \hat{\Sigma}_i M_i'$ the SVD of the $L_i$, with $i = 1, 2$. In this case, (10) becomes
\[
\begin{pmatrix}
R_{(UP)1}^{-T} & 0 \\
0 & R_{(UP)2}^{-T}
\end{pmatrix}
P^T U^T \tilde{Y} VO \begin{pmatrix}
R_{(VO)1}^{-1} & 0 \\
0 & R_{(VO)2}^{-1}
\end{pmatrix} = \begin{pmatrix} M_1 \hat{\Sigma}_1 M_1' & 0 \\ 0 & M_2 \hat{\Sigma}_2 M_2' \end{pmatrix} + Err. \tag{11}
\]
On the other hand, the left hand side of (11) also satisfies
\[
\begin{pmatrix}
R_{(UP)1}^{-T} & 0 \\
0 & R_{(UP)2}^{-T}
\end{pmatrix}
P^T U^T \tilde{Y} VO \begin{pmatrix}
R_{(VO)1}^{-1} & 0 \\
0 & R_{(VO)2}^{-1}
\end{pmatrix} = \begin{pmatrix}
R_{(UP)1}^{-T} (UP)_{11}^T & 0 \\
0 & R_{(UP)2}^{-T} (UP)_{22}^T
\end{pmatrix} Y ((VO)_{11} R_{(VO)1}^{-1} (VO)_{21} R_{(VO)2}^{-1})
\]
\[
= \begin{pmatrix}
Q_{(UP)1} & 0 \\
0 & Q_{(UP)2}
\end{pmatrix}^T \tilde{Y} \begin{pmatrix} Q_{(VO)1} & Q_{(VO)2} \end{pmatrix}. \tag{12}
\]
Combining (11) with (12), we obtain:
\[
\begin{pmatrix}
Q_{(UP)1} & Q_{(UP)2}
\end{pmatrix}^T \tilde{Y} \begin{pmatrix} Q_{(VO)1} & Q_{(VO)2} \end{pmatrix} = \begin{pmatrix}
M_1 \hat{\Sigma}_1 M_1' & 0 \\
0 & M_2 \hat{\Sigma}_2 M_2'
\end{pmatrix} + Err.
\]
Moving everything but $Y$ on the left hand side to the right by multiplying the inverse of each matrix and utilizing the fact that $(UP)_{1}$ and $(UP)_{2}$ are orthogonal to each other, we derive:
\[
\tilde{Y} = \begin{pmatrix}
Q_{(UP)1} & Q_{(UP)2}
\end{pmatrix} M_1 \hat{\Sigma}_1 M_1' \begin{pmatrix} 0 & 0 \\
0 & \hat{\Sigma}_2
\end{pmatrix} \begin{pmatrix} Q_{(VO)1} & Q_{(VO)2}
\end{pmatrix} M_1' M_2 \hat{\Sigma}_2 M_2' + Err. \tag{13}
\]
(13) combined with Theorem 1 imply that $\tilde{U}_1$ and $Q_{(UP)1} M_1$ are the first $k$ left singular vectors of two very similar matrices (different by the $Err$ term) and that they are close. Keeping this useful result in mind, we first turn to look at the big picture.
Our final goal is to approximate by a Gaussian variable the difference between $U_1$ and $\tilde{U}_1$ up to a rotation: $\tilde{U}_1 - U_1 M$ (we will define the unitary matrix $M$ explicitly later), which can be decomposed as:

$$\tilde{U}_1 - U_1 M = (\tilde{U}_1 - Q_{(UP)_1}M) + (Q_{(UP)_1}M - (UP)_1M) + ((UP)_1M - U_1 M).$$

(14)

We insert (8) into (14) to get

$$\tilde{U}_1 - U_1 M = (\tilde{U}_1 - Q_{(UP)_1}M) + (Q_{(UP)_1}M - (UP)_1M) + (\epsilon U_2 C_21 \Sigma_1^{-1} - \epsilon^2 U_2 B)M.$$  

Here, the Gaussian term is the second to last term on the right, so we want to prove all other terms are small. For that purpose, we move the Gaussian term to the left and take the component-wise matrix norm on both sides, to have:

$$\|\tilde{U}_1 - U_1 M - \epsilon U_2 C_21 \Sigma_1^{-1}M\|_{\text{max}} \leq \|\tilde{U}_1 - Q_{(UP)_1}M\|_{\text{max}} + \epsilon^2 \|U_2 B M\|_{\text{max}} + \|Q_{(UP)_1}M - (UP)_1M\|_{\text{max}}$$

(15)

$$= I + II + III.$$  

Observe that the left hand side of (15) is exactly what we want to bound in this theorem. The rest of the proof is divided into three parts to bound each term $I$, $II$, $III$ on the right hand side.

