PCA Supervised and Unsupervised Classifiers in Signal Processing

Catalina Cocianu[1], Luminita State[2], Panayiotis Vlamos[3]
Doru Constantin[2] and Corina Sararu[2]

[1] Department of Computer Science, Bucharest University of Economic Studies
Bucuresti, Romania

[2] Department of Computer Science, University of Pitesti, Pitesti, Romania

[3]Department of Computer Science, Ionian University, Corfu, Greece

Abstract. The aims of the research reported in this paper are to investigate the potential of principal directions-based approach in supervised and unsupervised frameworks. The structure of a class is represented in terms of the estimates of its principal directions computed from data, the overall dissimilarity of a particular object with a given class being given by the "disturbance" of the structure, when the object is identified as a member of this class. In case of unsupervised framework, the clusters are computed using the estimates of the principal directions. Our attempt uses arguments based on the principal components to refine the basic idea of k-means aiming to assure soundness and homogeneity to the resulted clusters. Each cluster is represented in terms of its skeleton given by a set of orthogonal and unit eigen vectors (principal directions) of sample covariance matrix, a set of principal directions corresponding to the maximum variability of the "cloud" from metric point of view. A series of conclusions experimentally established are exposed in the final section of the paper.

1 Introduction

Classical feature extraction and data projection methods have been extensively investigated in the pattern recognition and exploratory data analysis literature. Feature selection refers to a process whereby a data space is transformed into a feature space that, in theory, has precisely the same dimension as the original data space. However, the transformation is designed in such a way that a data set may be represented by a reduced number of effective features and yet retain most of the intrinsic information content of the data, that is the data set undergoes a dimensionality reduction.

Principal Component Analysis (PCA), also called Karhunen-Loeve transform is a well-known statistical method for feature extraction, data compression and multivariate data projection and so far it has been broadly used in a large series of signal and image processing, pattern recognition and data analysis applications. Principal component analysis allows the identification of a linear transform such that the axes of the resulted coordinate system correspond to the largest variability of the

Cocianu C., State L., Vlamos P., Doru C. and Sararu C. (2009).
PCA Supervised and Unsupervised Classifiers in Signal Processing.
In *Proceedings of the 9th International Workshop on Pattern Recognition in Information Systems*, pages 61-70
DOI: 10.5220/0002195000610070

investigated signal. The signal features corresponding to the new coordinate system are uncorrelated, that is, in case of normal models these components are independent. The advantages of using principal components reside from the fact that bands are uncorrelated and no information contained in one band can be predicted by the knowledge of the other bands, therefore the information contained by each band is maximum for the whole set of bits [3].

Principal components analysis seeks to explain the correlation structure of a set of predictor variables using a smaller set of linear combinations of these variables. The total variability of a data set produced by the complete set of *n* variables can often be accounted for primarily by a smaller set of *m* linear combinations of these variables, which would mean that there is almost as much information in the *m* components as there is in the original *n* variables. The principal components represent a new coordinate system, found by rotating the original system along the directions of maximum variability [7].

Classical PCA is based on the second-order statistics of the data and, in particular, on the eigen-structure of the data covariance matrix and accordingly, the PCA neural models incorporate only cells with linear activation functions. More recently, several generalizations of the classical PCA models to non-Gaussian models, the Independent Component Analysis (ICA) and the Blind Source Separation techniques (BSS) have become a very attractive and promising framework in developing more efficient image restoration algorithms [8].

In unsupervised classification, the classes are not known at the start of the process. The number of classes, their defining features and their objects have to be determined. The unsupervised classification can be viewed as a process of seeking valid summaries of data comprising classes of similar objects such that the resulted classes are well separated in the sense that objects are not only similar to other objects belonging to the same class, but also significantly different from objects in another classes. Occasionally, the summaries of a data set are expected to be relevant for describing a large collection of objects and allowing predictions or to discover hypotheses on the inner structures in the data.

Since similarity plays a key role for both clustering and classification purposes, the problem of finding relevant indicators to measure the similarity between two patterns drawn from the same feature space became of major importance. Recently, alternative methods as discriminant common vectors, neighborhood components analysis and Laplacianfaces have been proposed allowing the learning of linear projection matrices for dimensionality reduction [4], [10].

