Chapter 19

Graph Clustering

19.1 $K$-Way Clustering Using Normalized Cuts

We now consider the general case in which $K \geq 3$.

Two crucial issues need to be addressed (to the best of our knowledge, these points are not clearly articulated in the literature).

1. The choice of a matrix representation for partitions on the set of vertices.

   It is important that such a representation be scale-invariant.

   It is also necessary to state necessary and sufficient conditions for such matrices to represent a partition.
2. The choice of a metric to compare solutions.

We describe a partition \( (A_1, \ldots, A_K) \) of the set of nodes \( V \) by an \( N \times K \) matrix \( X = [X^1 \cdots X^K] \) whose columns \( X^1, \ldots, X^K \) are indicator vectors of the partition \( (A_1, \ldots, A_K) \).

Inspired by what we did in Section ??, we assume that the vector \( X^j \) is of the form

\[
X^j = (x^j_1, \ldots, x^j_N),
\]

where \( x^j_i \in \{a_j, b_j\} \) for \( j = 1, \ldots, K \) and \( i = 1, \ldots, N \), and where \( a_j, b_j \) are any two distinct real numbers.

The vector \( X^j \) is an indicator vector for \( A_j \) in the sense that, for \( i = 1, \ldots, N \),

\[
x^j_i = \begin{cases} a_j & \text{if } v_i \in A_j \\ b_j & \text{if } v_i \notin A_j. \end{cases}
\]
When \( \{a_j, b_j\} = \{0, 1\} \) for \( j = 1, \ldots, K \), such a matrix is called a *partition matrix* by Yu and Shi.

However, such a choice is premature, since it is better to have a scale-invariant representation to make the denominators of the Rayleigh ratios go away.

Since the partition \((A_1, \ldots, A_K)\) consists of nonempty pairwise disjoint blocks whose union is \(V\), some conditions on \(X\) are required to reflect these properties, but we will worry about this later.

As in Section ??, we seek conditions on the \(a_j\)s and the \(b_j\)s in order to express the normalized cut \(\text{Ncut}(A_1, \ldots, A_K)\) as a sum of Rayleigh ratios.

Then, we reformulate our optimization problem in a more convenient form, by chasing the denominators in the Rayleigh ratios, and by expressing the objective function in terms of the *trace* of a certain matrix.
This will reveal the important fact that the solutions of the relaxed problem are right-invariant under multiplication by a $K \times K$ orthogonal matrix.

Let $d = 1^\top D 1$ and $\alpha_j = \text{vol}(A_j)$, so that $\alpha_1 + \cdots + \alpha_K = d$.

Then, $\text{vol}(\overline{A}_j) = d - \alpha_j$, and as in Section ??, we have

\[
(X_j)^\top LX^j = (a_j - b_j)^2 \text{cut}(A_j, \overline{A}_j),
\]

\[
(X_j)^\top DX^j = \alpha_j a_j^2 + (d - \alpha_j) b_j^2.
\]
When \( K \geq 3 \), unlike the case \( K = 2 \), in general we have \( \text{cut}(A_j, \overline{A_j}) \neq \text{cut}(A_k, \overline{A_k}) \) if \( j \neq k \), and since

\[
\text{Ncut}(A_1, \ldots, A_K) = \sum_{j=1}^{K} \frac{\text{cut}(A_j, \overline{A_j})}{\text{vol}(A_j)},
\]

we would like to choose \( a_j, b_j \) so that

\[
\frac{\text{cut}(A_j, \overline{A_j})}{\text{vol}(A_j)} = \frac{(X^j)^\top L X^j}{(X^j)^\top D X^j}, \quad j = 1, \ldots, K,
\]

because this implies that

\[
\mu(X) = \text{Ncut}(A_1, \ldots, A_K) = \sum_{j=1}^{K} \frac{\text{cut}(A_j, \overline{A_j})}{\text{vol}(A_j)} = \sum_{j=1}^{K} \frac{(X^j)^\top L X^j}{(X^j)^\top D X^j}.
\]
We find the condition

\[ 2\alpha_j b_j (b_j - a_j) = db_j^2. \]

The above equation is trivially satisfied if \( b_j = 0 \).