### 4.1 Estimating $I$

From (2) and (13), we obtain that $(\tilde{U}_1, \tilde{U}_2)$ and $(Q_{(UP)_1}M_1, Q_{(UP)_2}M_2)$ are the left eigenspaces of $\tilde{Y}$ and $\tilde{Y} - Err$, respectively. Thus, we want to use Theorem 1 to bound $I$. For this purpose, we first calculate the key parameter $\delta$ in that theorem.

Recall $\tilde{Y} = Y + \epsilon W$. By Theorem 2 and Theorem 3, the $i$th largest singular value of $Y$ and that of $\tilde{Y}$ obey, with probability over $1 - e^{-(n\gamma^2)/2}$, that

$$|\lambda_i - \tilde{\lambda}_i| \leq \epsilon \|W\|_2 \leq (2 + \gamma)\epsilon, \quad \text{for } i = 1, \ldots, n.$$  

(16)

Equation (16) implies the following lower bound on $\delta$:

$$\delta \geq \sigma_{\text{min}}(\Sigma_1) - (2 + \gamma)\epsilon \equiv \delta_1.$$  

(17)

Whenever $\delta_1 > 0$, we can apply Theorem 1 to the two SVDs in (2) and (13), to obtain

$$\min_{L \text{ unitary}} \|Q_{(UP)_1}M_1 L - \tilde{U}_1\|_F \leq \frac{\sqrt{\|Err^T \tilde{U}_1\|_F^2 + \|Err^T \tilde{Y}_1\|_F^2}}{\delta}.$$  

(18)

The matrix $Err$, defined in (7) and modified in (10) and (13), is essentially a sum of several products of Gaussian matrices. Applying Lemma 1 on $Err$ and utilizing Lemma 3, we obtain the following bound:

$$\|Err\|_2 \leq \frac{\sqrt{2}E_1(\epsilon, \Sigma_1, k, n, \gamma)}{1 - E_2(\epsilon, \Sigma_1, k, n, \gamma)}.$$
holds with probability over $1 - 4e^{-(n-k)^2/2}$ whenever $\epsilon$ is small enough such that $E_2 < 1$. Here,

$$E_1(\epsilon, \Sigma, k, n, \gamma) = e^3\|\Sigma_1^{-1}\|^2(\alpha_1^2 \alpha_3 + 2\alpha_1^2 \alpha_2 + 2\alpha_1 \alpha_2 \alpha_3) + e^4\|\Sigma_1^{-1}\|^3(\alpha_1^2 \alpha_2^2 + 2\alpha_1^2 \alpha_2 \alpha_3) + e^5\|\Sigma_1^{-1}\|^4(\alpha_1^2 \alpha_3)$$

and

$$E_2 = e^2\|\Sigma_1^{-1}\|^2(\alpha_1^2 + 2e^3\|\Sigma_1^{-1}\|^2\alpha_2 + e^4\|\Sigma_1^{-1}\|^4\alpha_1^2 \alpha_2).$$

where $\alpha_1 = 1 + \gamma + \sqrt{k/n}$, $\alpha_2 = 2 + 2\gamma + 2\sqrt{k/n}$, and $\alpha_3 = 2 + \gamma$.

Now, the right hand side of (18) has the following bound:

$$\sqrt{\|\text{Err}^T \hat{U}_1\|_F^2 + \|\text{Err} \hat{V}_1\|_F^2} \leq \sqrt{2k}\|\text{Err}\|_2 \leq \frac{2\sqrt{k}E_1}{1 - E_2}. \tag{21}$$

We are now ready to define the rotation $M$ which first appears in (14). Comparing (18) with the first term of (14), we have

$$M = M_1L, \tag{22}$$

where the explicit form of $L$ is given in Theorem 1.