The aims of the research reported in this paper are to investigate the potential of principal directions-based approach in supervised and unsupervised frameworks. The structure of a class is represented in terms of the estimates of its principal directions computed from data, the overall dissimilarity of a particular object with a given class being given by the “disturbance” of the structure, when the object is identified as a member of this class. In case of unsupervised framework, the clusters are computed using the estimates of the principal directions. Our attempt uses arguments based on the principal components to refine the basic idea of k-means aiming to assure soundness and homogeneity to the resulted clusters. The clusters are represented in terms of skeletons given by sets of orthogonal and unit eigen vectors (principal directions) of each cluster sample covariance matrix. According to the well known result established by Karhunen and Loeve, a set of principal directions corresponds to the maximum variability of the “cloud” from metric point of view, and they are also almost optimal from informational point of view, the principal directions being recommended by the maximum entropy principle as the most reliable characteristics of the repartition.

A series of conclusions experimentally established are exposed in the final section of the paper.

2 Methodology Based on Principal Direction for Classification and Recognition Purposes

In probabilistic models for pattern recognition and classification, the classes are represented in terms of multivariate density functions, and an object coming from a certain class is modeled as a random vector whose repartition has the density function corresponding to this class. In cases when there is no statistical information concerning the set of density functions corresponding to the classes involved in the recognition process, usually estimates based on the information extracted from available data are used instead.

The principal directions of a class are given by a set of unit orthogonal eigen vectors of the covariance matrix. When the available data is represented by a set of objects $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_N$, belonging to a certain class C, the covariance matrix is estimated by the sample covariance matrix,

$$\hat{\boldsymbol{\Sigma}}_N = \frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_N\right)\left(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_N\right)^T \tag{1}$$

where $\hat{\boldsymbol{\mu}}_N = \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i$.

Let us denote by $\lambda_1^N \geq \lambda_2^N \geq ... \geq \lambda_n^N$ the eigen values and by $\boldsymbol{\psi}_1^N, ..., \boldsymbol{\psi}_n^N$ a set of orthonormal eigen vectors of $\hat{\Sigma}_N$.

If a new example $\mathbf{X}_{N+1}$ coming from the same class has to be included in the sample, the new estimate of the covariance matrix can be recomputed as,

$$\hat{\boldsymbol{\Sigma}}_{N+1} = \hat{\boldsymbol{\Sigma}}_N + \frac{1}{N+1}\left(\mathbf{X}_{N+1} - \hat{\boldsymbol{\mu}}_N\right)\left(\mathbf{X}_{N+1} - \hat{\mu}_N\right)^T - \frac{1}{N}\hat{\boldsymbol{\Sigma}}_N \tag{2}$$

Using first order approximations [11], the estimates of the eigen values and eigen vectors respectively are given by,

$$\lambda_i^{N+1} = \lambda_i^N + \left(\boldsymbol{\psi}_i^N\right)^T \Delta\hat{\boldsymbol{\Sigma}}_N \boldsymbol{\psi}_i^N = \left(\boldsymbol{\psi}_i^N\right)^T \hat{\boldsymbol{\Sigma}}_{N+1} \boldsymbol{\psi}_i^N \tag{3}$$

$$\boldsymbol{\psi}_i^{N+1} = \boldsymbol{\psi}_i^{N} + \sum_{\substack{j=1\\ j\neq i}}^{n} \frac{\left(\boldsymbol{\psi}_N^{j}\right)^T \Delta\hat{\boldsymbol{\Sigma}}_N \boldsymbol{\psi}_i^{N}}{\lambda_i^N - \lambda_j^N} \boldsymbol{\psi}_j^{N} \tag{4}$$

On the other hand, when an object has to be removed from the sample, then the estimate of the covariance matrix can be computed as (see appendix),

$$\hat{\boldsymbol{\Sigma}}_N = \hat{\boldsymbol{\Sigma}}_{N+1} + \Delta\boldsymbol{\Sigma}_{N+1} \tag{5}$$

where $\Delta\boldsymbol{\Sigma}_{N+1} = \frac{1}{N-1}\hat{\boldsymbol{\Sigma}}_{N+1} - \frac{N}{(N-1)(N+1)}\left(\mathbf{X}_{N+1} - \boldsymbol{\mu}_N\right)\left(\mathbf{X}_{N+1} - \boldsymbol{\mu}_N\right)^T$ and $\boldsymbol{\mu}_N = \frac{(N+1)\boldsymbol{\mu}_{N+1}}{N} - \frac{\mathbf{X}_{N+1}}{N}$.

