

# Guide to the MATLAB code for Coordinate-Wise sparse PCA

Amir Beck, Yakov Vaisbourd

## 1 General Description

This guide briefly describes the usage of the function presented in the paper:

Amir Beck and Yakov Vaisbourd, "Optimization Methods for Solving the Sparse PCA Problem".

In addition, Appendix [A](#) contains a complete list of the 20 gene expression data sets from the GeneChip oncology database that were used in order to conduct some numerical study in the aforementioned paper.

The package contains a single m-file with the function `cwPCA`, which is an implementation of the Greedy and Partial CW PCA algorithms presented in the paper. Both algorithms solve the sparse PCA problem which is given by

$$\begin{aligned} \max_{\mathbf{x} \in \mathbb{R}^n} \quad & \mathbf{x}^T \mathbf{A} \mathbf{x} \\ \text{(SPCA)} \quad \text{s.t.} \quad & \|\mathbf{x}\|_2 \leq 1, \\ & \|\mathbf{x}\|_0 \leq s, \end{aligned}$$

where the  $l_0$ -norm is defined as  $\|\mathbf{x}\|_0 = |\{i : x_i \neq 0\}|$  and  $\mathbf{A} \in \mathbb{R}^{n \times n}$  is the covariance matrix. The covariance matrix is given by  $\mathbf{A} = \mathbf{D}^T \mathbf{D}$  where  $\mathbf{D} \in \mathbb{R}^{m \times n}$  is the data matrix. The function `cwPCA` is designed to solve ([SPCA](#)) with either of the two matrices is given as an input.

The function `cwPCA` requires only two obligatory input arguments; the covariance or the data matrix and the sparsity level. The following list summarizes the additional arguments which are supported. This arguments should be provided by the user as a - name, value - pairs as it will be demonstrated in Section [2](#).

### List of optional arguments:

- `initial_support` - The initial support.  
*Possible Values:* any vector  $\mathbf{v} \in \{1, 2, \dots, n\}^s$   
*Default Value:*  $(1, 2, \dots, s)^T$
- `mat_type` - Indicates which type of matrix is given as an input; covariance/correlation matrix (CMat) or data matrix (DMat).  
*Possible Values:* 'CMat', 'DMat'  
*Default Value:* 'CMat'
- `type` - Indicates the algorithm type; Greedy CW (GCW) or Partial CW (PCW).  
*Possible Values:* 'GCW', 'PCW'  
*Default Value:* 'GCW'
- `max_iter` - Additional stopping criteria - maximum number of iterations. If applied then there is no guarantee that the solution is CW maximal.  
*Possible Values:* positive integer  
*Default Value:* 0 (Suppressed)
- `display` - Control output display.  
*Possible Values:* logical (true, false)  
*Default Value:* false

The function performs some input validation. Nevertheless, if a covariance/correlation matrix is given as an input, then it assumes that the matrix is symmetric positive semi definite.

## 2 Examples

Generate random data

```
m = 150;      % Number of samples
n = 1000;     % Number of variables
s = 100;      % Sparsity level
D = (m^(-0.5))*randn(m,n);      % Generate the zero mean
D = D - repmat(mean(D),m,1);    % data matrix
A = D'*D;     % Compute the covariance matrix
```

Some particular examples of applying the algorithms on the data just generated:

- Run the Greedy CW algorithm. Use the covariance matrix as input.

```
x = cwPCA(A, s);
```

- Run the Greedy CW algorithm. Use the data matrix as input. Initialize the algorithms with a random support.

```
SupInit = randperm(n);
x = cwPCA(D, s, 'mat_type', 'DMat', 'initial_support', ...
           SupInit(1:s));
```

- Run the Partial CW algorithm. Use the covariance matrix as input. Restrict to 5 iterations and display output at each iteration.

```
x = cwPCA(A, s, 'type', 'PCW', 'max_iter', 5, 'display', 'on');
```

Running the last command will produce the following output:

```
Starting cwPCA method with type: PCW
-----
|      | index | index | increase in  |
|   k  |  out  |  in   | function value|
|-----|
|   1  |   59  |  334  | 0.0042589086  |
|   2  |   73  |  672  | 0.0070413964  |
|   3  |   52  |  479  | 0.0027497472  |
|   4  |   77  |  716  | 0.0069070628  |
|   5  |   76  |  101  | 0.0012274473  |
|-----|
```

For each iteration the function will display the indices that were excluded and included in the support accompanied with the increase in the objective function of (SPCA).

## A GeneChip oncology dataset list

The following table includes the list of GeneChip datasets used in the numerical study. For most of the datasets a PubMed<sup>1</sup> identification is also included.

| #  | Study Name            | PubMed ID | Samples | Variables |
|----|-----------------------|-----------|---------|-----------|
| 1  | brain_nutt            | 12670911  | 50      | 12625     |
| 2  | brain_pomeroy         | 11807556  | 98      | 7129      |
| 3  | brain_rickman         | 11559565  | 51      | 7129      |
| 4  | brain_turkheimer      | 17140431  | 30      | 54675     |
| 5  | brain_wang            |           | 102     | 12625     |
| 6  | cervical_bachtiary    | 17020965  | 33      | 54675     |
| 7  | colon_ancona          |           | 47      | 22283     |
| 8  | headandneck_kuriakose | 15170515  | 44      | 12625     |
| 9  | kidney_copland        | GSE6344   | 20      | 22283     |
| 10 | leukemia_ferrando     | 12086890  | 39      | 7129      |
| 11 | leukemia_haslinger    | 15459216  | 28      | 12626     |
| 12 | leukemia_holleman     | 15295046  | 173     | 22283     |
| 13 | leukemia_soulier      | 15774621  | 104     | 22283     |
| 14 | leukemia_stegmaier    | 14770183  | 87      | 22283     |
| 15 | leukemia_teuffel      | 15257931  | 31      | 22283     |
| 16 | leukemia_yeoh         | 12086872  | 335     | 12625     |
| 17 | lung_beer             | 12118244  | 96      | 7129      |
| 18 | lung_bhattacharjee    | 11707567  | 254     | 12625     |
| 19 | lung_lu               | 17194181  | 36      | 12625     |
| 20 | lymphoma_piccaluga    | 17304354  | 60      | 54675     |

---

<sup>1</sup><http://www.ncbi.nlm.nih.gov/pubmed>