We plug (21) into (18), to obtain:

$$\|\hat{U}_1 - Q_UQ_P^T M\|_{\max} \leq \|\hat{U}_1 - Q_UQ_P^T M\|_F \leq \frac{2\sqrt{k}E_1}{\delta(1 - E_2)}.$$

4.2 ESTIMATING II

We start with breaking $II$ into two parts:

$$II = e^2\|U_2BM\|_{\max} \leq e^2(\max_{\|U_2C_{21}\|_{\max}} + \|U_2\|_{\max}\Sigma_1^{-1}\|_{\max}) = e^2(IV + V).$$

We estimate IV and V separately.

$$IV = \|\begin{pmatrix} C_{21}^T \\ C_{22}^T \end{pmatrix} \begin{pmatrix} U_1, U_2 \end{pmatrix} \begin{pmatrix} C_{12}^T \Sigma_1^{-1} C_{12}^{-1} \end{pmatrix} - U_1 C_{21}^T C_{12}^{-1} \|_{\max} \leq \|\Sigma_1^{-1}\|_2^2 \left( \|\begin{pmatrix} U_1, U_2 \end{pmatrix} \begin{pmatrix} C_{21}^T \\ C_{22}^T \end{pmatrix} \|_{\max} + k\|U_1\|_{\max}\|C_{21}^T C_{12}^{-1}\|_{\max} \right) \leq \|\Sigma_1^{-1}\|_2^2 \left( \|\begin{pmatrix} U_1, U_2 \end{pmatrix} \begin{pmatrix} C_{21}^T \\ C_{22}^T \end{pmatrix} \|_{\max} + k\|U_1\|_{\max}\|C_{21}^T C_{12}^{-1}\|_{\max} \right). \tag{23}$$

Observe that the entries of $(U_1, U_2) \begin{pmatrix} C_{21}^T \\ C_{22}^T \end{pmatrix}$ are i.i.d. $N(0, 1)$ and are independent of those in $C_{12}^T$. Therefore we can apply Lemma 2 to each entry $(U_1, U_2) \begin{pmatrix} C_{21}^T \\ C_{22}^T \end{pmatrix} C_{12}^T$ and those of $C_{21}^T C_{12}^{-1}$ to get, with probability at least $1 - 2e^{-(n-k)^2 + \ln(k(n+k))}$, with $0 < \beta < \frac{1}{2}$,

$$IV \leq (1 + k)(n - k)^{-\frac{1}{2} + \beta}\|\Sigma_1^{-1}\|^2_2.$$
For $V$, we first observe the following random upper bound,

$$V \leq k\|\Sigma^{-1}\|_2\|U_2C_{21}\|_{\max}\|C_{11}\|_{\max}. \tag{24}$$

Using (24), the union bound, as well as the following inequality,

$$\frac{2}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-t^2}dt \leq \sqrt{\pi}x,$$

we can estimate the probability that $V$ exceeds the value $\sqrt{2kn^{-1+\beta}\|\Sigma^{-1}\|_2^2}$,

$$P(V \geq \sqrt{2kn^{-1+\beta}\|\Sigma^{-1}\|_2^2}) \leq P(k\|\Sigma^{-1}\|_2\|U_2C_{21}\|_{\max}\|C_{11}\|_{\max} \geq 2kn^{-1+\beta}\|\Sigma^{-1}\|_2^2)$$

$$= k(n + k)P(C_{11}(1, 1) \geq \sqrt{2n^{-2+\beta}})$$

$$\leq e^{-n^\beta + \ln k(n+k)},$$

where $C_{11}(1, 1)$ is the first element of $C_{11}$. Therefore, with probability exceeding $1 - e^{-n^\beta + \ln k(n+k)}$, we have

$$V \leq 2kn^{-1+\beta}\|\Sigma^{-1}\|_2^2,$$

We combine the estimates of $IV$ and $V$ to get, with probability greater than $1 - 3e^{-1-(n-k)^\beta + \ln(k(n+k))}$,

$$II \leq e^2(1 + k)(1 + \frac{2}{n^{1/2}})(n - k)^{-\frac{1}{2}+\beta}\|\Sigma^{-1}\|_2^2 \equiv E_3. \tag{25}$$