If \( b_j \neq 0 \), then

\[ a_j = \frac{2\alpha_j - d}{2\alpha_j} b_j. \]

This choice seems more complicated than the choice \( b_j = 0 \), so we will opt for the choice \( b_j = 0, j = 1, \ldots, K \).

With this choice, we get

\[ (X^j)^\top DX^j = \alpha_j a_j^2. \]
Thus, it makes sense to pick

$$a_j = \frac{1}{\sqrt{\alpha_j}} = \frac{1}{\sqrt{\text{vol}(A_j)}}, \quad j = 1, \ldots, K,$$

which is the solution presented in von Luxburg [?]. This choice also corresponds to the scaled partition matrix used in Yu [?] and Yu and Shi [?].

When $N = 10$ and $K = 4$, an example of a matrix $X$ representing the partition of $V = \{v_1, v_2, \ldots, v_{10}\}$ into the four blocks
\[ \{A_1, A_2, A_3, A_4\} = \{\{v_2, v_4, v_6\}, \{v_1, v_5\}, \{v_3, v_8, v_{10}\}, \{v_7, v_9\}\}, \]

is shown below:

\[
X = \begin{pmatrix}
0 & a_2 & 0 & 0 \\
0 & a_1 & 0 & 0 \\
0 & 0 & a_3 & 0 \\
0 & a_1 & 0 & 0 \\
0 & 0 & a_2 & 0 \\
0 & a_1 & 0 & 0 \\
0 & 0 & 0 & a_4 \\
0 & 0 & a_3 & 0 \\
0 & 0 & 0 & a_4 \\
0 & 0 & a_3 & 0 \\
\end{pmatrix}.
\]

Let us now consider the problem of finding necessary and sufficient conditions for a matrix \(X\) to represent a partition of \(V\).

When \(b_j = 0\), the pairwise disjointness of the \(A_i\) is captured by the orthogonality of the \(X^i\):

\[
(X^i)^\top X^j = 0, \quad 1 \leq i, j \leq K, \ i \neq j. \quad (*)
\]
When we formulate our minimization problem in terms of Rayleigh ratios, conditions on the quantities \((X^i)^\top DX^i\) show up, and it is more convenient to express the orthogonality conditions using the quantities \((X^i)^\top DX^j\) instead of the \((X^i)^\top X^j\), because these various conditions can be combined into a single condition involving the matrix \(X^\top DX\).

Now, because \(D\) is a diagonal matrix with positive entries and because the nonzero entries in each column of \(X\) have the same sign, for any \(i \neq j\), the condition

\[
(X^i)^\top X^j = 0
\]

is equivalent to

\[
(X^i)^\top DX^j = 0. \tag{**}
\]
Observe that the orthogonality conditions (*) (and (**)) are equivalent to the fact that every row of \( X \) has at most one nonzero entry.

Each \( A_j \) is nonempty iff \( X^j \neq 0 \), and the fact that the union of the \( A_j \) is \( V \) is captured by the fact that each row of \( X \) must have some nonzero entry (every vertex appears in some block).

It is not immediately obvious how to state conveniently this condition in matrix form.

Since every row of any matrix \( X \) representing a partition has a single nonzero entry \( a_j \), we have

\[
X^\top X = \text{diag} \left( n_1 a_1^2, \ldots, n_K a_K^2 \right),
\]

where \( n_j \) is the number of elements in \( A_j \), the \( j \)th block of the partition.
Therefore, the condition for the columns of $X$ to be nonzero is

$$\det(X^\top X) \neq 0.$$  

Another condition which does not involve explicitly a determinant and is scale-invariant stems from the observation that not only

$$X^\top X = \text{diag}(n_1a_1^2, \ldots, n_Ka_K^2),$$

but

$$X^\top 1_N = \begin{pmatrix} n_1a_1 \\ \vdots \\ n_Ka_K \end{pmatrix},$$
and these equations imply that

\[
(X^\top X)^{-1}X^\top 1_N = \begin{pmatrix}
\frac{1}{a_1} \\
\vdots \\
\frac{1}{a_K}
\end{pmatrix},
\]

and thus

\[
X(X^\top X)^{-1}X^\top 1_N = 1_N. \tag{†}
\]

Note that because the columns of \( X \) are linearly independent, \((X^\top X)^{-1}X^\top\) is the pseudo-inverse \( X^+ \) of \( X \).

