SPECTRAL CLUSTERING, Lesson 2.
Normalized Laplacian and modularity matrices, contingency tables, RKHS

Marianna Bolla, DSc. Prof. BME Math. Inst.

October 23, 2020

1 SVD of contingency tables and correspondence matrices

[This section can be skipped.]

Now, more generally, our underlying objects will be contingency tables, i.e. rectangular arrays with nonnegative, real entries. For example, keyword–document matrices or microarrays are such. In microarrays, rows correspond to genes and columns to different conditions, while the corresponding entries are expression levels of genes under specific conditions (a 0-1 matrix is a special case of it). Let $\boldsymbol{C}$ be a contingency table on row set $Row = \{1, \dots, m\}$, column set $Col = \{1, \dots, n\}$, where $\boldsymbol{C}$ is $m \times n$ rectangular matrix of nonnegative real entries c_{ij}'s. Without loss of generality, we can assume that there are no identically zero rows or columns (otherwise they can be omitted). Here c_{ij} is some kind of association between the objects representing row i and column j, where 0 means no interaction at all. Usually, the entries of $\boldsymbol{C}$ are normalized, either with a uniform bound, say 1 (like probabilities), or the sum of the entries is 1 (reminiscent of a joint distribution). This normalization will have importance in Section 4, here it has no relevance, since the correspondence matrix to be introduced is invariant under scaling the entries of $\boldsymbol{C}$. Let the row-sums of $\boldsymbol{C}$ be

$$d_{row,i} = \sum_{j=1}^{n} c_{ij}, \quad i = 1, \dots, m \tag{1}$$

and the column-sums

$$d_{col,j} = \sum_{i=1}^{m} c_{ij}, \quad j = 1, \dots, n \tag{2}$$

which are collected in the main diagonal of the $m \times m$ diagonal matrix $\boldsymbol{D}_{row} = \operatorname{diag}(d_{row,1}, \dots, d_{row,m})$ and that of the $n \times n$ diagonal matrix $\boldsymbol{D}_{col} = \operatorname{diag}(d_{col,1}, \dots, d_{col,n})$, respectively.

For a given integer $1 \le k \le \min\{m, n\}$, we are looking for k-dimensional representatives $\mathbf{r}_1, \dots, \mathbf{r}_m \in \mathbb{R}^k$ of the rows and $\mathbf{q}_1, \dots, \mathbf{q}_n \in \mathbb{R}^k$ of the columns such that they minimize the objective function

$$Q_k = \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} \|\mathbf{r}_i - \mathbf{q}_j\|^2 \tag{3}$$

subject to

$$\sum_{i=1}^{m} d_{row,i} \mathbf{r}_i \mathbf{r}_i^T = \boldsymbol{I}_k \quad \textit{and} \quad \sum_{j=1}^{n} d_{col,j} \mathbf{q}_j \mathbf{q}_j^T = \boldsymbol{I}_k. \tag{4}$$

When minimized, the objective function Q_k favors k-dimensional placement of the rows and columns such that representatives of columns and rows with large association are forced to be close to each other. As we will see, this is equivalent to the problem of Correspondence Analysis.

Let us put both the objective function and the constraints in a more favorable form. Let $\boldsymbol{X}$ be the $m \times k$ matrix of rows $\mathbf{r}_1^T, \dots, \mathbf{r}_m^T$, and $\mathbf{x}_1, \dots, \mathbf{x}_k \in \mathbb{R}^n$ denote the columns of $\boldsymbol{X}$, for which fact we use the notation $\boldsymbol{X} = (\mathbf{x}_1, \dots, \mathbf{x}_k)$. Because of the constraint (4), the vectors $\boldsymbol{D}_{row}^{-1/2} \mathbf{x}_i$ $(i = 1, \dots, k)$ form an orthonormal system, hence, $\boldsymbol{D}_{row}^{-1/2} \boldsymbol{X}$ is a suborthogonal matrix. Therefore, the first part of the constraint can be formulated as $\boldsymbol{X}^T \boldsymbol{D}_{row} \boldsymbol{X} = \boldsymbol{I}_k$. Likewise, let $\boldsymbol{Y}$ be the $n \times k$ matrix of rows $\mathbf{q}_1^T, \dots, \mathbf{q}_n^T$, and $\mathbf{y}_1, \dots, \mathbf{y}_k \in \mathbb{R}^n$ denote the columns of $\boldsymbol{Y}$, i.e. $\boldsymbol{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_k)$. Hence, the second part of the constraint (4) can be formulated as $\boldsymbol{Y}^T \boldsymbol{D}_{col} \boldsymbol{Y} = \boldsymbol{I}_k$ and the matrix $\boldsymbol{D}_{col}^{-1/2} \boldsymbol{Y}$ is also suborthogonal.

With this notation, the objective function (3) is rewritten as

$$\begin{aligned} Q_k &= \sum_{i=1}^{m} d_{row,i} \|\mathbf{r}_i\|^2 + \sum_{j=1}^{n} d_{col,j} \|\mathbf{q}_j\|^2 - \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} \mathbf{r}_i^T \mathbf{q}_j \\ &= \sum_{\ell=1}^{k} \mathbf{x}_\ell^T \boldsymbol{D}_{row} \mathbf{x}_\ell + \sum_{\ell=1}^{k} \mathbf{y}_\ell^T \boldsymbol{D}_{col} \mathbf{y}_\ell - \sum_{\ell=1}^{k} \mathbf{x}_\ell^T \boldsymbol{C} \mathbf{y}_\ell \\ &= \operatorname{tr}(\boldsymbol{X}^T \boldsymbol{D}_{row} \boldsymbol{X}) + \operatorname{tr}(\boldsymbol{Y}^T \boldsymbol{D}_{col} \boldsymbol{Y}) - \operatorname{tr}(\boldsymbol{X}^T \boldsymbol{C} \boldsymbol{Y}) \\ &= 2k - \operatorname{tr}(\boldsymbol{X}^T \boldsymbol{C} \boldsymbol{Y}) = 2k - \operatorname{tr}[(\boldsymbol{D}_{row}^{1/2} \boldsymbol{X})^T (\boldsymbol{D}_{row}^{-1/2} \boldsymbol{C} \boldsymbol{D}_{col}^{-1/2}) (\boldsymbol{D}_{col}^{1/2} \boldsymbol{Y})] \end{aligned}$$

where the matrix $\boldsymbol{C}_{corr} = \boldsymbol{D}_{row}^{-1/2} \boldsymbol{C} \boldsymbol{D}_{col}^{-1/2}$ is introduced in [20] as the *normalized contingency table* or *correspondence matrix* corresponding to the contingency table $\boldsymbol{C}$.

The correspondence matrix has singular value decomposition, briefly SVD

$$\boldsymbol{C}_{corr} = \sum_{k=0}^{r-1} s_k \mathbf{v}_k \mathbf{u}_k^T, \tag{5}$$

where $r \le \min\{n, m\}$ is the rank of $\boldsymbol{C}_{corr}$, or equivalently (as there are no identically zero rows or columns), the rank of $\boldsymbol{C}$. Here $1 = s_0 \ge s_1 \ge \dots \ge s_{r-1} > 0$ are the non-zero singular values of $\boldsymbol{C}_{corr}$. They cannot exceed 1, since they are correlations. Furthermore, 1 is a single singular value if $\boldsymbol{C}_{corr}$ (or equivalently, $\boldsymbol{C}$) is non-decomposable). In this case $\mathbf{v}_0 = (\sqrt{d_{row,1}}, \dots, \sqrt{d_{row,m}})^T$ and $\mathbf{u}_0 = (\sqrt{d_{col,1}}, \dots, \sqrt{d_{col,n}})^T$ is the singular vector pair corresponding to $s_0 = 1$.

Note that the singular spectrum of a decomposable contingency table can be composed from the singular spectra of its non-decomposable parts, as well as their singular vector pairs. Therefore, in the future, the non-decomposability of the underlying contingency table will be assumed.

Theorem 1 (Representation theorem for contingency tables) *Let $\boldsymbol{C}$ be a non-decomposable contingency table with correspondence matrix $\boldsymbol{C}_{corr}$. Let $1 = s_0 > s_1 \geq \cdots \geq s_{r-1}$ be the positive singular values of $\boldsymbol{C}_{corr}$ with unit-norm singular vector pairs $\mathbf{v}_i, \mathbf{u}_i$ $(i = 0, \dots, r-1)$, and $k \leq r$ be a positive integer such that $s_{k-1} > s_k$. Then the minimum of (3) subject to (4) is $2k - \sum_{i=0}^{k-1} s_i$ and it is attained with the optimum row representatives $\mathbf{r}_1^*, \dots, \mathbf{r}_m^*$ and column representatives $\mathbf{q}_1^*, \dots, \mathbf{q}_n^*$ the transposes of which are row vectors of the matrices $\boldsymbol{X}^* = \boldsymbol{D}_{row}^{-1/2}(\mathbf{v}_0, \mathbf{v}_1, \dots, \mathbf{v}_{k-1})$ and $\boldsymbol{Y}^* = \boldsymbol{D}_{col}^{-1/2}(\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_{k-1})$, respectively.*

Proof 1 *In fact, we have to maximize*

$$\mathrm{tr}[(\boldsymbol{D}_{row}^{1/2}\boldsymbol{X})^T \boldsymbol{C}_{corr} (\boldsymbol{D}_{col}^{1/2}\boldsymbol{Y})]$$

under the constraints that $\boldsymbol{D}_{row}^{1/2}\boldsymbol{X}$ and $\boldsymbol{D}_{col}^{1/2}\boldsymbol{Y}$ are suborthogonal matrices.

Definition 1 *The vectors $\mathbf{r}_1^*, \dots, \mathbf{r}_n^*$ and $\mathbf{q}_1^*, \dots, \mathbf{q}_m^*$ giving the optimum in Theorem 1 are called optimum k-dimensional representatives of the rows and columns, while the transformed singular vectors $\boldsymbol{D}_{row}^{-1/2}\mathbf{v}_0, \dots, \boldsymbol{D}_{row}^{-1/2}\mathbf{v}_{k-1}$ and $\boldsymbol{D}_{col}^{-1/2}\mathbf{u}_0, \dots, \boldsymbol{D}_{col}^{-1/2}\mathbf{u}_{k-1}$ are called vector components of the contingency table, taking part in the k-dimensional representation of its rows and columns.*

We remark the following.

- Provided 1 is a single singular value (or equivalently, $\boldsymbol{C}$ is non-decomposable), the first columns of the matrices $\boldsymbol{X}^*$ and $\boldsymbol{Y}^*$ are $\boldsymbol{D}_{row}^{-1/2}\mathbf{v}_0$ and $\boldsymbol{D}_{col}^{-1/2}\mathbf{u}_0$, i.e. the constantly $\mathbf{1}$ vectors of $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively. Therefore they do not contribute significantly to the separation of the representatives, and the k-dimensional representatives are in a $(k-1)$-dimensional hyperplane of $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively.

- Note that the dimension k does not play an important role here, the vector components can be included successively up to a k such that $s_{k-1} > s_k$. We remark that the singular vectors can arbitrarily be chosen in the isotropic subspaces corresponding to possible multiple singular values, under the orthogonality conditions.

- As for the joint distribution view (when the rows and columns belong to the categories of two categorical variables), this representation has the following optimum properties: the closeness of categories of the same variable reflects the similarity between them, while the closeness of categories of the two different variables reflects their frequent simultaneous occurrence. For example, $\boldsymbol{C}$ being a microarray, the representatives of similar function genes as well as representatives of similar conditions are close to each other; likewise, representatives of genes that are responsible for a given condition are close to the representatives of those conditions.