### 4.3 ESTIMATING III

The following calculation shows how far away is $UP$ from unitary.

$$P^TU^TUP = P^TP = I + \epsilon^2 \begin{pmatrix} \Sigma_{11}^{-1} C_{21}^T C_{21} \Sigma_{11}^{-1} \\ 0 \\ C_{21} \Sigma_{11}^{-2} C_{21}^T \end{pmatrix} + O(\epsilon^3).$$

Hence,

$$(UP)^T(UP)_1 = I + \epsilon^2(\Sigma_{11} C_{21}^T C_{21} \Sigma_{11}^{-1}) + O(\epsilon^3).$$

Following a procedure similar to the one we used to obtain the bound in (21), we derive the distance between this covariance matrix and the identity matrix in Frobenius norm:

$$\|(UP)^T(UP)_1 - I\|_F \leq \sqrt{k}E_2(\epsilon, \Sigma, k, n, \gamma). \tag{26}$$

Here, $E_1$ was defined in (19). When $\epsilon$ is small enough such that $E_2 \leq \frac{1}{2\sqrt{k}}$, Theorem 2.2 in [1] shows that the distance $\|Q_{(UP)_1} - (UP)_1\|_2$ can be bounded by a function of $E_2$:

$$\|Q_{(UP)_1} - (UP)_1\|_2 \leq (\sqrt{2} + 1)\frac{\sqrt{k}E_2}{1 - \sqrt{k}E_2}.$$
Thus,

\[ III = \|Q_{(UP)_1}M - (UP)_11\|_{\text{max}} \]
\[ = \|(UP)_1(R_{(UP)_1}^{-1} - I)M\|_{\text{max}} \]
\[ \leq \sqrt{k}\|(UP)_1\|_{\text{max}}\|(R_{(UP)_1}^{-1} - I)M\|_2 \]
\[ = \sqrt{k}\|(UP)_1\|_{\text{max}}\|R_{(UP}_1^{-1}\|_2\|I - R_{(UP)_1}\|_2 \]
\[ \leq \sqrt{k}\|(UP)_1\|_{\text{max}}(\sigma_{\text{min}}((UP)_1))^{-1}\|Q_{(UP)_1} - (UP)_1\|_2. \]  (27)

To estimate (27), we first note that from (25) and the assumption that \( E_2 < \frac{1}{2\sqrt{k}} \), we obtain \(|\sigma_{\text{min}}((UP)_1^2) - 1| < \frac{1}{2}\). Therefore,

\[ \frac{1}{\sigma_{\text{min}}((UP)_1)} < \sqrt{2}. \]  (28)

We insert (26) and (28) into (27), to arrive at the bound:

\[ III \leq \sqrt{2k}(\sqrt{2} + 1)\frac{E_2}{1 - E_2}\|(UP)_1\|_{\text{max}}. \]

Furthermore, from (8) and Lemma 3, it is straightforward to estimate:

\[ \|(UP)_1\|_{\text{max}} \leq \|U_1\|_{\text{max}} + \epsilon\|U_2C_2\Sigma_1^{-1}\|_2 + \epsilon^2\|U_2B\|_2 \]
\[ \leq \|U_1\|_{\text{max}} + \epsilon\alpha_1\|\Sigma_1^{-1}\|_2 + \epsilon^2\|\Sigma_1^{-1}\|_2^2\alpha_1\alpha_2 \equiv \|U_1\|_{\text{max}} + E_4. \]  (29)

We combine the above two inequalities to get:

\[ III \leq \sqrt{2k}(\|U_1\|_{\text{max}} + E_4)(\sqrt{2} + 1)\frac{\sqrt{k}E_2}{1 - \sqrt{k}E_2}. \]

We now aggregate the estimates of I, II and III to get (4) and (5). \( \square \)

5 Application

We use the M-PSK (Phase Shift Keying) modulation classification problem as an example to show how our result is used to make a sampling strategy and a new classification method.