Consequently, if \( X^\top X \) is invertible, condition (†) can also be written as

\[
XX^+ 1_N = 1_N.
\]
However, it is well known that $XX^+$ is the orthogonal projection of $\mathbb{R}^K$ onto the range of $X$ (see Gallier [?], Section 14.1), so the condition $XX^+1_N = 1_N$ is equivalent to the fact that $1_N$ belongs to the range of $X$.

In retrospect, this should have been obvious since the columns of a solution $X$ satisfy the equation

$$a_1^{-1}X^1 + \cdots + a_K^{-1}X^K = 1_N.$$

We emphasize that it is important to use conditions that are invariant under multiplication by a nonzero scalar, because the Rayleigh ratio is scale-invariant, and it is crucial to take advantage of this fact to make the denominators go away.
If we let

\[ X = \left\{ [X^1 \ldots X^K] \mid X^j = a_j(x_i^j, \ldots, x_N^j), x_i^j \in \{1, 0\}, \right. \]
\[ \left. a_j \in \mathbb{R}, X^j \neq 0 \right\} \]

(note that the condition \( X^j \neq 0 \) implies that \( a_j \neq 0 \)),
then the set of matrices representing partitions of \( V \) into \( K \) blocks is

\[ \mathcal{K} = \left\{ X = [X^1 \ldots X^K] \mid X \in \mathcal{X}, \right. \]
\[ \left. (X^i)^\top DX^j = 0, \right. \]
\[ 1 \leq i, j \leq K, i \neq j, \]
\[ X(X^\top X)^{-1}X^\top 1 = 1 \right\}. \]
As in the case $K = 2$, to be rigorous, the solution are really $K$-tuples of points in $\mathbb{RP}^{N-1}$, so our solution set is really

$$\mathbb{P}(\mathcal{K}) = \left\{ (\mathbb{P}(X^1), \ldots, \mathbb{P}(X^K)) \mid [X^1 \cdots X^K] \in \mathcal{K} \right\}.$$  

In view of the above, we have our first formulation of $K$-way clustering of a graph using normalized cuts, called problem PNC1 (the notation PNCX is used in Yu [?], Section 2.1):
$K$-way Clustering of a graph using Normalized Cut, Version 1:
Problem PNC1

\[
\minimize \sum_{j=1}^{K} \frac{(X^j)^\top L X^j}{(X^j)^\top D X^j}
\]

subject to \((X^i)^\top D X^j = 0, \quad 1 \leq i, j \leq K, \ i \neq j, \ X(X^\top X)^{-1} X^\top 1 = 1, \ X \in \mathcal{X}.
\]

As in the case $K = 2$, the solutions that we are seeking are $K$-tuples $(\mathbb{P}(X^1), \ldots, \mathbb{P}(X^K))$ of points in $\mathbb{RP}^{N-1}$ determined by their homogeneous coordinates $X^1, \ldots, X^K$. 
Remark:

Because

\[(X^j)^\top LX^j = (X^j)^\top DX^j - (X^j)^\top WX^j = \text{vol}(A_j) - (X^j)^\top WX^j,\]

Instead of minimizing

\[\mu(X^1, \ldots, X^K) = \sum_{j=1}^K \frac{(X^j)^\top LX^j}{(X^j)^\top DX^j},\]

we can maximize

\[\epsilon(X^1, \ldots, X^K) = \sum_{j=1}^K \frac{(X^j)^\top WX^j}{(X^j)^\top DX^j},\]

since

\[\epsilon(X^1, \ldots, X^K) = K - \mu(X^1, \ldots, X^K).\]
Theoretically, minimizing $\mu(X^1, \ldots, X^K)$ is equivalent to maximizing $\epsilon(X^1, \ldots, X^K)$, but from a practical point of view, it is preferable to maximize $\epsilon(X^1, \ldots, X^K)$.