Let $\boldsymbol{\psi}_1^N,...,\boldsymbol{\psi}_n^N$ be set of principal directions of the class C computed using $\hat{\boldsymbol{\Sigma}}_N$. When the example $\mathbf{X}_{N+1}$ is identified as a member of the class *C*, then the disturbance implied by extending C is expressed as,

$$D = \frac{1}{n}\sum_{k=1}^{n} d\left(\boldsymbol{\psi}_k^N, \boldsymbol{\psi}_k^{N+1}\right) \tag{6}$$

where *d* is the Euclidian distance and $\boldsymbol{\psi}_1^{N+1},...,\boldsymbol{\psi}_n^{N+1}$ are the principal directions computed using $\hat{\boldsymbol{\Sigma}}_{N+1}$.

Let $H = \{C_1, C_2, ..., C_M\}$ be a set of classes, where the class C_j contains N_j elements. The new object **X** is allotted to C_j, one of the classes for which

$$D = \frac{1}{n}\sum_{k=1}^{n} d\left(\boldsymbol{\psi}_{k,j}^{N_j}, \boldsymbol{\psi}_{k,j}^{N_j+1}\right) = \min_{1\leq p\leq M} \frac{1}{n}\sum_{k=1}^{n} d\left(\boldsymbol{\psi}_{k,p}^{N_p}, \boldsymbol{\psi}_{k,p}^{N_p+1}\right) \tag{7}$$

In order to protect against misclassifications, due to insufficient "closeness" to any class, we implement this recognition technique using a threshold *T*>0 such that the example **X** is allotted to C_j only if relation (7) holds and *D*<*T*.

Briefly, the recognition procedure, P1, is described below [3].

```
Input: $H = \{C_1, C_2, ..., C_M\}$
Repeat
i←1
Step 1: Let X be a new sample. Classify X according to (7)
Step 2: If $\exists j, 1 \leq j \leq M$ such that X is allotted to $C_j$, then
2.1.re-compute the characteristics of $C_j$ using (2), (3) and (4)
2.2. i++
Step 3: If i<PN go to Step 1
```

```
        Else
        3.1. For i=$\overline{1,M}$ , compute the characteristics of class $C_i$ using **M**.
        3.2. go to Step 1.
Until the last new example was classified
Output: The new set $\{C_1, C_2, ..., C_M\} \cup CR$
```

In unsupervised classification, the clusters are computed by identifying the natural grouping trends existing in data. Our approach based on principal directions, P2, is described as follows [12]. The input is represented by the data to be classified $\aleph = \{X_1, X_2, ..., X_N\}$, the number of clusters *M,* and the set of initial seeds $P_1, P_2, .., P_M$.

Parameters are:

- θ , the threshold value to control the cluster size; $\theta \in (0,1)$
- *nr*, the threshold value for the cluster homogeneity;
- *Cond*, the stopping condition, expressed in terms of the threshold value *NoRe*, for the number of re-allotted data;
- ρ , the control parameter, $\rho \in (0,1)$, to control the number of re-allotted data.

Initializations.

$t \leftarrow 0$; $P_1, P_2, .., P_M$ are taken as initial centers of the clusters $C_1^0, C_2^0, ..., C_M^0$ respectively.

Step 1. Generate the set of initial clusters, $C^0 = \{C_1^0, C_2^0, ..., C_M^0\}$

The data $X_1, X_2, ..., X_N$ are allotted to the initial clusters according to the minimum distance to the cluster centers.

Step2. Compute the set of cluster skeletons, $S^t = \{S_1^t, ..., S_M^t\}$, where $S_k^t = (\psi_{k,1}^t, \psi_{k,2}^t, ..., \psi_{k,n}^t)$ is the skeleton of the cluster *k* at the moment *t*.

```
Step3.
Repeat
t=t+1;
S^t= S^{t-1}; C^t= C^{t-1}
```

Compute the new set of clusters according to the minimum distance to the skeletons of the current clusters. For each cluster C_k^{t-1} compute C_k^t by performing the following operations.