In the so-called adaptive modulation system, the modulation type is varying with time. When the condition of channels (such as fading or interference) changes, the transmitter seeks to find the modulation type which best adapts to the new environment. Meanwhile, if the receiver is not informed of these changes, a modulation classification procedure needs to be carried out as soon as the signal is received. Here, we consider the classification problem for the widely used MPSK type of modulations which conveys the data by changing the phase of a carrier wave. Here \( M \) stands for the number of available phases and usually takes value 2, 4, 8, 16 or 32. An MPSK signal \( s(t) \) has the following mathematical representation:

\[ s(t) = \sum_k p_T(t - nT \cos(2\pi f_c t + \theta_n + \theta_c) + w(t), \]  (30)
where \( T \) is called the symbol period/duration; \( p_T \) is a function supported on \([0, T]\) and is called baseband waveform; the frequency \( f_c \) is called carrier frequency, and \( \theta_c \) is the carrier phase. All these parameters, whether assumed known or not, are fixed for the duration of the signal. \( w(t) \) denotes an additive white Gaussian noise with two sided power spectral density \( N_0/2 \). The information is encoded in the phase parameter \( \theta_n \in \Theta_M = \{ 2\pi iM, i = 0, 1, ..., M - 1 \} \). \( M \) denotes the possible choice of phase shifts. If \( M = 2^m \), then each symbol period can encode \( m \) binary bits. The most popular cases are \( M = 2 \) and \( M = 4 \), where the modulation is called 2-PSK and 4-PSK or more often BPSK and QPSK, respectively. The classification among M-PSK modulations is defined to be the process of detecting the right value of \( M \) from a received noisy signal \( s(t) \).

Among the literature about MPSK classifications (see e.g. \([6, 7, 9]\)), there exist many different types of model settings, ranging from the fully blind classifications, i.e., where all the parameters in (30) are unknown, to the constrained classification problems where \( M \) is the only unknown parameter. For illustration purposes, we consider a simple, yet nontrivial, partially blind model assuming that \( T \) and \( N_0 \) are known, \( p_T = \chi[0, T] \), \( f_c = k/T \) for some unknown integer \( k \), \( \theta_c \) is unknown, and \( \theta_n \) is chosen uniformly from \( \Theta \).

Suppose we digitize the data in the following way. We take \( LN \) uniform samples from \( N \) periods of \( s(t) \) and store them in an \( L \times N \) matrix \( Y \), where \( Y(l,n) \) denote the \( l \)th sample of the \( n \)th period, which has the expression:

\[
Y(l,n) = s \left( (L(n-1)+l) \frac{T}{L} \right) = \cos \left( 2\pi f_c (L(n-1)+l) \frac{T}{L} + \theta_n + \theta_c \right) + W(l,n).
\]

(31)

where \( W \) is a Gaussian noise matrix with i.i.d. entries.

When noise is absent (i.e., \( W = 0 \)), each column of \( Y \) is a function of \( \theta_n \). Since \( \theta_n \) has at most \( M \) choices, the columns of \( Y \) also have only \( M \) patterns. If we deem each column as an \( L \) dimensional data point, then it merely is a high dimensional embedding of a zero dimensional parameter space. In other words, the \( L \) dimensional graph of \( Y \) consists only of \( M \) points. When noise is added, these \( M \) points becomes \( M \) clusters. Hence, the classification problem of finding the correct \( M \) is reduced to a clustering problem of finding the total number of clusters.

The complexity of all well known clustering methods, such as k-means and Mean shift grows exponentially with dimension. Therefore, for large data sets, a preprocessing step of dimension reduction is necessary. The dimension reduction procedure has another two advantages over other well known methods (e.g., \([6, 7, 9]\)):

- no carrier removal procedure is needed, relaxing the digitization rate to below the Nyquist rate of the carrier frequency.
- classification and detection are completed simultaneously.

For our problem setting, PCA is the most suitable technique for the following reasons:
• The signal has Gaussian noise and PCA is just a linear-Gaussian latent variable model.

• This is a linear model when we deem \((\cos(\theta), \sin(\theta))\) as parameters of \(Y\) (we will provide more discussion on this observation later). In telecommunications, the 2-D graph of all evaluations of \((\cos(\theta), \sin(\theta))\) with \(\theta \in \Theta\) is formally called the constellation diagram of the MPSK modulation (see Figure 1), and can be used as an identifier of the modulation type.