This is because minimizing solutions of $\mu$ are obtained from (unit) eigenvectors corresponding to the $K$ smallest eigenvalues of $L_{\text{sym}} = D^{-1/2}LD^{-1/2}$ (by multiplying these eigenvectors by $D^{1/2}$).

However, numerical methods for computing eigenvalues and eigenvectors of a symmetric matrix do much better at computing largest eigenvalues.

Since $L_{\text{sym}} = I - D^{-1/2}WD^{-1/2}$, the eigenvalues of $L_{\text{sym}}$ listed in increasing order correspond to the eigenvalues of $I - L_{\text{sym}} = D^{-1/2}WD^{-1/2}$ listed in decreasing order.
Furthermore, \( v \) is an eigenvector of \( L_{\text{sym}} \) for the \( i \)th smallest eigenvalue \( \nu_i \) iff \( v \) is an eigenvector of \( I - L_{\text{sym}} \) for the \((N + 1 - i)\)th largest eigenvalue \( \nu_i \).

Therefore, it is preferable to find the \textit{largest} eigenvalues of \( I - L_{\text{sym}} = D^{-1/2}WD^{-1/2} \) and their eigenvectors.

In fact, since the eigenvalues of \( L_{\text{sym}} \) are in the range \([0, 2]\), the eigenvalues of \( 2I - L_{\text{sym}} = I + D^{-1/2}WD^{-1/2} \) are also in the range \([0, 2]\) (that is, \( I + D^{-1/2}WD^{-1/2} \) is positive semidefinite).

Let us now show how our original formulation (PNC1) can be converted to a more convenient form, by chasing the denominators in the Rayleigh ratios, and by expressing the objective function in terms of the \textit{trace} of a certain matrix.
Proposition 19.1. For any orthogonal $K \times K$ matrix $R$, any symmetric $N \times N$ matrix $A$, and any $N \times K$ matrix $X = [X^1 \cdots X^K]$, the following properties hold:

1. $\mu(X) = \text{tr}(\Lambda^{-1}X^\top LX)$, where
   \[
   \Lambda = \text{diag}((X^1)^\top DX^1, \ldots, (X^K)^\top DX^K).
   \]

2. If $(X^1)^\top DX^1 = \cdots = (X^K)^\top DX^K = \alpha^2$, then
   \[
   \mu(X) = \mu(XR) = \frac{1}{\alpha^2} \text{tr}(X^\top LX).
   \]

3. The condition $X^\top AX = \alpha^2 I$ is preserved if $X$ is replaced by $XR$.

4. The condition $X(X^\top X)^{-1}X^\top 1 = 1$ is preserved if $X$ is replaced by $XR$.

Now, by Proposition 19.1(1) and the fact that the conditions in PNC1 are scale-invariant, we are led to the following formulation of our problem:
minimize $\text{tr}(X^\top LX)$
subject to $(X^i)^\top DX^j = 0, \quad 1 \leq i, j \leq K, \quad i \neq j,$
$(X^j)^\top DX^j = 1, \quad 1 \leq j \leq K,$
$X(X^\top X)^{-1}X^\top 1 = 1,$
$X \in \mathcal{X}.$

Conditions on lines 2 and 3 can be combined in the equation

$$X^\top DX = I,$$

and, we obtain the following formulation of our minimization problem:
$K$-way Clustering of a graph using Normalized Cut, Version 2: Problem PNC2

\[
\begin{align*}
\text{minimize} & \quad \text{tr}(X^\top LX) \\
\text{subject to} & \quad X^\top DX = I, \\
& \quad X(X^\top X)^{-1}X^\top 1 = 1, \quad X \in \mathcal{X}.
\end{align*}
\]