- One frequently studied example of a rectangular array is the keyword–document matrix. Here the entries are associations between documents and words. Based on network data, the entry in the ith row and jth column is the relative frequency of word j in document i. Latent semantic indexing looks for real scores of the documents and keywords such that the score of a any document be proportional to the total scores of the keywords occurring in it, and vice versa, the score of any keyword be proportional to the total scores of the documents containing it. Not surprisingly, the solution is given by the SVD of the contingency table, where the document- and keyword-scores are the coordinates of the left and right singular vectors corresponding to its largest non-trivial singular value which gives the constant of proportionality. This idea is generalized in [32] in the following way. We can think of the above relation between keywords and documents as the relation with respect to the most important topic (or context, or factor). After this, we are looking for another scoring with respect to the second topic, up to k (where k is a positive integer not exceeding the rank of the table). The solution is given by the singular vector pairs corresponding to the k largest singular values of the table. The problem is also related to the Pagerank, see e.g. [39].

- In another view, a contingency table can be considered as part of the weight matrix of a bipartite graph on vertex set $Row \cup Col$. this bipartite graph However, it would be hard to always distinguish between these two types of vertices, we rather use the framework of correspondence analysis, and formulate our statements in terms of rows and columns.

2 Normalized Laplacian spectra

Let $G = (V, \boldsymbol{W}, \boldsymbol{S})$ be a weighted graph on the vertex-set V $(|V| = n)$, where both the edges and vertices have nonnegative weights. The edge-weights are entries of $\boldsymbol{W}$, whereas the diagonal matrix $\boldsymbol{S} = \operatorname{diag}(s_1, \dots, s_n)$ contains the positive vertex-weights in its main diagonal. Without loss of generality, we can assume that the entries in $\boldsymbol{W}$ and $\boldsymbol{S}$ both sum to 1. For the time being, the vertex-weights have nothing to do with the edge-weights. These individual weights are assigned to the vertices subjectively. For example, in a social network, the edge-weights are similarities between the vertices based on the strengths of their pairwise connections (like frequency of co-starring of artists), while vertex-weights embody the individual strengths of the vertices in the network (like the actors' individual abilities). We will further motivate this idea in later.

Now, we look for k-dimensional representatives $\mathbf{r}_1, \dots, \mathbf{r}_n$ of the vertices so that they minimize the objective function $Q_k = \sum_{i<j} w_{ij} \|\mathbf{r}_i - \mathbf{r}_j\|^2$ subject to

$$\sum_{i=1}^{n} s_i \mathbf{r}_i \mathbf{r}_i^T = \boldsymbol{I}_k.$$

With the notation and considerations of Lesson 1,

$$\begin{aligned}\min_{\sum_{i=1}^n s_i \mathbf{r}_i \mathbf{r}_i^T = \boldsymbol{I}_k} Q_k &= \min_{\boldsymbol{X}^T \boldsymbol{S} \boldsymbol{X} = \boldsymbol{I}_k} \operatorname{tr}(\boldsymbol{X}^T \boldsymbol{L} \boldsymbol{X}) \\ &= \min_{\boldsymbol{X}^T \boldsymbol{S} \boldsymbol{X} = \boldsymbol{I}_k} \operatorname{tr}[(\boldsymbol{S}^{1/2}\boldsymbol{X})^T (\boldsymbol{S}^{-1/2}\boldsymbol{L}\boldsymbol{S}^{-1/2})(\boldsymbol{S}^{1/2}\boldsymbol{X})] \\ &= \sum_{i=0}^{k-1} \lambda_i(\boldsymbol{L}_S) = \sum_{i=1}^{k-1} \lambda_i(\boldsymbol{L}_S)\end{aligned}$$

where $\boldsymbol{L}_S = \boldsymbol{S}^{-1/2}\boldsymbol{L}\boldsymbol{S}^{-1/2}$ is the Laplacian, normalized by $\boldsymbol{S}$, and because of the constraints, $\boldsymbol{S}^{1/2}\boldsymbol{X}$ is a suborthogonal matrix. Obviously, $\boldsymbol{L}_S$ is also positive semidefinite with eigenvalues $0 = \lambda_0(\boldsymbol{L}_S) \le \lambda_1(\boldsymbol{L}_S) \le \dots \le \lambda_{n-1}(\boldsymbol{L}_S)$ and corresponding orthonormal eigenvectors $\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_{n-1}$. Furthermore, 0 is a single eigenvalue if and only if G is connected. The optimum k-dimensional representation is obtained by the row vectors of the matrix $\boldsymbol{S}^{-1/2}(\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_{k-1})$.

The special case, when the vertex-weights are the generalized degrees, that is $\boldsymbol{S} = \boldsymbol{D}$, has a distinguished importance.

Definition 2 *The matrix*

$$\boldsymbol{L}_D = \boldsymbol{D}^{-1/2}\boldsymbol{L}\boldsymbol{D}^{-1/2} = \boldsymbol{I}_n - \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$$

is called the normalized Laplacian of the edge-weighted graph $G = (V, \boldsymbol{W})$.

Remark 1 *Now, we enumerate some simple statements concerning the normalized Laplacian spectrum.*

(i) Since the matrix $\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$ is the correspondence matrix corresponding to the symmetric contingency table $\boldsymbol{W}$, the singular values of this matrix are in the $[0,1]$ interval: they are special correlations, the largest one being 1 (see Section 1). Consequently, the eigenvalues of $\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$ are in the $[-1,1]$ interval, while those of $\boldsymbol{I}_n - \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$ in the $[0,2]$ interval. Let

$$0 = \lambda_0 \le \lambda_1 \le \dots \le \lambda_{n-1} \le 2$$

denote the spectrum of the normalized Laplacian $\boldsymbol{L}_D$.

(ii) Trivially, 0 is a single eigenvalue of $\boldsymbol{L}_D$ if and only if G is connected (i.e. $\boldsymbol{W}$ is irreducible), and in this case, the corresponding unit-norm eigenvector is the $\sqrt{\mathbf{d}} = (\sqrt{d_1}, \dots, \sqrt{d_n})^T$ vector. Furthermore, the normalized Laplacian spectrum of a disconnected graph is the union of those of its connected components.

(iii) Since $\sum_{i=0}^{n-1} \lambda_i = \operatorname{tr}(\boldsymbol{L}_D) = n$, the following estimations for the smallest and largest positive normalized Laplacian eigenvalues of the connected edge-weighted graph $G = (V, \boldsymbol{W})$ on n vertices hold:

$$\lambda_1 = \min_{i \in \{1,\dots,n-1\}} \lambda_i \le \frac{1}{n-1}\sum_{i=1}^{n-1} \lambda_i = \frac{n}{n-1} \le \max_{i \in \{1,\dots,n-1\}} \lambda_i = \lambda_{n-1}.$$

Note that both of the above inequalities hold with equality at the same time, if and only if G is the complete graph (see the forthcoming Example (A)).

(iv) *For a simple graph G, which is not the complete graph, $\lambda_1 \le 1$ holds. For the proof see [25].*

(v) *Provided G is connected, 2 is an eigenvalue if and only if G is a bipartite graph (i.e. its vertices can be divided into two parts such that there are no edges within these two vertex-subsets, or equivalently, after permuting its rows and columns in the same way, $\boldsymbol{W}$ contains two zero diagonal blocks). The proof for simple graphs is found in [25], and for general edge-weighted graphs follows from the following considerations. The bipartedness of a connected G is equivalent to the fact that its weight matrix $\boldsymbol{W}$ is irreducible, but decomposable. With the arguments of (i), this property extends to the correspondence matrix, which has therefore a multiple singular value 1. Since 1 is a single eigenvalue of $\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$ (thanks to G connected), it must have the -1 as an eigenvalue, which results in the eigenvalue 2 of $\boldsymbol{L}_D$.*

Next, we enlist the normalized Laplacian spectra of some simple graphs. Fially, for regular graphs, the Laplacian and normalized Laplacian eigenvalues are constant multiples of each other.

(A) Since the complete graph C_n is $(n-1)$-regular, its normalized Laplacian eigenvalues are the $\frac{1}{n-1}$ multiples of the Laplacian ones. Therefore,

$$\lambda_0 = 0,\ \lambda_1 = \cdots = \lambda_{n-1} = \frac{n}{n-1}.$$

(B) For the path graph P_n,

$$\lambda_i = 1 - \cos\frac{i\pi}{n-1}, \quad i = 0, 1, \dots, n-1.$$

Note that the largest eigenvalue is 2, since P_n is bipartite. Indeed, let us label the vertices in their natural succession. Then the vertices of odd and even labels constitute the two independent vertex subsets of the bipartite graph. We also remark that for large n, P_n is almost regular, in other words, its degree-matrix is close to $2I_n$. Therefore the normalized Laplacian eigenvalues of P_n are asymptotically $\frac{1}{2}$ multiples of the Laplacian ones, see Example (b) of Lesson 1.

(C) For the d-dimensional hypercube Q_d on 2^d vertices, based on the adjacency spectrum (derived in [40]), the normalized Laplacian eigenvalues are the numbers $\frac{2i}{d}$ with multiplicity $\binom{d}{i}$, $i = 0, 1, \dots, d$. Hence, 2 is a single eigenvalue of Q_d, which fact is not surprising, since Q_d is bipartite again.

(D) The normalized Laplacian eigenvalues of the complete bipartite graph K_{n_1,n_2} are

$$\lambda_0 = 0,\ \lambda_1 = \cdots = \lambda_{n_1+n_2-2} = 1,\ \lambda_{n_1+n_2-1} = 2.$$

(E) Especially, the star graph S_d on $d+1$ vertices is the $K_{1,d}$ graph, therefore its eigenvalues are

$$\lambda_0 = 0,\ \lambda_1 = \dots, \lambda_{d-1} = 1,\ \lambda_d = 2.$$

Normalized Laplacian was used for spectral clustering in several papers (see, e.g., [11, 18]). Those results will be based on the observation that the spectral decomposition (briefly, SD) of $\boldsymbol{L}_D$ solves the following quadratic placement problem.

Theorem 2 (Representation theorem for edge- and vertex-weighted graphs) *Let $G = (V, \boldsymbol{W})$ be a connected edge-weighted graph with normalized Laplacian $\boldsymbol{L}_D$. Let $0 = \lambda_0 < \lambda_1 \le \dots \le \lambda_{n-1}$ be the eigenvalues of $\boldsymbol{L}_D$ with corresponding unit-norm eigenvectors $\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_{n-1}$. Let $k < n$ be a positive integer such that $\lambda_{k-1} < \lambda_k$. Then the minimum of Q_{k-1} subject to*

$$\sum_{i=1}^{n} d_i \mathbf{r}_i \mathbf{r}_i^T = \boldsymbol{I}_{k-1} \quad \textit{and} \quad \sum_{i=1}^{n} d_i \mathbf{r}_i = \mathbf{0}$$

is $\sum_{i=1}^{k-1} \lambda_i$ and it is attained with the optimum $(k-1)$-dimensional representatives $\mathbf{r}_1^*, \dots, \mathbf{r}_n^*$ the transposes of which are row vectors of $\boldsymbol{X}^* = \boldsymbol{D}^{-1/2}(\mathbf{u}_1, \dots, \mathbf{u}_{k-1})$.