1. Add the elements $X_i \in \aleph$ not belonging to C_k^t , and $k = \arg \min_{1 \le cl \le M} D(X_i, S_{cl}^t)$.

2. Remove the elements $X_i \in C_k^t$ for which $D(X_i, S_k^t) > \min_{1 \le cl \le M} D(X_i, S_{cl}^t)$

3. Test on the homogeneity of C_k^t :

```
3.1. Compute the new center $c_k^t = \frac{1}{\left|C_k^t\right|} \sum_{X \in C_k^t} X$ .
3.2. If $\left|F_1 \cup F_2\right| > nr$ then $C_k^t$ is not homogenous and it is homogenous otherwise,
     where $F_1 = \left\{ X \in C_k^t \middle/ \left\| X - c_k^t \right\|_2 > \theta \max_{X \in C_k^t} \left\| X - c_k^t \right\|_2 \right\}$ and
     $F_2 = \left\{ X \in C_k^t / \exists j \neq k, D\left(X, S_k^t\right) > D\left(X, S_j^t\right) \right\}$
4.   Extend $C_k^t$ in case it is homogenous by adding each $X_i \in \aleph$ for which
     $D\left(X_i, S_k^t\right) = \min_{1 \le cl \le M} D\left(X_i, S_{cl}^t\right)$.
5.   In case the test decides that $C_k^t$ is not homogenous, the cluster $C_k^t$ is corrected by
     re-allotting the set of the most $\rho\left|F_1 \cup F_2\right|$ disturbing elements from $F_1 \cup F_2$ that
     is the elements of the maximum distance to $S_k^t$ .
6.   Re-compute $S_k^t$ , the skeleton of the new $C_k^t$
7.   Re-allot the elements of $C_k^{t-1} \setminus C_k^t$ according to the minimum distance to cluster's
     skeleton
8.   Compute the new set of skeletons $\mathscr{S}^t$
Until Cond
```

3 Tests on the Proposed Signal Classification and Recognition Methods

Several tests on the recognition procedure P1 were performed on different classes of signals. The results proved very good performance in terms of the recognition error.

The results of a test on a two-class problem in signal recognition are presented in Fig. 1, Fig. 2, and Fig. 3. The samples are extracted from the signals depicted in Fig. 1. In Fig. 2 are represented the initial samples. The correct recognition of 20 new examples coming from these two classes using P1 failed in 2 cases. The correctly recognized examples are presented in Fig. 3. The performance was improved significantly when the volume of the initial samples increases. Using the leaving-one-out method, the values of the resulted empirical mean error are less than 0.05 (more than 95% new examples are correctly recognized). In order to apply the leaving-one-out method, some first order approximations for the covariance matrices, eigen values and eigen vectors had to be derived. The recursive equations based on first order approximations allow to avoid the re-computing of covariance matrices, eigen vectors and eigen values. The computations are provided in the appendix.

A series of tests were performed on P2 and they pointed out that in spite of its higher complexity as compare to k-means, significantly increased accuracy is obtained.

For instance, in case of 5 classes of data of dimensionality 6, using the first 2 principal directions, the results obtained in the compressed space are presented in Fig. 4. The examples were generated by sampling from the normal distributions for each class. The matrix having as entries the Mahalanobis distances between classes is,

$$10^3 * \begin{pmatrix} 0 & 3.3655 & 1.5008 & 2.5378 & 2.8579 \\ 3.3655 & 0 & 0.6417 & 4.1304 & 2.0341 \\ 1.5008 & 0.6417 & 0 & 1.7881 & 1.9008 \\ 2.5378 & 4.1304 & 1.7881 & 0 & 2.2807 \\ 2.8579 & 2.0341 & 1.9008 & 2.2807 & 0 \end{pmatrix}.$$

Table 1.