Therefore, the idea of our method is using PCA to map the high dimensional data to the two-dimensional constellation diagram on which clustering algorithm is applied to find the true cluster quantity.

Now we provide more detail to legislate the linear model observation. By definition, the matrix \(Y\) has the following decomposition:

\[
Y = U\Sigma V^T + W,
\]

where for \(l = 1,..,L, \, j = 1,2, \, \text{and} \, n = 1,..,N,\)

\[
U_{l,j} = \sqrt{\frac{2}{L}} \cos(2\pi f_c \frac{LT}{L} + \theta_c - (j - 1)\frac{\pi}{2}),
\]

\[
V_{n,j} = \sqrt{\frac{2}{N}} \cos(\theta_n + (j - 1)\frac{\pi}{2}), \quad (32)
\]

\[
\Sigma = \begin{pmatrix} \frac{\sqrt{LN}}{2} & 0 \\ 0 & \frac{\sqrt{LN}}{2} \end{pmatrix}. \quad (33)
\]

It is immediate to verify that when \(L \nmid f_c T\), we have \(U^T U = I_2\), and \(V\) is nearly orthogonal since we have assumed that \(\theta_n\) are chosen uniformly at random from the parameter space.
(rigorous calculation of the distance between $V$ and an orthogonal matrix can be found in [8]). This decomposition clearly shows how the noiseless part of $Y$ linearly depends on $(\cos(\theta), \sin(\theta))$. When noise is absent, PCA does the job of separating $U$ from $\Sigma V^T$, where the latter is just a scaled version of the constellation. With noise is added, as explained earlier, each individual point in the graph turns into a cluster (Figure 2). Even though it is quite obvious to a human observer how many clusters there are in Figure 2 without any prior knowledge, the existing clustering algorithms are surprisingly inferior by requiring either the number of clusters or the cluster radius as input. When the number of clusters is unknown as in our setting, many previous work suggest to do brute force on all the possible numbers (which are 2, 4, 8, 16, 32, maybe also 64, 128), and compare the clustering results in some ad hoc way to decide which is more likely. A more pleasant way to do this is by finding the cluster radius, which is exactly the place to use our results in this paper.

Before applying our results, we first normalize the singular values of $Y$ by setting $\tilde{Y} = \frac{2}{\sqrt{LN}} Y$, so that the singular values of $\tilde{Y}$ do not change with the matrix size,

$$\tilde{Y} = UV^T + \frac{2}{\sqrt{LN}} W = UV^T + \frac{2}{\sqrt{L}} \cdot \frac{1}{\sqrt{N}} W.$$  \hspace{1cm} (34)

In hardware implementations, a larger value of $L$ is usually more difficult to realize than that of $N$, because $L$ corresponds to the sampling rate and $N$ is the sample duration. Hence, we assume $L \leq N$. A generalized version of Theorem 4 for rectangular matrix can be similarly built, by padding zeros to form an $N \times N$ matrix. Since now the factor $\frac{1}{\sqrt{N}} W$ in (34) is a normalized Gaussian matrix, the other factor, $\frac{2}{\sqrt{L}}$, denotes the energy of noise, corresponding to $\epsilon$ in Theorem 4. As $L \to \infty$, $\epsilon \to 0$. Theorem 4 implies that for a given matrix, and a given confidence level, there exists a threshold $\alpha$ such that, as long as $\frac{2}{L} = \epsilon < \alpha$, we can assume the first two eigenvectors of $\tilde{Y}$ have multivariate normal distribution. In general, $\alpha$ is a function of both $L$ and $N$, which makes the inequality
Figure 3: Clustering result: Small circles denote the low dimension representation of the data points returned by PCA. The eight big circles denote the theoretically predicted cluster radius.

\[ \frac{2}{L} < \alpha(L,N) \] difficult to solve. Fortunately, in our model, where \( \sqrt{N}\|V\|_{\text{max}} \) is uniformly bounded for all \( N \), it can be verified from (5) that \( \alpha \) is universal for all sizes of \( Y \). Thus solving the simple inequality \( \frac{2}{\sqrt{L}} < \alpha \) gives us the feasible region of \( L > \frac{4}{\alpha^2} \).