Proof 2 *Observe that instead of $\boldsymbol{X}$, the augmented $n \times k$ matrix $\tilde{\boldsymbol{X}}$ can as well be used, which is obtained from $\boldsymbol{X}$ by inserting the column $\mathbf{x}_0 = \mathbf{1}$ of all 1's. In fact, $\mathbf{x}_0 = \boldsymbol{D}^{-1/2}\mathbf{u}_0$, where $\mathbf{u}_0 = \sqrt{\mathbf{d}}$ is the eigenvector corresponding to the eigenvalue 0 of $\boldsymbol{L}_D$. Then*

$$\begin{aligned} \min_{\substack{\sum_{i=1}^n d_i \mathbf{r}_i \mathbf{r}_i^T = \boldsymbol{I}_{k-1} \\ \sum_{i=1}^n d_i \mathbf{r}_i = \mathbf{0}}} Q_{k-1} &= \min_{\substack{\boldsymbol{X}^T \boldsymbol{D} \boldsymbol{X} = \boldsymbol{I}_{k-1} \\ \boldsymbol{X}^T \boldsymbol{D} \mathbf{1} = \mathbf{0}}} \operatorname{tr}[(\boldsymbol{D}^{1/2}\boldsymbol{X})^T \boldsymbol{L}_D (\boldsymbol{D}^{1/2}\boldsymbol{X})] \\ &= \min_{\tilde{\boldsymbol{X}}^T \boldsymbol{D} \tilde{\boldsymbol{X}} = \boldsymbol{I}_k} \operatorname{tr}[(\boldsymbol{D}^{1/2}\tilde{\boldsymbol{X}})^T \boldsymbol{L}_D (\boldsymbol{D}^{1/2}\tilde{\boldsymbol{X}})] = \sum_{i=1}^{k-1} \lambda_i . \end{aligned} \tag{6}$$

Here we used that

$$\begin{aligned} \operatorname{tr}[(\boldsymbol{D}^{1/2}\tilde{\boldsymbol{X}})^T \boldsymbol{L}_D (\boldsymbol{D}^{1/2}\tilde{\boldsymbol{X}})] &= \sum_{\ell=0}^{k-1} (\boldsymbol{D}^{1/2}\mathbf{x}_\ell)^T \boldsymbol{L}_D (\boldsymbol{D}^{1/2}\mathbf{x}_\ell) \\ &= \sum_{\ell=1}^{k-1} (\boldsymbol{D}^{1/2}\mathbf{x}_\ell)^T \boldsymbol{L}_D (\boldsymbol{D}^{1/2}\mathbf{x}_\ell) \end{aligned}$$

because of the relation $\boldsymbol{L}_D(\boldsymbol{D}^{1/2}\mathbf{x}_\ell) = \boldsymbol{L}_D\sqrt{\mathbf{d}} = \mathbf{0}$, since $\sqrt{\mathbf{d}} = \mathbf{u}_0$ is the unit-norm eigenvector of $\boldsymbol{L}_D$ corresponding to the eigenvalue 0. It is important that G is connected and $\boldsymbol{W}$ is normalized such that $\sum_{i=1}^n d_i = 1$.

3 Modularity spectra

The *modularity matrix* $\boldsymbol{M}$ was defined by [46, 47] for simple graphs and naturally extends to edge-weighted graphs (see [21]) as

$$\boldsymbol{M} = \boldsymbol{W} - \mathbf{d}\mathbf{d}^T \tag{7}$$

which is the negative of the so-called Q-Laplacian introduced in [64]. It is easy to see that 0 is always an eigenvalue of $\boldsymbol{M}$ with corresponding eigendirection $\mathbf{1}$. However, it is not true that the modularity spectrum of a disconnected graph is the union of modularity spectra of its components, and the above minima are not related immediately to the eigenvalues of this modularity matrix. In case of simple graphs, $\boldsymbol{M}$ is usually indefinite, and it is negative definite for complete or complete multipartite graphs. In [22] we proved that the complete and complete multipartite graphs are the only ones for which the largest modularity eigenvalue is 0.

In [21] we introduced the following normalized version of the modularity matrix.

Definition 3 *Let $G = (V, \boldsymbol{W})$ be an edge-weighted graph with the entries of $\boldsymbol{W}$ summing up to 1. The matrix*

$$\boldsymbol{M}_D = \boldsymbol{D}^{-1/2}\boldsymbol{M}\boldsymbol{D}^{-1/2} = \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2} - \sqrt{\mathbf{d}}\sqrt{\mathbf{d}}^T \tag{8}$$

is called normalized modularity matrix of G.

As we have established, the eigenvalues of $\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$ are in the $[-1, 1]$ interval; the largest eigenvalue is always 1 with corresponding unit-norm eigenvector $\sqrt{\mathbf{d}}$. The only non-zero eigenvalue of the rank 1 term $\sqrt{\mathbf{d}}\sqrt{\mathbf{d}}^T$ is also 1 with the same eigenvector. Therefore, the spectrum of the matrix $\boldsymbol{M}_D$ is the same as the spectrum of $\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$, with the only exception that – due to the subtraction of the term $\sqrt{\mathbf{d}}\sqrt{\mathbf{d}}^T$ – the eigenvalue 1 of $\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$ becomes an eigenvalue 0 of $\boldsymbol{M}_D$ with eigenvector $\sqrt{\mathbf{d}}$. Hence, the spectrum of $\boldsymbol{M}_D$ is in $[-1, 1]$ and includes the 0.

These considerations give an exact relation between the normalized Laplacian and modularity matrix: $\boldsymbol{M}_D = \boldsymbol{I} - \boldsymbol{L}_D - \sqrt{\mathbf{d}}\sqrt{\mathbf{d}}^T$. If the eigenvalues of $\boldsymbol{L}_D$ are $0 = \lambda_0 \le \lambda_1 \cdots \le \lambda_{n-1} \le 2$, then the spectrum of $\boldsymbol{M}_D$ consists of the numbers $1 - \lambda_i$ $(i = 1, \dots, n-1)$ and the zero with corresponding eigenvector $\sqrt{\mathbf{d}}$. Further, the multiplicity of 0 is one more than the multiplicity of the eigenvalue 1 of $\boldsymbol{L}_D$. The multiplicity of 1 is one less than multiplicity of the eigenvalue 0 of $\boldsymbol{L}_D$; hence, 1 cannot be an eigenvalue of $\boldsymbol{M}_D$ if G is connected.

In terms of the normalized modularity matrix, the minimization problem (6) can be formulated as a maximization task in the following way.

$$\begin{aligned}
&\max_{\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X}=\boldsymbol{I}_{k-1}} \operatorname{tr}[(\boldsymbol{D}^{1/2}\boldsymbol{X})^T\boldsymbol{M}_D(\boldsymbol{D}^{1/2}\boldsymbol{X})]\\
=&\max_{\substack{\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X}=\boldsymbol{I}_{k-1}\\ \boldsymbol{X}^T\boldsymbol{D}\mathbf{1}=\mathbf{0}}} \operatorname{tr}[(\boldsymbol{D}^{1/2}\boldsymbol{X})^T(\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2})(\boldsymbol{D}^{1/2}\boldsymbol{X})]\\
=&k-1-\min_{\substack{\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{X}=\boldsymbol{I}_{k-1}\\ \boldsymbol{X}^T\boldsymbol{D}\mathbf{1}=\mathbf{0}}} \operatorname{tr}[(\boldsymbol{D}^{1/2}\boldsymbol{X})^T(\boldsymbol{I}_n-\boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2})(\boldsymbol{D}^{1/2}\boldsymbol{X})]\\
=&k-1-\min_{\substack{\sum_{i=1}^n d_i\mathbf{r}_i\mathbf{r}_i^T=\boldsymbol{I}_{k-1}\\ \sum_{i=1}^n d_i\mathbf{r}_i=\mathbf{0}}} Q_{k-1}.
\end{aligned}$$

The maximum is $k - 1 - \sum_{i=1}^{k-1}\lambda_i = \sum_{i=1}^{k-1}(1 - \lambda_i)$, that is the sum of the $k - 1$ largest eigenvalues of $\boldsymbol{M}_D$.

4 Representation of joint distributions

[This section can be skipped.]

In this section we would like to give an abstract description of the issues discussed in the previous sections in terms of two-variate distributions. With the help of joint distributions, representation can be discussed in a more general framework, of which graphs and contingency tables are special finite cases. This section is rather of theoretical importance, however, some parts of it will appear later when we will take limits of graphs and contingency tables, and consider the continuous limit objects as kernels of integral operators taking conditional expectation with respect to the joint distributions. Here and in the next section we will intensively use the theory of Hilbert spaces and linear operators between them; further, distribution of random vectors.

4.1 General setup

Let (ξ, η) be a pair of real-valued random variables – neither of them being constant with probability 1 – defined over the product space $\mathcal{X} \times \mathcal{Y}$ having joint distribution $\mathbb{W}$ with marginals $\mathbb{P}$ and $\mathbb{Q}$, respectively. Assume that the dependence between ξ and η is regular, i.e. their joint distribution $\mathbb{W}$ is absolutely continuous with respect to the product measure $\mathbb{P} \times \mathbb{Q}$, and let w denote the Radon–Nikodym derivative of $\mathbb{W}$ with respect to $\mathbb{P} \times \mathbb{Q}$, see [54] for details. In case of discrete or absolutely continuous distributions we will soon introduce more friendly versions of this derivative.

In the spirit of [24], let $H = L^2(\xi)$ and $H' = L^2(\eta)$ be the set of random variables which are functions of ξ and η and have zero expectation and finite variance with respect to $\mathbb{P}$ and $\mathbb{Q}$, respectively. Both H and H' are Hilbert spaces with the covariance as inner product; further, they are embedded as subspaces into the L^2 space defined likewise by the (ξ, η) pair over the product space. (Note that we consider Borel-measurable functions which are also measurable with respect to the so-called σ-algebras generated by ξ and η, but we do not want to introduce superfluous notions that will not be used later. For example, in case of discrete, especially categorical variables with finitely many values, a function of such a variable takes on as many values as the original one, with the same probabilities.)

4.2 Integral operators between L^2 spaces

Let $K : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a kernel such that for it

$$\int_{\mathcal{X}} \int_{\mathcal{Y}} K^2(x, y)\, \mathbb{P}(dx)\, \mathbb{Q}(dy) < \infty \tag{9}$$

holds. With the kernel K, a linear operator (integral operator) $A : H' \to H$ is defined in the following way: to the random variable $\phi \in H'$ the linear operator A assigns the random variable $\psi \in H$ such that

$$\psi(x) = (A\phi)(x) = \int_{\mathcal{Y}} K(x, y)\phi(y)\, \mathbb{Q}(dy), \quad x \in \mathcal{X}.$$

By the linearity of A, ψ has zero expectation, and it is easy to check that has finite variance; further

$$\|\psi\| \le \|K\|_2 \cdot \|\phi\| < \infty,$$

where $\|\psi\|$ and $\|\phi\|$ denote the standard deviation (squareroot of the variance) of the random variables ψ and ϕ, respectively, while $\|K\|_2$ is the squareroot of the finite integral in (9). Further, for the operator norm of A, the following holds:

$$\|A\| = \sup_{\|\phi\|=1} \|A\phi\| \leq \|K\|_2.$$

It is known that the above L^2 spaces are separable Hilbert spaces, and in view of (9), A is a Hilbert–Schmidt, therefore a compact (in other words, completely continuous) linear operator with SVD:

$$A = \sum_{i=1}^{\infty} s_i \langle ., \phi_i \rangle_{H'} \psi_i.$$

Here $\langle ., . \rangle$ denotes the real inner product (covariance) in the corresponding Hilbert space; further, the nonnegative real numbers $s_1 \geq s_2 \geq \cdots \geq 0$ are the singular values with the zero as the only possible point of accumulation, and the corresponding function pairs ψ_i, ϕ_i can be chosen (even in case of multiple singular values) so that $\{\psi_i\}_{i=1}^{\infty} \subset H$ and $\{\phi_i\}_{i=1}^{\infty} \subset H'$ form complete orthonormal systems. Since A is also a Hilbert–Schmidt operator, $\sum_{i=1}^{\infty} s_i^2 = \|K\|_2^2 < \infty$. Such an SVD is essentially unique (apart from function pairs corresponding to multiple singular values). It is easy to see that the adjoint of A (which is, in fact, the transpose, as we only deal with real valued random variables) has the SVD

$$A^* = \sum_{i=1}^{\infty} s_i \langle ., \psi_i \rangle_H \phi_i$$

and

$$A\phi_i = s_i \psi_i, \quad A^* \psi_i = s_i \phi_i, \quad i = 1, 2, \ldots$$

further, s_1 is the spectral norm of both A and A^*.