The sample	$\aleph_1$	$\aleph_2$	$\aleph_3$	$\aleph_4$
Number of misclassified examples by our method	0	0	1	0
Number of misclassified examples by k-means	1	0	280	3
Number of iterations	2	2	3	2

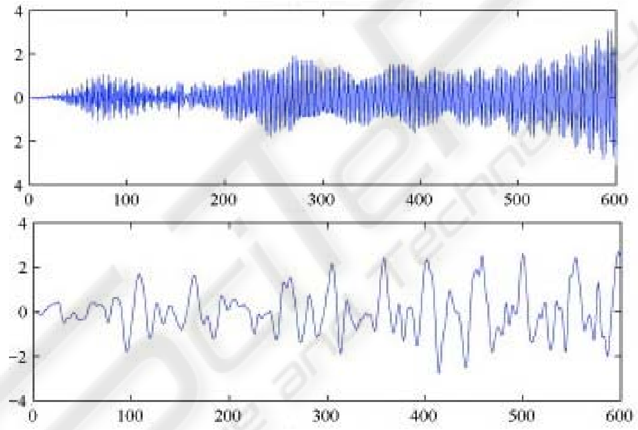

Fig. 1.

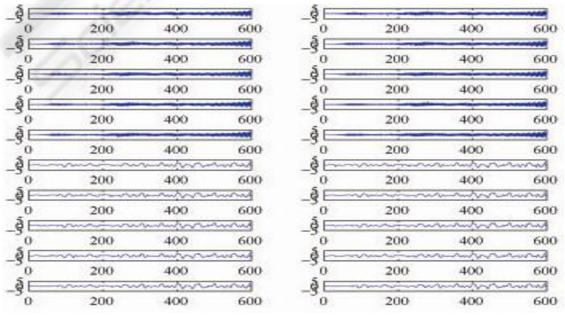

Fig. 2.

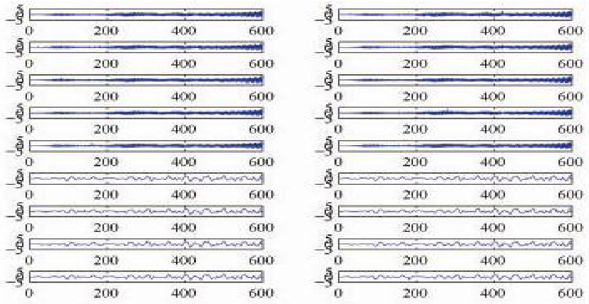

Fig. 3.

The results in case of the sample $\aleph_3$ are shown in Figures 4a, 4b, and 4c.

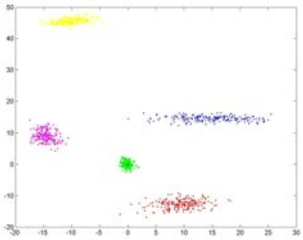

Fig. 4a. The actual classification.

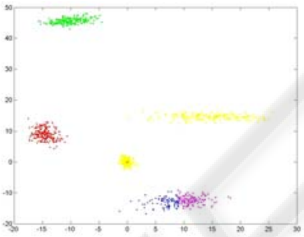

Fig. 4b. The clusters resulted by applying k-means.

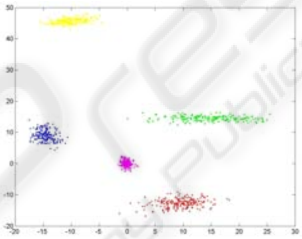

Fig. 4c. The clusters resulted by applying the proposed method.

References

1. Chatterjee, C., Roychowdhury, V.P., Chong, E.K.P.: On Relative Convergence Properties of PCA Algorithms, IEEE Trans. on Neural Networks, vol.9,no.2 (1998).
2. Cocianu, C., State, L., Rosca, I., Vlamos, P: A New Adaptive Classification Scheme Based on Skeleton Information, Proceedings of SIGMAP 2007 (2007).
3. Diamantaras, K.I., Kung, S.Y.: Principal Component Neural Networks: theory and applications, John Wiley &Sons, (1996).
4. Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R.: Neighbourhood Component Analysis, Proceedings of the Conference on Advances in Neural Information Processing Systems (2004).
5. Gordon, A.D.: Classification, Chapman&Hall/CRC, 2nd Edition (1999).
6. Haykin, S., Neural Networks A Comprehensive Foundation, Prentice Hall,Inc. (1999).
7. Hastie, T., Tibshirani, R., Friedman,J.: The Elements of Statistical Learning Data Mining, Inference, and Prediction, Springer (2001).
8. Hyvarinen, A., Karhunen, J., Oja, E. Independent Component Analysis, John Wiley &Sons (2001).
9. Larose, D.T. Data Mining. Methods and Models, Wiley-Interscience, A John Wiley and Sons, Inc Publication, Hoboken, New Jersey (2006).
10. Liu, J., and Chen, S. Discriminant common vectors versus neighborhood components analysis and Laplacianfaces: A comparative study in small sample size problem. Image and Vision Computing 24 (2006).