With any feasible \( L \), we can apply the Mean shift clustering method with Gaussian kernel. By some elementary calculations, one can derive that the 95% percentile of the Gaussian noise on each data point in the embedded space equals \( 2.45\sqrt{2N_0(1-2/N)/(LN)} \). We set this number to be the radius.

Since the rank of \( Y \) for BPSK signals is one, then trying to invert the second singular value in Theorem 4 makes the problem near singular and cause large error in the second dimension as shown in Figure 2. Fortunately, this case is easy to be recognized by simply examining whether the second singular value of \( Y \) is much smaller than the first one.

In our first experiment, we generate a QPSK signal with carrier frequency of 1GHZ, symbol rate 10MHZ, and damped by AWGN with \( SNR = 10 \). We use the sampling rate 21 samples per symbol (much lower than the Nyquist rate of the carrier frequency, and satisfies \( L \nmid f_cT \)) and sample 200 symbols. In Figure 3, the result of PCA is plotted, together with a circle whose radius is the 95% percentile predicted by the theorem, and whose centers are those found by the meanshift algorithm. We can see that the prediction of radius is quite accurate.

In our second experiment, we let the SNR decrease and examine the performance of the above algorithm. An classification is deemed as successful only when the number of clusters returned by the Mean shift algorithm \( \hat{M} \) strictly equal to the true type \( M \). The result is plotted in Figure 4. As expected, when noise grows, the eigenvector distribution deviates from the Gaussian distribution and the predicted radius becomes too small for the algorithm to find the correct \( M \).

6 CONCLUSION

In this paper, we provided a condition under which the perturbation of the principal eigenvectors of a matrix under Gaussian noise has a near-Gaussian distribution. The condition is nonasymptotic and is useful in application. We provided an simple example
Figure 4: The success rate for BPSK, QPSK, and 8PSK modulations with respect the SNR of audio signal classification problem to illustrate how our theorem can be used to make sampling strategy and to form new classification technique. More details about this new classification scheme is discussed in [4].

ACKNOWLEDGEMENT

We are deeply grateful to Wojciech Czaja for his generous help on this article. Research presented in this paper was supported in part by Laboratory of Telecommunication Sciences. We gratefully acknowledge this support.

References

[1] A. Barrlund, Perturbation bounds for the LDLH and LU decompositions. BIT, vol. 31, pp. 358-363, 1991.

[2] E. Bura, R. Pfeiffer, On the distribution of the left singular vectors of a random matrix and its applications. Statistics and Probability Letters, vol. 78, pp. 2275-2280, 2008.

[3] D. Chandler, W. M. Kahan, The Rotation of Eigenvectors by a Perturbation. SIAM J. Numer. Anal., vol. 7, pp. 1-46, 1970.

[4] W. Czaja, R. Wang, MPSK classification by PCA. In preparation.

[5] F. Dopico, A Note on SinΘ Theorems for singular subspace variations. BIT, vol. 40, no. 2, pp. 395-403, 2000.

[6] C. Huang, A. Polydoros, Likelihood methods for Mpsk modulation classification. IEEE Transaction on Communications, vol. 43, no. 2-4, pp. 1493-1504, 1995.
[7] S. S. Soliman, S.-Z. Hsue, Signal classification using statistical moments. IEEE Trans. Commun., vol. 40, no. 5, pp. 908-916, 1992.

[8] E.M. Stein, R. Shakarchi, Real Analysis, Measure Theory, Integration, & Hilbert Spaces. vol. 3, p.91, 2005.

[9] A. Swami, B. M. Sadler, Hierarchical digital modulation classification using cumulants. IEEE Trans. Commun., vol. 48, no. 3, pp 416-429, 2000.

[10] S. J. Szarek, Condition numbers of random matrices. J. Complexity, vol. 7, no. 2, pp. 131-149, 1991.

[11] V. Vu, Singular vectors under random perturbation. Random Struct. Algorithms, vol. 39, pp. 526-538, 2011.

[12] P. A. Wedin, Perturbation bounds in connection with singular value decomposition. BIT Numer. Math., vol. 12, pp. 99-111, 1972.

[13] H. Weyl, Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung). Math. Ann. Vol. 71, no. 4, pp. 441-479, 1912.