Now we proceed to the symmetric case: the joint distribution over the product space is symmetric, i.e. $w(x, y) = w(y, x)$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$. In this case ξ and η are identically distributed (not independent as their joint distribution is $\mathbb{W}$) and hence, each random variable in H has a counterpart in H' and vice versa, such that they are identically distributed; therefore H and H' are isomorphic in terms of the distributions, too. By the Hilbert–Schmidt theorem, the selfadjoint compact linear operator $A : H' \to H$ has SD

$$A = \sum_{i=1}^{\infty} \lambda_i \langle ., \psi_i' \rangle_{H'} \psi_i$$

with real eigenvalues whose only possible point of accumulation is the zero ($\lim_{i\to\infty} \lambda_i = 0$ if the eigenvalues are countably infinitely many), and corresponding orthonormal eigenvectors $\psi_1, \psi_2, \ldots$ such that ψ_i and ψ_i' are identically distributed with joint distribution $\mathbb{W}$. Of course, A also has a singular value decomposition, where the singular values are the absolute values of the eigenvalues; if $\lambda_i > 0$, then $s_i = \lambda_i$, $\psi_i = \phi_i$ and they coincide with the unit-norm eigenfunction; if $\lambda_i < 0$, then $s_i = -\lambda_i$, $\psi_i = -\phi_i$ and any of them can be the unit-norm eigenfunction corresponding to λ_i.

4.3 When the kernel is the joint distribution itself

The following linear operators taking conditional expectation between the two marginals (with respect to the joint distribution) will play a crucial role in the future, see also [24]. In the general, not necessarily symmetric case, they are defined as

$$P_{\mathcal{X}} : H' \to H, \quad \psi = P_{\mathcal{X}}\phi = \mathbb{E}(\phi \,|\, \xi), \quad \psi(x) = \int_{\mathcal{Y}} w(x,y)\phi(y)\,\mathbb{Q}(dy)$$

and

$$P_{\mathcal{Y}} : H \to H', \quad \phi = P_{\mathcal{Y}}\psi = \mathbb{E}(\psi \,|\, \eta), \quad \phi(y) = \int_{\mathcal{X}} w(x,y)\psi(x)\,\mathbb{P}(dx).$$

It is easy to see that $P_{\mathcal{X}}^* = P_{\mathcal{Y}}$ and vice versa, because of the relation

$$\langle P_{\mathcal{X}}\phi, \psi\rangle_H = \langle P_{\mathcal{Y}}\psi, \phi\rangle_{H'} = \mathrm{Cov}_{\mathbb{W}}(\psi, \phi), \tag{10}$$

where $\mathrm{Cov}_{\mathbb{W}}$ is the so-called covariance function with respect to the joint distribution $\mathbb{W}$, defined as

$$\mathrm{Cov}_{\mathbb{W}}(\psi, \phi) = \int_{\mathcal{X}\times\mathcal{Y}} \psi(x)\phi(y)\mathbb{W}(dx,dy) = \int_{\mathcal{X}}\int_{\mathcal{Y}} \psi(x)\phi(y)w(x,y)\mathbb{Q}(dy)\mathbb{P}(dx).$$

Assume that

$$\int_{\mathcal{X}}\int_{\mathcal{Y}} w^2(x,y)\mathbb{Q}(dy)\mathbb{P}(dx) < \infty. \tag{11}$$

In case of discrete distributions with joint distribution $\{w_{ij}\}$ and marginals $\{p_i\}$ ($p_i = \sum_j w_{ij}$) and $\{q_j\}$ ($q_j = \sum_i w_{ij}$), (11) means that

$$\sum_{i\in\mathcal{X}}\sum_{j\in\mathcal{Y}} \left(\frac{w_{ij}}{p_i q_j}\right)^2 p_i q_j = \sum_{i\in\mathcal{X}}\sum_{j\in\mathcal{Y}} \frac{w_{ij}^2}{p_i q_j} < \infty,$$

while in case of absolutely continuous distributions with joint p.d.f. $f(x,y)$ and marginal p.d.f.s $f_1(x)$ ($f_1(x) = \int f(x,y)\,dy$) and $f_2(y)$ ($f_2(y) = \int f(x,y)\,dx$), (11) means that or

$$\int_{\mathcal{X}}\int_{\mathcal{Y}} \left(\frac{f(x,y)}{f_1(x)f_2(y)}\right)^2 f_1(x)f_2(y)\,dx\,dy = \int_{\mathcal{X}}\int_{\mathcal{Y}} \frac{f^2(x,y)}{f_1(x)f_2(y)}\,dx\,dy < \infty.$$

Under these conditions $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$ are Hilbert–Schmidt operators, and therefore compact, with SVD

$$P_{\mathcal{X}} = \sum_{i=1}^{\infty} s_i \langle ., \phi_i\rangle_{H'}\psi_i, \quad P_{\mathcal{Y}} = \sum_{i=1}^{\infty} s_i \langle ., \psi_i\rangle_{H}\phi_i \tag{12}$$

where for the singular values $1 > s_1 \geq s_2 \geq \cdots \geq 0$ holds, since the operators $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$ are in fact orthogonal projections from one marginal onto the other, but the projections are restricted to the subspaces H and H', respectively. We remark that denoting by ψ_0 and ϕ_0 the constantly 1 random variables, $\mathbb{E}(\phi_0|\xi) = \psi_0$ and $\mathbb{E}(\psi_0|\eta) = \phi_0$, however, this pair is not considered as a function pair with singular value $s_0 = 1$, since they have no zero expectation.

Consequently, we will subtract 1 from the kernel, but with this new kernel, $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$ will define the same integral operators.

Especially, if $\mathbb{W}$ is symmetric (H and H' are isomorphic in terms of the distributions too), then in view of (10), $P_{\mathcal{X}} = P_{\mathcal{Y}}$ is a selfadjoint (symmetric) linear operator, since

$$\langle P_{\mathcal{X}}\phi, \psi\rangle_H = \mathrm{Cov}_{\mathbb{W}}(\phi, \psi) = \mathrm{Cov}_{\mathbb{W}}(\psi, \phi) = \langle P_{\mathcal{Y}}\psi, \phi\rangle_{H'}.$$

The SD of $P_{\mathcal{X}} : H' \to H$ is

$$P_{\mathcal{X}} = \sum_{i=1}^{\infty} \lambda_i \langle ., \psi_i'\rangle_{H'} \psi_i.$$

Here for the eigenvalues, $|\lambda_i| \leq 1$ holds, and the eigenvalue–eigenfunction equation looks like

$$P_{\mathcal{X}}\psi_i' = \lambda_i \psi_i$$

where ψ_i and ψ_i' are identically distributed, whereas their joint distribution is $\mathbb{W}$ ($i = 1, 2, \dots$).

4.4 Maximal correlation and optimal representations

From now on, we will intensively use separation theorems for the singular values and eigenvalues. In view of these, the SVD gives the solution of the following task of *maximal correlation*, posed by [33] and [53]. We are looking for $\psi \in H$, $\phi \in H'$ such that their correlation is maximum with respect to their joint distribution $\mathbb{W}$. Using (10),

$$\max_{\|\psi\|=\|\phi\|=1} \mathrm{Cov}_{\mathbb{W}}(\psi, \phi) = s_1$$

and it is attained on the non-trivial ψ_1, ϕ_1 pair. In the finite, symmetric case, maximal correlation is related to some conditional probabilities in [19].

This task is equivalent to the following one:

$$\min_{\|\psi\|=\|\phi\|=1} \|\psi - \phi\|^2 = \min_{\|\psi\|=\|\phi\|=1} (\|\psi\|^2 + \|\phi\|^2 - 2\mathrm{Cov}_{\mathbb{W}}(\psi, \phi)) = 2(1 - s_1). \quad (13)$$

Correspondence analysis is on the one hand, a special case of the problem of maximal correlation being $\mathcal{X}$ and $\mathcal{Y}$ finite sets, but on the other hand, it is a generalization in the extent that we are successively finding maximal correlations under some orthogonality conditions.

The product space is now an $m \times n$ contingency table with row set $\mathcal{X} = \{1, \dots, m\}$ and column set $\mathcal{Y} = \{1, \dots, n\}$, whereas the entries $w_{ij} \geq 0$ ($i = 1, \dots, m$; $j = 1, \dots, n$) embody the joint distribution over the product space, with row-sums $p_1, \dots, p_m$ and column-sums $q_1, \dots, q_n$ as marginals.

Hence, the effect of $P_{\mathcal{X}} : H' \to H$, $P_{\mathcal{X}}\phi = \psi$ is the following:

$$\psi(i) = \frac{1}{p_i} \sum_{j=1}^{n} w_{ij}\phi(j) = \sum_{j=1}^{n} \frac{w_{ij}}{p_i q_j} \phi(j) q_j, \quad i = 1, \dots, m. \quad (14)$$

Therefore, $P_{\mathcal{X}}$ is an integral operator with kernel $K_{ij} = \frac{w_{ij}}{p_i q_j}$ (instead of integration, we have summation with respect to the marginal measure $\mathbb{Q}$).

Consider the SVD

$$P_{\mathcal{X}} = \sum_{k=1}^{r-1} s_k \langle ., \phi_k \rangle_{H'} \psi_k$$

where r is the now finite rank of the contingency table ($r \le \min\{n, m\}$). The singular value $s_0 = 1$ with the trivial $\psi_0 = 1$, $\phi_0 = 1$ factor pair is disregarded as their expectation is 1 with respect to the $\mathbb{P}$- and $\mathbb{Q}$-measures, respectively, therefore, the summation starts from 1. If we used the kernel $K_{ij} - 1$, we could eliminate the trivial factors. Assume that there is no other singular value 1, i.e. our contingency table is non-decomposable. Then, by the orthogonality, the subsequent left- and right-hand side singular functions have zero expectation with respect to the $\mathbb{P}$- and $\mathbb{Q}$-measures, and they solve the following successive maximal correlation problem. For $k = 1, \dots, r-1$, in step k we want to find $\max \mathrm{Corr}_{\mathbb{W}}(\psi, \phi)$ subject to

$$\mathrm{Var}_{\mathbb{P}}(\psi) = \mathrm{Var}_{\mathbb{Q}}(\phi) = 1, \quad \mathrm{Cov}_{\mathbb{P}}(\psi, \psi_i) = \mathrm{Cov}_{\mathbb{Q}}(\phi, \phi_i) = 0, \quad 0 = 1, \dots, k-1.$$

Note that the last condition for $i = 0$ is equivalent to

$$\mathbb{E}_{\mathbb{P}}(\psi) = \mathbb{E}_{\mathbb{Q}}(\phi) = 0.$$

The maximum is s_k and it is attained on the ψ_k, ϕ_k pair.

Now, we are able to define the joint representation of the general Hilbert-spaces H, H' with respect to the joint measure $\mathbb{W}$ in the following way.

Definition 4 *We say that the pair $(\mathbf{X}, \mathbf{Y})$ of k-dimensional random vectors with components in H and H', respectively, form a k-dimensional representation of the product space endowed with the measure $\mathbb{W}$ if $\mathbb{E}_{\mathbb{P}}\mathbf{X}\mathbf{X}^T = \boldsymbol{I}_k$ and $\mathbb{E}_{\mathbb{Q}}\mathbf{Y}\mathbf{Y}^T = \boldsymbol{I}_k$ (i.e. the components of $\mathbf{X}$ and $\mathbf{Y}$ are uncorrelated with zero expectation and unit variance, respectively). Further, the cost of this representation is defined as*

$$Q_k(\mathbf{X}, \mathbf{Y}) = \mathbb{E}_{\mathbb{W}} \|\mathbf{X} - \mathbf{Y}\|^2.$$

The pair $(\mathbf{X}^*, \mathbf{Y}^*)$ is an optimal representation if it minimizes the above cost.