11. State, L., Cocianu, C., Vlamos, P, Stefanescu, V. PCA-Based Data Mining Probabilistic and Fuzzy Approaches with Applications in Pattern Recognition, Proceedings of ICSOFT 2006 (2006).
12. State, L., Cocianu, C., Rosca, I., Vlamos, P: A New Learning Algorithm for Classification in the Reduced Space, Proceedings of ICEIS 2008 (2008).

Appendix

Lemma. Let $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_K$ be an *n*-dimensional Bernoullian sample. We denote by $\hat{\boldsymbol{\mu}}_N = \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i$, $\hat{\boldsymbol{\Sigma}}_N = \frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_N\right)\left(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_N\right)^T$, and let $\left\{\lambda_i^N\right\}_{1\le i\le n}$ be the eigen values and $\left\{\boldsymbol{\psi}_i^N\right\}_{1\le i\le n}$ a set of orthogonal unit eigen vectors of $\hat{\boldsymbol{\Sigma}}_N$, $2 \le N \le K-1$. In case the eigen values of $\hat{\boldsymbol{\Sigma}}_{N+1}$ are pair wise distinct, the following first order approximations hold,

$$\lambda_i^N = \lambda_i^{N+1} + \left(\boldsymbol{\psi}_i^{N+1}\right)^T \Delta\boldsymbol{\Sigma}_{N+1}\boldsymbol{\psi}_i^{N+1} \tag{8}$$

$$\boldsymbol{\psi}_i^N = \boldsymbol{\psi}_i^{N+1} + \sum_{\substack{j=1\\ j\ne i}}^{n} \frac{\left(\boldsymbol{\psi}_j^{N+1}\right)^T \Delta\boldsymbol{\Sigma}_{N+1}\boldsymbol{\psi}_i^{N+1}}{\lambda_i^{N+1} - \lambda_j^{N+1}}\boldsymbol{\psi}_j^{N+1} \tag{9}$$

where $\Delta\boldsymbol{\Sigma}_{N+1} = \hat{\boldsymbol{\Sigma}}_N - \hat{\boldsymbol{\Sigma}}_{N+1}$.

Proof Using the perturbation theory, we get, $\hat{\boldsymbol{\Sigma}}_N = \hat{\boldsymbol{\Sigma}}_{N+1} + \Delta\boldsymbol{\Sigma}_{N+1}$ and, $\boldsymbol{\psi}_i^N = \boldsymbol{\psi}_i^{N+1} + \Delta\boldsymbol{\psi}_i^{N+1}$, $\lambda_i^N = \lambda_i^{N+1} + \Delta\lambda_i^{N+1}$, $1 \le i \le n$. Then,

$$\Delta\boldsymbol{\Sigma}_{N+1} = \frac{1}{N-1}\hat{\boldsymbol{\Sigma}}_{N+1} - \frac{N}{(N-1)(N+1)}\left(\mathbf{X}_{N+1} - \boldsymbol{\mu}_N\right)\left(\mathbf{X}_{N+1} - \boldsymbol{\mu}_N\right)^T \tag{10}$$

where $\boldsymbol{\mu}_N = \frac{(N+1)\boldsymbol{\mu}_{N+1}}{N} - \frac{\mathbf{X}_{N+1}}{N}$

$$\left(\hat{\boldsymbol{\Sigma}}_{N+1} + \Delta\boldsymbol{\Sigma}_{N+1}\right)\left(\boldsymbol{\psi}_i^{N+1} + \Delta\boldsymbol{\psi}_i^{N+1}\right) = \left(\lambda_i^{N+1} + \Delta\lambda_i^{N+1}\right)\left(\boldsymbol{\psi}_i^{N+1} + \Delta\boldsymbol{\psi}_i^{N+1}\right) \tag{11}$$