Because problem PNC2 requires the constraint $X^\top DX = I$ to be satisfied, it does not have the same set of solutions as problem PNC1.
Nevertheless, problem PNC2 is equivalent to problem PNC1, in the sense that for every minimal solution $(X^1, \ldots, X^K)$ of PNC1,

$$(((X^1)^{\top}DX^1)^{-1/2}X^1, \ldots, ((X^K)^{\top}DX^K)^{-1/2}X^K)$$

is a minimal solution of PNC2 (with the same minimum for the objective functions), and that for every minimal solution $(Z^1, \ldots, Z^K)$ of PNC2, $(\lambda_1Z^1, \ldots, \lambda_KZ^K)$ is a minimal solution of PNC1, for all $\lambda_i \neq 0$, $i = 1, \ldots, K$ (with the same minimum for the objective functions).

In other words, problems PNC1 and PNC2 have the same set of minimal solutions as $K$-tuples of points $(\mathbb{P}(X^1), \ldots, \mathbb{P}(X^K))$ in $\mathbb{RP}^{N-1}$ determined by their homogeneous coordinates $X^1, \ldots, X^K$. 
Formulation PNC2 reveals that finding a minimum normalized cut has a geometric interpretation in terms of the graph drawings discussed in Section 4.1.

Indeed, PNC2 has the following equivalent formulation: Find a minimal energy graph drawing $X$ in $\mathbb{R}^K$ of the weighted graph $G = (V, W)$ such that:

1. The matrix $X$ is orthogonal with respect to the inner product $\langle - , - \rangle_D$ in $\mathbb{R}^N$ induced by $D$, with

   $\langle x, y \rangle_D = x^\top D y, \quad x, y \in \mathbb{R}^N.$

2. The rows of $X$ are nonzero; this means that no vertex $v_i \in V$ is assigned to the origin of $\mathbb{R}^K$ (the zero vector $0_K$).

3. Every vertex $v_i$ is assigned a point of the form $(0, \ldots, 0, a_j, 0, \ldots, 0)$ on some axis (in $\mathbb{R}^K$).

4. Every axis in $\mathbb{R}^K$ is assigned at least some vertex.
Condition 1 can be reduced to the standard condition for graph drawings \((R^\top R = I)\) by making the change of variable \(Y = D^{1/2}X\) or equivalently \(X = D^{-1/2}Y\). Indeed,

\[
\text{tr}(X^\top LX) = \text{tr}(Y^\top D^{-1/2}LD^{-1/2}Y),
\]

so we use the normalized Laplacian \(L_{\text{sym}} = D^{-1/2}LD^{-1/2}\) instead of \(L\),

\[
X^\top DX = Y^\top Y = I,
\]

and conditions (2), (3), (4) are preserved under the change of variable \(Y = D^{1/2}X\), since \(D^{1/2}\) is invertible.

However, conditions (2), (3), (4) are “hard” constraints, especially condition (3).
In fact, condition (3) implies that the columns of $X$ are orthogonal with respect to both the Euclidean inner product and the inner product $\langle -, - \rangle_D$, so condition (1) is redundant, except for the fact that it prescribes the norm of the columns, but this is not essential due to the projective nature of the solutions.

The main problem in finding a good relaxation of problem PNC2 is that it is very difficult to enforce the condition $X \in \mathcal{X}$.

Also, the solutions $X$ are not preserved under arbitrary rotations, but only by very special rotations which leave $\mathcal{X}$ invariant (they exchange the axes).
The first natural relaxation of problem PNC2 is to drop the condition that $X \in \mathcal{X}$, and we obtain the

**Problem (⋆₂)**

\[
\begin{align*}
\text{minimize} & \quad \text{tr}(X^\top LX) \\
\text{subject to} & \quad X^\top DX = I, \\
& \quad X(X^\top X)^{-1}X^\top 1 = 1.
\end{align*}
\]