Theorem 3 (Representation theorem for joint distributions) *Let $\mathbb{W}$ be a joint distribution with marginals $\mathbb{P}$ and $\mathbb{Q}$. Assume that among the singular values of the conditional expectation operator $P_{\mathcal{X}} : H' \to H$ (see (12)) there are at least k positive ones and denote by $1 > s_1 \ge s_2 \ge \dots \ge s_k > 0$ the largest ones. The minimum cost of a k-dimensional representation is $2\sum_{i=1}^{k}(1 - s_i)$ and it is attained with $\mathbf{X}^* = (\psi_1, \dots, \psi_k)$ and $\mathbf{Y}^* = (\phi_1, \dots, \phi_k)$, where ψ_i, ϕ_i is the singular function pair corresponding to the singular value s_i $(i = 1, \dots, k)$.*

Proof 3

$$\begin{aligned}
\mathbb{E}_{\mathbb{W}}(\mathbf{X} - \mathbf{Y})^T(\mathbf{X} - \mathbf{Y}) &= \mathbb{E}_{\mathbb{W}}(\mathbf{X}^T\mathbf{X}) + \mathbb{E}_{\mathbb{W}}(\mathbf{Y}^T\mathbf{Y}) - \mathbb{E}_{\mathbb{W}}\mathbf{X}^T\mathbf{Y} - \mathbb{E}_{\mathbb{W}}\mathbf{Y}^T\mathbf{X} \\
&= \mathbb{E}_{\mathbb{P}}(\mathrm{tr}[\mathbf{X}\mathbf{X}^T]) + \mathbb{E}_{\mathbb{Q}}(\mathrm{tr}[\mathbf{Y}\mathbf{Y}^T]) - 2\sum_{i=1}^{k} \mathbb{E}_{\mathbb{W}}(X_i Y_i) \\
&= 2k - 2\sum_{i=1}^{k} \mathrm{Cov}(X_i Y_i).
\end{aligned}$$

Applying the statement for the singular values of the conditional expectation operator, the required statement is obtained.

We remark that in case of a finite $\mathcal{X}$ and $\mathcal{Y}$, the solution corresponds to the SVD of the correspondence matrix. Though, the correspondence matrix seemingly does not have the same normalization as the kernel, but our numerical algorithm for the SVD of a rectangular matrix is capable to find orthogonal eigenvectors in the usual Euclidean norm, which corresponds to the Lebesgue measure and not to the $\mathbb{P}$- or $\mathbb{Q}$-measures. Observe that in case of an irreducible contingency table, $s_i < 1$ $(i = 1, \dots, k)$, therefore the minimum cost is strictly positive.

In the symmetric case we can also define a representation. Now the $\mathbf{X}, \mathbf{X}'$ pair is identically distributed, but not independent as they are connected with the symmetric joint measure $\mathbb{W}$.

Definition 5 *We say that the k-dimensional random vector $\mathbf{X}$ with components in H forms a k-dimensional representation of the product space $H \times H'$ (H and H' are isomorphic) endowed with the symmetric measure $\mathbb{W}$ (and marginal $\mathbb{P}$) if $\mathbb{E}_{\mathbb{P}} \mathbf{X}\mathbf{X}^T = \boldsymbol{I}_k$ (i.e. the components of $\mathbf{X}$ are uncorrelated with zero expectation and unit variance). Further, the cost of this representation is defined as*

$$Q_k(\mathbf{X}) = \mathbb{E}_{\mathbb{W}} \|\mathbf{X} - \mathbf{X}'\|^2,$$

where $\mathbf{X}$ and $\mathbf{X}'$ are identically distributed and the joint distribution of their coordinates X_i and X_i' is $\mathbb{W}$ $(i = 1, \dots, k)$, while X_i and X_j' are uncorrelated if $i \neq j$. The random vector $\mathbf{X}^*$ is an optimal representation if it minimizes the above cost.

Theorem 4 (Representation theorem for symmetric joint distributions) *Let $\mathbb{W}$ be a symmetric joint distribution with marginal $\mathbb{P}$. Assume that among the eigenvalues of the conditional expectation operator $P_{\mathcal{X}} : H' \to H$ (H and H' are isomorphic) there are at least k positive ones and denote by $1 > \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k > 0$ the largest ones. Then the minimum cost of a k-dimensional representation is $2 \sum_{i=1}^k (1 - \lambda_i)$ and it is attained by $\mathbf{X}^* = (\psi_1, \dots, \psi_k)$ where ψ_i is the eigenfunction corresponding to the eigenvalue λ_i $(i = 1, \dots k)$.*

In case of a finite $\mathcal{X}$ (vertex set of an edge-weighted graph), we have a weighted graph with edge-weights w_{ij} $(\sum_{i=1}^n \sum_{j=1}^n w_{ij} = 1)$. The operator $P_{\mathcal{X}}$ deprived of the trivial factor corresponds to its normalized modularity matrix with eigenvalues in the [-1,1] interval (1 cannot be an eigenvalue if the underlying graph is connected), and eigenfunctions which are the transformed eigenvectors. As the numerical algorithm gives an orthonormal set of eigenvectors in Euclidean norm, some back-transformation is needed to get uncorrelated components with unit variance, therefore we use the normalized modularity matrix instead of the kernel $K_{ij} = \frac{w_{ij}}{d_i d_j}$ expected from (14), where $d_i = \sum_{j \in \mathcal{X}} w_{ij}$, $i \in \mathcal{X}$, the generalized degrees of the vertices.

We remark that the above formula for the kernel corresponds to the so-called copula transformation of the joint distribution $\mathbb{W}$ into the unit square, see [**?**]. This idea appears when vertex- and edge-weighted graphs are transformed into piecewise constant functions over $[0,1] \times [0,1]$, see the definition of graphons later. This transformation can be done in the general non-symmetric and non-finite cases too. Also observe that neither the kernel nor the contingency table or graph is changed under measure preserving transformations of $\mathcal{X}$, see the theory of exchangeable sequences and arrays.

We also remark that any or both of the starting random variables ξ, η can as well be a random vector (with real components). For example, if they have p- and q-dimensional Gaussian distribution respectively, than their maximum correlation is the largest canonical correlation between them, and it is realized by appropriate linear combinations of the components of ξ and η, respectively. Moreover, we can find canonical correlations one after the other with corresponding function pairs (under some orthogonality constraints), as many as the rank of the cross-covariance matrix of ξ and η. In fact, the whole decomposition relies on the SVD of a matrix calculated from this cross-covariance matrix and the individual covariance matrices of ξ and η. Hence, in many cases, the maximum correlation problem can be dealt more generally by means of SVD or SD.

5 Treating nonlinearities via reproducing kernel Hilbert spaces

There are fairy tales about some fictitious spaces, where everything is 'smooth' and 'linear'. Such spaces really exist, the hard part is that we should adopt them to our data. Good news is that it is not necessary to actually map our data into them, it suffices to treat only a kernel function, but the bad news is that the kernel must be appropriately selected so that the underlying nonlinearity could be detected.

Reproducing kernel Hilbert spaces were introduced in the middle of the 20th century by [9, 50], and others, but the theory itself is an elegant application of already known theorems of functional analysis, first of all the Riesz–Fréchet theorem and the theory of integral operators, see the works of [56, 57] tracing back to the beginning of the 20ieth century. Later on, in the last decades of the 20ieth century and even in our days, reproducing kernel Hilbert spaces are several times reinvented and applied in modern statistical methods and data mining, for example in [13, 58]. But what is the mystery of reproducing kernels and what is the diabolic kernel trick? We would like to reveal this secret and show the technical advantages of this artificially constructed creature. We will start with the motivation for using this concept of data representation.

A popular approach to data clustering (sometimes this is called spectral clustering) is the following. Our data points $\mathbf{x}_1, \dots, \mathbf{x}_n$ are already in a metric space, called *data space*, but cannot be well classified by the k-means algorithm for no integer $0 < k < n$. For example, there are obviously two clusters of points in $\mathbb{R}^2$, but they are separated by an annulus and the k-means algorithm with $k = 2$ is not able to find them, see [52]. However, we can map the points into a usually higher dimensional, or more abstract space with a non-linear mapping so that the images are already well clustered in the new, so-called *feature space*. At the end of this section, we will give a mapping of the points into $\mathbb{R}^3$ using a second degree polynomial kernel, because we want to make a separation with second degree curves linear.

In practical higher-dimensional problems, when we do not have the faintest idea about the clusters and no visualization is possible, unfortunately, we cannot give such mappings explicitly. Moreover, the feature space usually has much higher dimension than the original one, which fact is frequently referred to as the *curse of dimensionality*. However, the point of the kernel method to be introduced is that it is not even necessary to perform the mapping, it suffices to select a kernel – based on the inner product of the original points – that is no longer a linear kernel, but a more complicated, still admissible kernel (the exact meaning is given in Definition 7), and defines a new inner product within the feature space. Then we process statistical algorithms that need only this kernel and nothing else.

The feature space above is the counterpart of a so-called *reproducing kernel Hilbert space* that we will introduce right now together with the correspondence between it and the feature space. For example, finite dimensional Euclidean spaces (vector spaces with the usual inner product, e.g. $\mathbb{R}^p$) or the $L^2(\mathcal{X})$ space of real-valued, square integrable functions with respect to some finite measure on the compact set $\mathcal{X}$ (where the inner product of two functions is the integral of their product on $\mathcal{X}$) are such.

5.1 Notion of the reproducing kernel

A stronger condition imposed on a Hilbert space $\mathcal{H}$ of functions $\mathcal{X} \to \mathbb{R}$ (where $\mathcal{X}$ is an arbitrary set, for the time being) is that the following so-called evaluation mapping be a continuous, or equivalently, a bounded linear functional. The evaluation mapping $L_x : \mathcal{H} \to \mathbb{R}$ works on an $f \in \mathcal{H}$ such that

$$L_x(f) = f(x). \tag{15}$$

Definition 6 *A Hilbert space $\mathcal{H}$ of (real) functions on the set $\mathcal{X}$ is a reproducing kernel Hilbert space, briefly RKHS, if the point evaluation functional L_x of (15) exists and is continuous for all $x \in \mathcal{X}$.*

The name reproducing kernel comes from the fact that – by the Riesz–Fréchet representation theorem – the result of such a continuous mapping can be written as an inner product. This theorem states that a Hilbert space (in our case $\mathcal{H}$) and its dual (in our case the set of $\mathcal{H} \to \mathbb{R}$ continuous linear functionals, e.g. L_x) are isometrically isomorphic. Therefore, to any L_x there uniquely corresponds a $K_x \in \mathcal{H}$ such that

$$L_x(f) = \langle f, K_x \rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H}. \tag{16}$$

Since K_x is itself an $\mathcal{X} \to \mathbb{R}$ function, it can be evaluated at any point $y \in \mathcal{X}$. We define the bivariate function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as

$$K(x, y) := K_x(y) \tag{17}$$

and call it the *reproducing kernel* for the Hilbert space $\mathcal{H}$. Then using formulas (15), (16), and (17), we get that on the one hand,

$$K(x, y) = K_x(y) = L_y(K_x) = \langle K_x, K_y \rangle_{\mathcal{H}},$$

and on the other hand,

$$K(y, x) = K_y(x) = L_x(K_y) = \langle K_y, K_x \rangle_{\mathcal{H}}.$$

By the symmetry of the (real) inner product it follows that the reproducing kernel is symmetric and it is also reproduced as the inner product of special functions in the RKHS:

$$K(x, y) = \langle K_x, K_y \rangle_{\mathcal{H}} = \langle K(x, .), K(., y) \rangle_{\mathcal{H}}, \tag{18}$$

hence, K is positive definite (for the precise notion see the forthcoming Definition 7). This is the diabolic kernel trick. In this way, any point is represented by its similarity to all other points, see [8, 61] for more details.