Using first order approximations, from (11) we get,

$$\begin{aligned} &\lambda_i^{N+1}\boldsymbol{\psi}_i^{N+1} + \hat{\boldsymbol{\Sigma}}_{N+1}\Delta\boldsymbol{\psi}_i^{N+1} + \Delta\boldsymbol{\Sigma}_{N+1}\boldsymbol{\psi}_i^{N+1} \cong \\ &\quad \cong \lambda_i^{N+1}\boldsymbol{\psi}_i^{N+1} + \lambda_i^{N+1}\Delta\boldsymbol{\psi}_i^{N+1} + \Delta\lambda_i^{N+1}\boldsymbol{\psi}_i^{N+1} \end{aligned} \tag{12}$$

hence,

$$\begin{aligned} &\left(\boldsymbol{\psi}_i^{N+1}\right)^T \hat{\boldsymbol{\Sigma}}_{N+1}\Delta\boldsymbol{\psi}_i^{N+1} + \left(\boldsymbol{\psi}_i^{N+1}\right)^T \Delta\boldsymbol{\Sigma}_{N+1}\boldsymbol{\psi}_i^{N+1} \cong \\ &\quad \cong \lambda_i^{N+1}\left(\boldsymbol{\psi}_i^{N+1}\right)^T \Delta\boldsymbol{\psi}_i^{N+1} + \Delta\lambda_i^{N+1}\left\|\boldsymbol{\psi}_i^{N+1}\right\|^2 \end{aligned} \tag{13}$$

Using $\lambda_i^{N+1}\left(\boldsymbol{\psi}_i^{N+1}\right)^T = \left(\boldsymbol{\psi}_i^{N+1}\right)^T \hat{\boldsymbol{\Sigma}}_{N+1}$ we obtain ,

$$\lambda_i^{N+1}\left(\boldsymbol{\psi}_i^{N+1}\right)^T \Delta\boldsymbol{\psi}_i^{N+1} + \left(\boldsymbol{\psi}_i^{N+1}\right)^T \Delta\boldsymbol{\Sigma}_{N+1}\boldsymbol{\psi}_i^{N+1} \cong \lambda_i^{N+1}\left(\boldsymbol{\psi}_i^{N+1}\right)^T \Delta\boldsymbol{\psi}_i^{N+1} + \Delta\lambda_i^{N+1} \tag{14}$$

hence $\Delta\lambda_i^{N+1} = \left(\boldsymbol{\psi}_i^{N+1}\right)^T \Delta\mathbf{\Sigma}_{N+1}\boldsymbol{\psi}_i^{N+1}$ that is,

$$\lambda_i^N = \lambda_i^{N+1} + \left(\boldsymbol{\psi}_i^{N+1}\right)^T \Delta\mathbf{\Sigma}_{N+1}\boldsymbol{\psi}_i^{N+1} \tag{15}$$

The first order approximations of the orthonormal eigen vectors of $\hat{\mathbf{\Sigma}}_N$ can be derived using the expansion of each vector $\Delta\boldsymbol{\psi}_i^{N+1}$ in the basis represented by the orthonormal eigen vectors of $\hat{\mathbf{\Sigma}}_{N+1}$,

$$\Delta\boldsymbol{\psi}_i^{N+1} = \sum_{j=1}^{n} b_{i,j}\boldsymbol{\psi}_j^{N+1} \tag{16}$$

where

$$b_{i,j} = \left(\boldsymbol{\psi}_j^{N+1}\right)^T \Delta\boldsymbol{\psi}_i^{N+1} \tag{17}$$