Actually, since the discrete solutions $X \in \mathcal{X}$ that we are ultimately seeking are solutions of problem PNC1, the preferred relaxation is the one obtained from problem PNC1 by dropping the condition $X \in \mathcal{X}$, and simply requiring that $X^j \neq 0$, for $j = 1, \ldots, K$: 
Problem \((\ast_1)\)

minimize \[ \sum_{j=1}^{K} \frac{(X^j)^\top LX^j}{(X^j)^\top DX^j} \]

subject to \((X^i)^\top DX^j = 0, X^j \neq 0\) \(1 \leq i, j \leq K, i \neq j,\)
\[ X(X^\top X)^{-1}X^\top 1 = 1. \]

Now that we dropped the condition \(X \in \mathcal{X}\), it is not clear that \(X^\top X\) is invertible in \((\ast_1)\) and \((\ast_2)\).

However, since the columns of \(X\) are nonzero and \(D\)-orthogonal, they must be linearly independent, so \(X\) has rank \(K\) and and \(X^\top X\) is invertible.
As we explained before, every solution $Z = [Z^1, \ldots, Z^K]$ of problem ($\ast_1$) yields a solution of problem ($\ast_2$) by normalizing each $Z^j$ by $((Z^j)^\top D Z^j)^{1/2}$, and conversely for every solution $Z = [Z^1, \ldots, Z^K]$ of problem ($\ast_2$), the $K$-tuple $[\lambda_1 Z^1, \ldots, \lambda_K Z^K]$ is a solution of problem ($\ast_1$), where $\lambda_j \neq 0$ for $j = 1, \ldots, K$.

Furthermore, by Proposition 19.1, for every orthogonal matrix $R \in O(K)$ and for every solution $X$ of ($\ast_2$), the matrix $XR$ is also a solution of ($\ast_2$).

Since Proposition 19.1(2) requires that all $(X^j)^\top DX^j$ have the same value in order to have $\mu(X) = \mu(XR)$, in general, if $X$ is a solution of ($\ast_1$), the matrix $XR$ is not necessarily a solution of ($\ast_1$).
However, every solution $X$ of $(\ast_2)$ is also a solution of $(\ast_1)$, for every $R \in O(K)$, $XR$ is a solution of both $(\ast_2)$ and $(\ast_1)$, and since $(\ast_1)$ is scale-invariant, for every diagonal invertible matrix $\Lambda$, the matrix $XRA$ is a solution of $(\ast_1)$.

In summary, every solution $Z$ of problem $(\ast_2)$ yields a family of solutions of problem $(\ast_1)$; namely, all matrices of the form $ZRA$, where $R \in O(K)$ and $\Lambda$ is a diagonal invertible matrix.

We will take advantage of this fact in looking for a discrete solution $X$ “close” to a solution $Z$ of the relaxed problem $(\ast_2)$.

Observe that a matrix is of the form $RA$ with $R \in O(K)$ and $\Lambda$ a diagonal invertible matrix iff its columns are nonzero and pairwise orthogonal.
Recall that if $X^\top X$ is invertible (which is the case), condition $X(X^\top X)^{-1}X^\top 1 = 1$ is equivalent to $XX^+1 = 1$, which is also equivalent to the fact that $1$ is in the range of $X$.

If we make the change of variable $Y = D^{1/2}X$ or equivalently $X = D^{-1/2}Y$, the condition that $1$ is in the range of $X$ becomes the condition that $D^{1/2}1$ is in the range of $Y$, which is equivalent to

$$YY^+D^{1/2}1 = D^{1/2}1.$$ 

However, since $Y^\top Y = I$, we have

$$Y^+ = Y^\top,$$

so we get the equivalent problem
Problem (**\text{2})

minimize \( \text{tr}(Y^\top D^{-1/2}L D^{-1/2}Y) \)
subject to
\[ Y^\top Y = I, \]
\[ YY^\top D^{1/2}1 = D^{1/2}1. \]

We pass from a solution \( Y \) of problem (**\text{2}) to a solution \( Z \) of problem (*\text{2}) by \( Z = D^{-1/2}Y \).