Vice versa, if we are given a positive definite kernel function on $\mathcal{X} \times \mathcal{X}$ at the beginning, then there exists an RKHS such that with appropriate elements of it, the inner product relation (18) holds. (In fact, we are not given, we just select an appropriate kernel function.) The mystery of RKHS just lies in this converse statement.

For this purpose, let us first define the most important types of kernel functions and discuss how more and more complicated ones can be derived from the simplest ones.

Definition 7 *A symmetric two-variate function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called positive definite kernel (equivalently, admissible, valid, or Mercer kernel) if for any $n \in \mathbb{N}$ and $x_1, \dots, x_n \in \mathcal{X}$, the symmetric matrix of entries $K(x_i, x_j) = K(x_j, x_i)$ $(i, j = 1, \dots n)$ is positive semidefinite.*

We remark that a symmetric real matrix is positive semidefinite if and only if it is a Gram matrix, and hence, its entries become inner products, but usually not of the entries in its arguments. However, the simplest kernel function, the so-called *linear kernel*, does this job. It is defined as

$$K_{\mathrm{lin}}(x, y) = \langle x, y \rangle_{\mathcal{X}},$$

if $\mathcal{X}$ is subset of a Euclidean space.

From a valid kernel, one can get other valid kernels with the following operations:

1. If $K_1, K_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ are positive definite kernels, then the kernel K defined by $K(x, y) = K_1(x, y) + K_2(x, y)$ is also positive definite.

2. If $K_1, K_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ are positive definite kernels, then the kernel K defined by $K(x, y) = K_1(x, y)K_2(x, y)$ is also positive definite. Especially, if K is a positive definite kernel, then so does cK with any $c > 0$.

The first statement is trivial, the second one follows from the following proposition.

Proposition 1 *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be $n \times n$ symmetric, positive semidefinite matrices. Then the matrix $\boldsymbol{C}$ of entries $c_{ij} = a_{ij} b_{ij}$ $(i, j = 1, \dots n)$ is also symmetric, positive semidefinite.*

Proof 4 *The symmetry of $\boldsymbol{C}$ is trivial. Any symmetric, positive semidefinite matrix is a Gram-matrix, and hence, can be considered as the covariance matrix of an n-dimensional random vector. For the simplicity, let $\mathbf{X} = (X_1, \dots, X_n)^T \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{A})$ and $\mathbf{Y} = (Y_1, \dots, Y_n)^T \sim \mathcal{N}_n(\mathbf{0}, \boldsymbol{B})$ be independent Gaussian random vectors. We define the random vector $\mathbf{Z}$ in the following way:*

$$\mathbf{Z} := (X_1 Y_1, \dots, X_n Y_n)^T.$$

It is easy to see that $\mathbb{E}(\mathbf{Z}) = \mathbf{0}$, since $\mathbb{E}(X_i Y_i) = \mathrm{Cov}(X_i, Y_i) + \mathbb{E}(X_i) \cdot \mathbb{E}(Y_i) = 0$, $i = 1, \dots, n$. If we verify that the covariance matrix of $\mathbf{Z}$ is $\boldsymbol{C}$ then we are ready *as a covariance matrix is always positive semidefinite. Indeed, the ij-th entry of* $\mathbb{E}(\mathbf{Z}\mathbf{Z}^T)$ *is*

$$\mathbb{E}(X_iY_iX_jY_j) = \mathbb{E}([X_iX_j]\cdot[Y_iY_j]) = \mathbb{E}(X_iX_j)\cdot\mathbb{E}(Y_iY_j) = a_{ij}\cdot b_{ij} = c_{ij},$$

where we again used that the expectation of the product of the two independent random variables in the brackets is the product of their expectations.

We remark that $\mathbf{Z}$ is not multivariate Gaussian, and it is not necessary that $\mathbf{X}$ and $\mathbf{Y}$ be so, just because of their independence, the above calculations are valid (as any component of $\mathbf{X}$ is independent of any component of $\mathbf{Y}$, their covariances are also zeros). It should be noted that the above kind of matrix product is called Hadamard product or Schur product and denoted by $\boldsymbol{A}\circ\boldsymbol{B}$, whereas Proposition 1 is referred to as Schur's theorem, with a purely algebraic proof.

Consequently, if h is a polynomial with positive coefficients and $K:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ is a positive definite kernel, then the kernel $K_h:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ defined by

$$K_h(x,y) = h(K(x,y)) \tag{19}$$

is also positive definite. Since the exponential function can be approximated by polynomials with positive coefficients and the positive definiteness is closed under pointwise convergence, the same is true if h is the exponential function: $h(x)=e^x$, perhaps some transformation of it.

Putting these facts together and using the formula

$$\|x-y\|^2 = \langle x,x\rangle + \langle y,y\rangle - 2\langle x,y\rangle, \tag{20}$$

one can easily verify that the following, so-called *Gaussian kernel* is positive definite:

$$K_{\text{Gauss}}(x,y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}, \tag{21}$$

where $\sigma>0$ is a parameter. Indeed, in view of (20), this kernel can be written as the product of two positive definite kernels in the following way:

$$K_{\text{Gauss}}(x,y) = K_1(x,y)K_2(x,y),$$

where

$$K_1(x,y) = e^{-\frac{\langle x,x\rangle+\langle y,y\rangle}{2\sigma^2}},$$

and

$$K_2(x,y) = e^{\frac{\langle x,y\rangle}{\sigma^2}}.$$

Here K_2 is positive definite as it is the exponential function of the positive definite kernel $\frac{1}{\sigma^2}K_{\text{lin}}$. To show that K_1 is positive definite, by definition, we have to verify that for any $n\in\mathbb{N}$ and $x_1,\dots,x_n\in\mathcal{X}$, the symmetric matrix of entries

$$K_1(x_i,x_j) = e^{-\frac{\langle x_i,x_i\rangle}{2\sigma^2}}\cdot e^{-\frac{\langle x_j,x_j\rangle}{2\sigma^2}}, \quad i,j=1,\dots n$$

is positive semidefinite. But it is a rank 1 matrix (dyad), its only non-zero eigenvalue being equal to its trace, which is positive.

If $\mathcal{X}=\{x_1,\dots,x_n\}$ and $\boldsymbol{S}$ is an $n\times n$ symmetric similarity matrix comprised of the pairwise similarities between the entries of $\mathcal{X}$, then the kernel K defined by the $n \times n$ symmetric, positive definite matrix $e^{\lambda \boldsymbol{S}}$ is called *diffusion kernel*, where $0 < \lambda < 1$ is parameter (sometimes called decay factor). Let us recapitulate that the eigenvalues of the $e^{\lambda \boldsymbol{S}}$ matrix are the numbers $e^{\lambda \lambda_i}$ $(i = 1, \dots, n)$, where λ_i's are real eigenvalues of $\boldsymbol{S}$. Therefore the diffusion kernel is always strictly positive definite, even if $\boldsymbol{S}$ is not positive semidefinite. Above the aforementioned ones, there are a lot of other kernels, see [8, 52] for details.

5.2 RKHS corresponding to a kernel

Now we are able to formulate the converse statement. Recall that, by the Riesz–Fréchet representation theorem, an RKHS defines a positive definite kernel. In the other direction, for any positive definite kernel we can find a unique RKHS such that the relation formulated in (18) holds with appropriate elements of it. The following theorem is due to [9] who attributed it to E. H. Moore.

Theorem 5 *For any positive definite kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ there exists a unique, possibly infinite-dimensional Hilbert space $\mathcal{H}$ of functions on $\mathcal{X}$, for which K is a reproducing kernel.*

If we want to emphasize that the RKHS corresponds to the kernel K, we will denote it by $\mathcal{H}_K$. The proof of the Theorem 5 can be found in [8, 61], among others. It is based on the observation that the function space $\mathrm{Span}\{K_x = K(x, .) \,|\, x \in \mathcal{X}\}$ uniquely defines a pre-Hilbert space that can be completed into a Hilbert space. This will provide the unique RKHS $\mathcal{H}_K$.

However, we may wish to realize the elements of an RKHS $\mathcal{H}_K$ in a more straightforward Hilbert space $\mathcal{F}$. Assume that there is a (usually not linear) map $\phi : \mathcal{X} \to \mathcal{F}$ such that when $x \in \mathcal{X}$ is mapped into $\phi(x) \in \mathcal{F}$, then

$$K(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{F}}$$

is the desired positive definite kernel. At the same time, in view of (18),

$$K(x, y) = \langle K_x, K_y \rangle_{\mathcal{H}_K}$$

where recall that $K_x = K(x, .)$ is an $\mathcal{X} \to \mathbb{R}$ function, hence, cannot be identical to $\phi(x)$, but they can be connected with the following transformation. Let T be a linear operator from $\mathcal{F}$ to the space of functions $\mathcal{X} \to \mathbb{R}$ defined by

$$(Tf)(y) = \langle f, \phi(y) \rangle_{\mathcal{F}}, \quad y \in \mathcal{X},\ f \in \mathcal{F}.$$

Then

$$(T\phi(x))(y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{F}} = K(x, y) = K_x(y),$$

therefore

$$T\phi(x) = K_x, \quad \forall x \in \mathcal{X} \tag{22}$$

and hence, $\mathcal{H}_K$ becomes the range of T. This was the informal proof of the more precise statements about this correspondence.

5.3 Two examples of an RKHS

Here we give the theoretical construction for $\mathcal{H}_k$ and $\mathcal{F}$, together with the functions K_x and the features $\phi(x)$, in two special cases.

(a) Let K be the continuous kernel of a positive definite Hilbert–Schmidt operator which is an integral operator working on the $L^2(\mathcal{X})$ space, where $\mathcal{X}$ is a compact set in $\mathbb{R}$ for simplicity (it could be $\mathbb{R}^p$ as well). Now the positive definiteness of K means that

$$\int_{\mathcal{X}}\int_{\mathcal{X}} K(x,y)f(x)f(y)\,dx\,dy \geq 0, \quad \forall f \in L^2(\mathcal{X}),$$

and for the integral operator to be Hilbert–Schmidt, K must be in the $L^2(\mathcal{X} \times \mathcal{X})$ space, that is

$$\int_{\mathcal{X}}\int_{\mathcal{X}} K^2(x,y)\,dx\,dy < \infty$$

holds for it.

It is well known (see, e.g. [57]) that this operator is compact, hence has a discrete spectrum whose only possible point of accumulation is the 0. Because of the symmetry of K, the integral operator is also self-adjoint, and for it, the Hilbert–Schmidt theorem is applicable. This guarantees that the operator has nonnegative real eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ and corresponding eigenfunctions $\psi_1, \psi_2, \dots$. It also follows that $\sum_{i=1}^{\infty} \lambda_i^2 = \int_{\mathcal{X}}\int_{\mathcal{X}} K^2(x,y)\,dx\,dy < \infty$. By the Mercer theorem, if K is a continuous kernel of a positive definite integral operator on $L^2(\mathcal{X})$, where $\mathcal{X}$ is some compact space, then it can be expanded into the following uniformly convergent series:

$$K(x,y) = \sum_{i=1}^{\infty} \lambda_i \psi_i(x)\psi_i(y), \quad \forall x,y \in \mathcal{X}$$

by the eigenfunctions and the eigenvalues of the integral operator.