Using the orthonormality, we get,

$$\begin{aligned} 1 = \left\|\boldsymbol{\psi}_i^{N+1} + \Delta\boldsymbol{\psi}_i^{N+1}\right\|^2 \cong \left\|\boldsymbol{\psi}_i^{N+1}\right\|^2 + 2\left(\boldsymbol{\psi}_i^{N+1}\right)^T\left(\Delta\boldsymbol{\psi}_i^{N+1}\right) = \\ = 1 + 2\left(\boldsymbol{\psi}_i^{N+1}\right)^T\left(\Delta\boldsymbol{\psi}_i^{N+1}\right) \end{aligned} \tag{18}$$

that is

$$b_{i,i} = \left(\boldsymbol{\psi}_i^{N+1}\right)^T \Delta\boldsymbol{\psi}_i^{N+1} = 0 \tag{19}$$

Using (11), the approximation,

$$\hat{\mathbf{\Sigma}}_{N+1}\Delta\boldsymbol{\psi}_i^{N+1} + \Delta\mathbf{\Sigma}_{N+1}\boldsymbol{\psi}_i^{N+1} \cong \lambda_i^{N+1}\Delta\boldsymbol{\psi}_i^{N+1} + \Delta\lambda_i^{N+1}\boldsymbol{\psi}_i^{N+1} \tag{20}$$

holds for each $1 \le i \le n$.

For $1 \le j \ne i \le n$, from (20) we obtain the following equations,

$$\begin{aligned} \left(\boldsymbol{\psi}_j^{N+1}\right)^T \hat{\mathbf{\Sigma}}_{N+1}\Delta\boldsymbol{\psi}_i^{N+1} + \left(\boldsymbol{\psi}_j^{N+1}\right)^T \Delta\mathbf{\Sigma}_{N+1}\boldsymbol{\psi}_i^{N+1} \cong \\ \cong \lambda_i^{N+1}\left(\boldsymbol{\psi}_j^{N+1}\right)^T \Delta\boldsymbol{\psi}_i^{N+1} + \Delta\lambda_i^{N+1}\left(\boldsymbol{\psi}_j^{N+1}\right)^T \boldsymbol{\psi}_i^{N+1} \end{aligned} \tag{21}$$

$$\left(\boldsymbol{\psi}_j^{N+1}\right)^T \hat{\mathbf{\Sigma}}_{N+1}\Delta\boldsymbol{\psi}_i^{N+1} + \left(\boldsymbol{\psi}_j^{N+1}\right)^T \Delta\mathbf{\Sigma}_{N+1}\boldsymbol{\psi}_i^{N+1} \cong \lambda_i^{N+1}\left(\boldsymbol{\psi}_j^{N+1}\right)^T \Delta\boldsymbol{\psi}_i^{N+1} \tag{22}$$

$$\lambda_j^{N+1}\left(\boldsymbol{\psi}_j^{N+1}\right)^T \Delta\boldsymbol{\psi}_i^{N+1} + \left(\boldsymbol{\psi}_j^{N+1}\right)^T \Delta\mathbf{\Sigma}_{N+1}\boldsymbol{\psi}_i^{N+1} \cong \lambda_i^{N+1}\left(\boldsymbol{\psi}_j^{N+1}\right)^T \Delta\boldsymbol{\psi}_i^{N+1} \tag{23}$$

From (23) we get,

$$b_{i,j} = \left(\boldsymbol{\psi}_j^{N+1}\right)^T \Delta\boldsymbol{\psi}_i^{N+1} = \frac{\left(\boldsymbol{\psi}_j^{N+1}\right)^T \Delta\mathbf{\Sigma}_{N+1}\boldsymbol{\psi}_i^{N+1}}{\lambda_i^{N+1} - \lambda_j^{N+1}} \tag{24}$$

Consequently, the first order approximation of the eigen vectors of $\hat{\mathbf{\Sigma}}_N$ are,

$$\boldsymbol{\psi}_i^{N+1} + \Delta\boldsymbol{\psi}_i^{N+1} \cong \boldsymbol{\psi}_i^{N+1} + \sum_{\substack{j=1 \\ j\ne i}}^{n} \frac{\left(\boldsymbol{\psi}_j^{N+1}\right)^T \Delta\mathbf{\Sigma}_{N+1}\boldsymbol{\psi}_i^{N+1}}{\lambda_i^{N+1} - \lambda_j^{N+1}}\boldsymbol{\psi}_j^{N+1} \tag{25}$$