It is not a priori obvious that the minimum of \( \text{tr}(Y^\top L_{\text{sym}} Y) \) over all \( N \times K \) matrices \( Y \) satisfying \( Y^\top Y = I \) is equal to the sum \( \nu_1 + \cdots + \nu_K \) of the first \( K \) eigenvalues of \( L_{\text{sym}} = D^{-1/2}LD^{-1/2} \).
Fortunately, the Poincaré separation theorem (Proposition 12.24) guarantees that the sum of the $K$ smallest eigenvalues of $L_{\text{sym}}$ is a lower bound for $\text{tr}(Y^\top L_{\text{sym}} Y)$.

Furthermore, if we temporarily ignore the second constraint, the minimum of problem (**) is achieved by any $K$ unit eigenvectors $(u_1, \ldots, u_K)$ associated with the smallest eigenvalues

$$0 = \nu_1 \leq \nu_2 \leq \ldots \leq \nu_K$$

of $L_{\text{sym}}$.

We may assume that $\nu_2 > 0$, namely that the underlying graph is connected (otherwise, we work with each connected component), in which case $Y^1 = D^{1/2} \mathbf{1} / \|D^{1/2} \mathbf{1}\|_2$, because $\mathbf{1}$ is in the nullspace of $L$. 
Since $Y^1 = D^{1/2}1 / \|D^{1/2}1\|_2$, the vector $D^{1/2}1$ is in the range of $Y$, so the condition

$$YY^\top D^{1/2}1 = D^{1/2}1$$

is also satisfied. Then, $Z = D^{-1/2}Y$ with $Y = [u_1 \ldots u_K]$ yields a minimum of our relaxed problem ($*_2$) (the second constraint is satisfied because $1$ is in the range of $Z$).

By Proposition 17.6, the vectors $Z^j$ are eigenvectors of $L_{tw}$ associated with the eigenvalues $0 = \nu_1 \leq \nu_2 \leq \ldots \leq \nu_K$.

Recall that $1$ is an eigenvector for the eigenvalue $\nu_1 = 0$, and $Z^1 = 1 / \|D^{1/2}1\|_2$. 
Because, \((Y^i)^\top Y^j = 0\) whenever \(i \neq j\), we have

\[(Z^i)^\top DZ^j = 0, \text{ whenever } i \neq j.\]

This implies that \(Z^2, \ldots, Z^K\) are all orthogonal to \(D1\), and thus, that each \(Z^j\) has both some positive and some negative coordinate, for \(j = 2, \ldots, K\).

The conditions \((Z^i)^\top DZ^j = 0\) do not necessarily imply that \(Z^i\) and \(Z^j\) are orthogonal (w.r.t. the Euclidean inner product), but we can obtain a solution of Problems \((\ast_2)\) and \((\ast_1)\) achieving the same minimum for which distinct columns \(Z^i\) and \(Z^j\) are simultaneously orthogonal and \(D\)-orthogonal, by multiplying \(Z\) by some \(K \times K\) orthogonal matrix \(R\) on the right.
Indeed, if $Z$ is a solution of $(\star_2)$ obtained as above, the $K \times K$ symmetric matrix $Z^\top Z$ can be diagonalized by some orthogonal $K \times K$ matrix $R$ as

$$Z^\top Z = R\Sigma R^\top,$$

where $\Sigma$ is a diagonal matrix, and thus,

$$R^\top Z^\top Z R = (ZR)^\top Z R = \Sigma,$$

which shows that the columns of $ZR$ are orthogonal.

By Proposition 19.1, $ZR$ also satisfies the constraints of $(\star_2)$ and $(\star_1)$, and $\text{tr}((ZR)^\top L(ZR)) = \text{tr}(Z^\top LZ)$. 
Remark: Since $Y$ has linearly independent columns (in fact, orthogonal) and since $Z = D^{-1/2}Y$, the matrix $Z$ also has linearly independent columns, so $Z^T Z$ is positive definite and the entries in $\Sigma$ are all positive.

Also, instead of computing $Z^T Z$ explicitly and diagonalizing it, the matrix $R$ can be found by computing an SVD of $Z$. 