The RKHS defined by K is the following:

$$\mathcal{H}_K = \{f : \mathcal{X} \to \mathbb{R} \,:\, f(x) = \sum_{i=1}^{\infty} c_i\psi_i(x) \quad \texttt{s.t.} \quad \sum_{i=1}^{\infty} \frac{c_i^2}{\lambda_i} < \infty\}.$$

If $g(x) = \sum_{i=1}^{\infty} d_i\psi_i(x)$ – where $\sum_{i=1}^{\infty} \frac{d_i^2}{\lambda_i} < \infty$ – is another function in $\mathcal{H}_K$, then $f(x) + g(x)$ also corresponds to $\mathcal{H}_K$, due to $(c_i + d_i)^2 \leq 2(c_i^2 + d_i^2)$; the constant multiple of $f(x)$ also corresponds to $\mathcal{H}_K$, therefore $\mathcal{H}_K$ is a subspace of $L^2(\mathcal{X})$. The inner product of f and g is

$$\langle f, g\rangle_{\mathcal{H}_k} = \sum_{i=1}^{\infty} \frac{c_i d_i}{\lambda_i}.$$

Then one can easily verify that

$$K_x = K(x,.) = \sum_{i=1}^{\infty} \lambda_i\psi_i(x)\psi_i$$

is in $\mathcal{H}_K$. Indeed, it operates as

$$K_x(z) = \sum_{i=1}^{\infty} \lambda_i \psi_i(x)\psi_i(z)$$

and

$$\sum_{i=1}^{\infty} \frac{\lambda_i^2 \psi_i^2(x)}{\lambda_i} = K(x,x) < \infty.$$

Therefore,

$$\langle K_x, K_y \rangle_{\mathcal{H}_K} = \sum_{i=1}^{\infty} \frac{\lambda_i \psi_i(x) \lambda_i \psi_i(y)}{\lambda_i} = K(x,y). \tag{23}$$

Here the feature space $\mathcal{F}$ is the following counterpart of $\mathcal{H}_K$: it consists of infinite dimensional vectors

$$\phi(x) = (\sqrt{\lambda_1}\psi_1(x), \sqrt{\lambda_2}\psi_2(x), \dots), \quad x \in \mathcal{X}$$

and the inner product is naturally defined by

$$\langle \phi(x), \phi(y) \rangle_{\mathcal{F}} = \sum_{i=1}^{\infty} \lambda_i \psi_i(x)\psi_i(y),$$

which, in view of (23), is really equal to $\langle K_x, K_y \rangle_{\mathcal{H}_K}$.

In fact, here there are the functions $\sqrt{\lambda_1}\psi_1, \sqrt{\lambda_2}\psi_2, \dots$, which form an orthonormal basis in $\mathcal{H}_k$, and because of this transformation, a function $f \in L^2(\mathcal{X})$ is in $\mathcal{H}_k$ if $\|f\|^2_{\mathcal{H}_k} = \sum_{i=1}^{\infty} \frac{c_i^2}{\lambda_i} < \infty$. This condition restricts $\mathcal{H}_k$ to special functions between which the inner product is also adopted to the affine basis transformation. As the eigenfunctions of a Hilbert–Schmidt operator are continuous, an $f \in \mathcal{H}_K$ is also a continuous function. To further characterize the elements of $\mathcal{H}_K$, let us use the Banach–Steinhaus theorem.

$$\sup_{x \in \mathcal{X}} \|L_x(f)\| = \sup_{x \in \mathcal{X}} |f(x)| < \infty$$

holds for any $f \in \mathcal{H}_K$, as f is continuous on the compact set $\mathcal{X}$, by the Theorem of Weierstass. Under these circumstances, the Banach–Steinhaus uniform boundedness principle states that

$$\sup_{x \in \mathcal{X}} \|L_x\| < \infty,$$

that is, with some positive constant B,

$$\sup_{x \in \mathcal{X}} \sup_{\|f\|_{\mathcal{H}_K}=1} |f(x)| \le B < \infty.$$

Consequently, functions with fixed norm $\|f\|_{\mathcal{H}_K}$ are uniformly bounded, and the uniform bound is proportional to their $\mathcal{H}_K$-norm. Therefore, the global behavior of functions in $\mathcal{H}_K$ effects their local behavior, at least, it bounds the functions on their whole domain. Thus, they are – in certain sense – smooth functions. This property is due to the fact that these functions are strongly determined by the common kernel.

(b) Now $\mathcal{X}$ is a Hilbert space of finite dimension, say R^p, and its elements will be denoted by boldface $\mathbf{x}$, stressing that they are vectors. If we used K_{lin} on $\mathcal{X} \times \mathcal{X}$, then $K_{\mathbf{x}} = \langle \mathbf{x}, . \rangle_{\mathcal{X}}$, and by the Riesz–Fréchet representation theorem, $\phi(\mathbf{x}) = \mathbf{x}$ would reproduce the kernel, as $K_{\mathrm{lin}}(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{X}}$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$. Now the RKHS induced by K_{lin} is identified with the feature space, which is $\mathcal{X} = \mathbb{R}^p$ itself.

In case of more sophisticated kernels, $\mathcal{H}_K$ contains non-linear functions, and therefore, the features $\phi(\mathbf{x})$ can be realized usually in much higher dimension than that of $\mathcal{X}$. For example, let us consider the so-called polynomial kernel

$$K_{\mathrm{poly}}(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{X}} + c)^d, \quad c \geq 0,\, d \in \mathbb{N}, \quad (\mathbf{x}, \mathbf{y} \in \mathbb{R}^p)$$

obtained by using a special h in (19). It has $\binom{p+d}{d} \approx p^d$ eigenfunctions that span the space of p-variate polynomials with total degree d (this number is actually less if $c = 0$, i.e. the polynomials are of homogeneous degree).

In the following example of [52, 61], $p = 2$, $d = 2$, but instead of a $\binom{4}{2}$ =6-dimensional feature space, a 3-dimensional one will do, since the choice $c = 0$ is possible. Indeed, for $\mathbf{x} = (x_1, x_2) \in \mathcal{X} = \mathbb{R}^2$ let

$$\phi(\mathbf{x}) := (x_1^2, x_2^2, \sqrt{2} x_1 x_2),$$

hence, $\mathcal{F} \subset \mathbb{R}^3$. The idea comes from that we want to separate data points allocated along two concentric circles, and therefore $\mathbb{R}^2 \to \mathbb{R}$ quadratic functions are applied. The separating circle with equation $x_1^2 + x_2^2 = r^2$ (with a radius r between the radii of the two concentric circles) becomes a plane in the new coordinate system. The original $\mathbf{x}$'s in $\mathbb{R}^2$ were separated by an annulus, whereas projecting $\phi(\mathbf{x})$'s onto an appropriate two-dimensional plane of $\mathbb{R}^3$, a linear separation can be achieved. The two clusters can be separated by the k-means algorithm, as well.

Let us see, exactly what kernel is applied here. Using the usual inner product in R^3,

$$\langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_{\mathcal{F}} = x_1^2 y_1^2 + x_2^2 y_2^2 + 2 x_1 x_2 y_1 y_2 = (x_1 y_1 + x_2 y_2)^2 = \langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{X}}^2,$$

hence, the new kernel is the square of the linear one, which is also positive definite (polynomial kernel with $c = 0$, $d = 2$).

The RKHS $\mathcal{H}_K$ corresponding to the feature space $\mathcal{F}$ now consists of homogeneous degree quadratic functions $\mathbb{R}^2 \to \mathbb{R}$, with the functions $f_1 : (x_1, x_2) \to x_1^2$, $f_2 : (x_1, x_2) \to x_2^2$, and $f_3 : (x_1, x_2) \to \sqrt{2} x_1 x_2$ forming an orthonormal basis in $\mathcal{H}_k$ such that $K(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{3} f_i(\mathbf{x}) f_i(\mathbf{y})$. However, by the correspondence (22), the elements of $\mathcal{H}_K$ can be imagined as elements $\phi(\mathbf{x})$ in $\mathbb{R}^3$. Anyway, we do not need to navigate neither in the RKHS, nor in the feature space $\mathcal{F}$, but what we only need, is the new kernel:

$$K(\mathbf{x}, \mathbf{y}) = [K_{\mathrm{lin}}(\mathbf{x}, \mathbf{y})]^2, \quad \forall \mathbf{x}, \mathbf{y} \in \mathcal{X}.$$

In another setup, we may start with the above quadratic kernel K and build up the RKHS $\mathcal{H}_K$ as the span and completion of the following (homogeneous degree quadratic) $\mathcal{X} \to \mathbb{R}$ functions:

$$K_{\mathbf{x}} = \sum_{i=1}^{3} f_i(\mathbf{x}) f_i = x_1^2 f_1 + x_2^2 f_2 + \sqrt{2} x_1 x_2 f_3, \quad \mathbf{x} \in \mathcal{X}.$$

For example, $f_1 = K_{(1,0)}$, $f_2 = K_{(0,1)}$, and $f_3 = \frac{1}{2\sqrt{2}}(K_{(1,1)} - K_{(-1,1)})$. Then $\mathcal{F} = \{\phi(\mathbf{x}) \,|\, \mathbf{x} \in \mathbb{R}^2\} \subset \mathbb{R}^3$, and

$$\langle K_{\mathbf{x}}, K_{\mathbf{y}} \rangle_{\mathcal{H}_K} = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle_{\mathcal{F}} = K(\mathbf{x}, \mathbf{y}),$$

in accord with the theory, producing some kind of representation for special non-linear $\mathbb{R}^2 \to \mathbb{R}$ functions.

Observe that in both examples $\phi(x)$ is a vector with coordinates which are the basis vectors of the RKHS evaluated at x. In the first exercise $\phi(x)$ is an infinite, whereas in the second one, a finite dimensional vector. Note that $\mathcal{H}_K$ is an affine and sparsified version of the Hilbert space of $\mathcal{X} \to \mathbb{R}$ functions, between which the inner product is adopted to the requirement that it would reproduce the kernel.

5.4 Kernel – based on a sample – and the empirical feature map

In practical applications, we usually have a finite sample $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n)$. Based on it, the *empirical feature map* $\hat{\phi} : \mathcal{X} \to \mathbb{R}^n$ can be constructed in the following way (see e.g. [61]):

$$\hat{\phi}(\mathbf{x}) = \boldsymbol{K}^{-1/2} \phi_n(\mathbf{x}), \tag{24}$$

with $\phi_n(\mathbf{x}) = (K(\mathbf{x}, \mathbf{x}_1), \dots, K(\mathbf{x}, \mathbf{x}_n))^T$, the counterpart of $K(\mathbf{x}, .)$ based on the n-element set $\mathcal{X}$, and the $n \times n$ symmetric real matrix $\boldsymbol{K} = (K_{ij})$ of entries $K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$, $i, j = 1, \dots, n$. Assume that $\boldsymbol{K}$ is positive definite, otherwise (if positive semidefinite with at least one zero eigenvalue) we will use generalized inverse when calculating $\boldsymbol{K}^{-1/2}$. Let us apply (24) for $\mathbf{x}_i$'s. Since

$$\phi_n(\mathbf{x}_i) = \boldsymbol{K} \mathbf{e}_i,$$

where $\mathbf{e}_i$ is the ith unit vector in $\mathbb{R}^n$ (it has 0 coordinates, except the ith one which is equal to 1), the relation

$$\hat{\phi}(\mathbf{x}_i) = \boldsymbol{K}^{-1/2} \phi_n(\mathbf{x}_i) = \boldsymbol{K}^{1/2} \mathbf{e}_i$$

holds. Further,

$$\langle \hat{\phi}(\mathbf{x}_i), \hat{\phi}(\mathbf{x}_j) \rangle = (\boldsymbol{K}^{1/2} \mathbf{e}_i)^T (\boldsymbol{K}^{1/2} \mathbf{e}_j) = \mathbf{e}_i^T \boldsymbol{K} \mathbf{e}_j = K_{ij}, \quad i, j = 1, \dots, n.$$

Howsoever we cannot see well in the artificially constructed spaces, this whole abstraction was not in vain. Observe that for the data points $\mathbf{x}_i$'s we need not even calculate $\hat{\phi}(\mathbf{x}_i)$'s, the spectral clustering of these images can be done based on their pairwise distances:

$$\begin{aligned} \|\hat{\phi}(\mathbf{x}_i) - \hat{\phi}(\mathbf{x}_j)\|^2 &= \langle \hat{\phi}(\mathbf{x}_i), \hat{\phi}(\mathbf{x}_j) \rangle + \langle \hat{\phi}(\mathbf{x}_i), \hat{\phi}(\mathbf{x}_i) \rangle - 2 \langle \hat{\phi}(\mathbf{x}_j), \hat{\phi}(\mathbf{x}_j) \rangle \\ &= K(\mathbf{x}_i, \mathbf{x}_i) + K(\mathbf{x}_j, \mathbf{x}_j) - 2K(\mathbf{x}_i, \mathbf{x}_j) \end{aligned}$$

$(i, j = 1, \dots, n)$. Thus, to evaluate the pairwise distances between any pairs of the n features, merely the kernel values are needed. Sometimes the kernel is some transformation of a similarity matrix of n objects, even if we do not have them as finite-dimensional points. In other cases, we have finite dimensional measurements on the objects, but merely the $n \times n$ empirical covariance matrix is stored. If our data are multivariate Gaussian, this matrix suffices for further processing, in other cases, we can calculate a polynomial or Gaussian kernel based on it, with the explanation that it may be the true similarity between our non-Gaussian data which are, in fact, in an abstract space. For example, if $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle^2$, then the $n \times n$ similarity matrix of the features, applying a Gaussian kernel afterwards, gives matrix entries

$$K_{\text{Gauss}}(\hat{\phi}(\mathbf{x}_i), \hat{\phi}(\mathbf{x}_j)) = e^{-\frac{\|\hat{\phi}(\mathbf{x}_i) - \hat{\phi}(\mathbf{x}_j)\|^2}{2\sigma^2}} = e^{-\frac{\langle \mathbf{x}_i, \mathbf{x}_i \rangle^2 + \langle \mathbf{x}_j, \mathbf{x}_j \rangle^2 - 2\langle \mathbf{x}_i, \mathbf{x}_j \rangle^2}{2\sigma^2}}$$

which can be further processed through Laplacian based clustering, see Chapter **??**.

In this way, linear methods are applicable in an implicitly constructed space, instead of having to use non-linear methods in the original one. Here we only use the kernel which is calculated from the inner products of the data points through several transformations. The philosophy behind the above techniques is that sometimes sophisticated, composite kernels are more capable to reveal the structure of our data or to classify them, especially if they are not from a Gaussian distribution or consist of different type of measurements (e.g. location, brightness, color, texture). Just like in geometry, where the Euclidean distance is not necessarily the best choice, it is not always the linear kernel which is most useful in data representation.

But what kind of a kernel to be used? This is the important question. Many authors, e.g. [62, 63], recommend the Gaussian kernel. For the data points $\mathbf{x}_1, \dots, \mathbf{x}_n$ to be classified they construct the Gaussian kernel and the $n \times n$ symmetric, positive definite kernel matrix is considered as weight matrix $\boldsymbol{W}$ of a graph (in [49] the authors use zero diagonal). Then they perform spectral clustering based on the Laplacian or normalized Laplacian matrix corresponding to $\boldsymbol{W}$, see Lesson 1. In this way, applying the k-means algorithm for the so obtained k-dimensional (in fact, $(k-1)$-dimensional) representatives, they obtain nice clusters. This is because the data points of $\mathcal{X}$ are mapped into a feature space $\mathcal{F}$ such that the only implicitly known images $\phi(\mathbf{x}_i)$ $(i = 1, \dots, n)$ define a graph similarity, starting with which, the usual representation based spectral clustering works well. Hence, the graph construction is just an intermediate step for the subsequent metric clustering. Even if we are given a graph in advance, we may calculate the k-dimensional representatives of the vertices (with a relatively small k, based on the Laplacian eigenvectors), and then we classify them using kernel methods (e.g. substituting them into the Gaussian kernel).

The advantage of the Gaussian kernel is that it is also translation-invariant. The kernel $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is *translation-invariant* if

$$K(\mathbf{x}, \mathbf{y}) = k(\mathbf{x} - \mathbf{y})$$

with some $\mathbb{R}^p \to \mathbb{R}$ function k. In this case, the feature space has infinite dimension and the RKHS determined by a translation-invariant K is described by Fourier theory, see [34]. Since the Fourier transforms of functions in $\mathcal{H}_K$ decay rapidly, the induced RKHS consists of smooth functions. Now, let K be a p-dimensional Gaussian kernel with parameter $\sigma > 0$, defined in (21). Then

$$K(\mathbf{x}, \mathbf{y}) = k(\mathbf{x} - \mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}} \quad (\mathbf{x}, \mathbf{y} \in \mathbb{R}^p),$$

$K_{\mathbf{x}}$'s are the so-called radial basis functions, and the functions of the induced RKHS are convolutions of functions of $L^2(\mathbb{R}^p)$ with $e^{-\frac{\|\mathbf{x}\|^2}{\sigma^2}}$. The RKHS decreases from $L^2(\mathbb{R}^p)$ to $\emptyset$ as σ increases from 0 to ∞. In this way, Gaussian kernels may be used as smoothing functions, for example, in the ACE (Alternating Conditional Expectation) algorithm elaborated for the generalized non-parametric regression problem, see [24] for details.

If the underlying space $\mathcal{X}$ is a probability space of random variables with finite variance, and the product space is endowed with a joint distribution (see Section 4), we may look for the so-called $\mathcal{F}$-correlation of two random variables which is the largest possible correlation between their ϕ-maps in the feature space. In [13] it is proved that if $\mathcal{F}$ is the feature space corresponding to the RKHS defined by the Gaussian kernel on $\mathbb{R}$ (with any positive parameter σ), then the $\mathcal{F}$-correlation of two random variables is zero if and only if they are independent. Therefore, by mapping our data into the feature space, usual linear methods – like Principal Component Analysis or Canonical Correlation Analysis – become non-linear ones, and able to find independent components instead of uncorrelated ones, which fact has significance if our data come from a non-Gaussian distribution. This is the base of the so-called Kernel Independent Component Analysis, see [13, 60] for details. Note that with a finite-dimensional feature space, $\mathcal{F}$-correlation cannot characterize independence.

Finally, let us remark that because of the smoothness of the functions in an RKHS, not only the ϕ-maps of the data points $\mathbf{x}_i$'s are available, but also $\phi(\mathbf{x})$ for an $\mathbf{x}$ in the small neighborhood of a data point. There is the Nyström formula of similar flavor than (24) to do so, see [14, 15]. It can also be helped if the kernel is just some similarity function between pairs of data (not in a metric space), and though it is symmetric, not positive definite. In this case we can approximate it with a positive definite one. The technique applied is similar to that of the Multidimensional Scaling and other low rank approximations. In addition, we can better work with a low rank matrix; especially, if n is 'large' we need to find only some leading eigenvalues and eigenvectors of the $n \times n$ matrix $\boldsymbol{K}$. Using Gaussian kernel, the entries of $\boldsymbol{K}$ corresponding to pairs of data points that are far away, are negligibly 'small' and made zero, which fact gives rise to use algorithms developed for SD of sparse matrices, see e.g. [1].

Summarizing, RKHS techniques can be useful if we want to recover non-linear separation in our data. What we can do in general is that we calculate a linear kernel based on the sample and use admissible transformations to define newer and newer positive definite kernels, for example, a polynomial kernel of degree d if we guess that our data points can be separated with some curve of degree d. If only the relative position of the data points is of importance, we can build a Gaussian kernel based on their pairwise distances. The diffusion kernel is advisable to use in situations, when only a distance matrix of the objects is given and it is not Euclidean. In this case, either we use Multidimensional Scaling to embed the objects into a Euclidean space (and then use linear, polynomial, or Gaussian kernels), or else we select a diffusion kernel based on the not necessarily positive definite similarity matrix obtained from the given distance matrix (e.g. the entries are transformed by some monotonous decreasing function). Possibly, the similarities are correlations between the variables (if we want to classify the variables in a multidimensional dataset).

Even if we know the type of the kernel to be used, we must adapt its parameters to the data. With the new kernel and the pairwise distances, either Multidimensional Scaling or Laplacian based spectral clustering can be performed here.

We remark that for image segmentation purposes, [59] uses two or more Gaussian kernels: one contains the Euclidean distances of the pixels, and the others those of their brightness, color, texture, etc. Eventually, they take the product of the two or more positive definite kernels. If we multiply kernels, it means that entries in the same position of the kernel matrices are multiplied together. During the whole calculation for n data points, we only use the $n \times n$ symmetric, positive definite, usually sparse kernel matrix. In [65] the authors also recommend kernelization, but they do not specify the kernel to be used.

6 Application to image segmentation

In an image segmentation problem, we used the normalized modularity matrix and the eigenvectors corresponding to its $k-1$ largest eigenvalues to find k clusters of 2304 pixels based on a Gaussian kernel, see Section 5 (the distances of the pixels depended not only on their spacial distances). More precisely, we assigned the points $\mathbf{x}_1, \dots, \mathbf{x}_{2304} \in \mathbb{R}^3$ to the pixels, the coordinates of which characterize the spacial location, color, and brightness of the pixels. With the positive parameter σ, the similarity between pixels i and j was $w_{ij} = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}}$ for $i \neq j$. Figure 1 shows the original picture, and the picture when the pixels were colored according to their cluster memberships with number of clusters 3,4, and 5. Since the largest absolute value eigenvalues of the 2304×2304 normalized modularity matrix are

$$0.137259,\ 0.0142548,\ 0.000925228,\ -0.000670733,\ -0.000670674,\ \dots$$

and the number of the positive eigenvalues is three, with a gap after the second one, the 3- or 4-cluster solution seems the most reasonable. It is an intriguing question – unsolved so far – whether the dimension k of the original data points can be detected in the spectrum of $\boldsymbol{M}_D$ when n is large.

Note that instead of the usual Gaussian kernel we can use product kernel as follows. Divide the coordinates i of $\mathbf{x}_i$ into groups $G_1, \dots, G_m$ and denote by $\mathbf{x}_i^{(\ell)}$ the subvector of $\mathbf{x}_i$ with coordinates in G_ℓ, for $\ell = 1, \dots, m$ (e.g., according to the Fourier frequencies). Then

$$w_{ij} = \prod_{\ell=1}^{m} e^{-\frac{\|\mathbf{x}_i^{(\ell)} - \mathbf{x}_j^{(\ell)}\|^2}{2\sigma_\ell^2}}, \quad i \neq j,$$

where σ_ℓ can be 'smaller' if we want to enhance the importance of coordinates in group ℓ (e.g., at smaller Fourier frequencies).

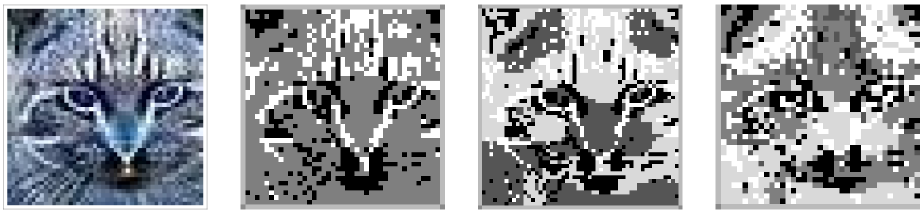

Figure 1: The original picture and the pixels colored with 3, 4, and 5 different colors according to their cluster memberships.

References

[1] Achlioptas D and McSherry F 2007 Fast computation of low-rank matrix approximations. *J. ACM* **54** (2), Article 9.

[2] Aldous DJ 1981 Reresentations for partially exchangeable arrays of random variables. *J. Multivariate Anal.* **11** (4), 581–598.

[3] Alon N 1986 Eigenvalues and expanders. *Combinatorica* **6** (2), 83–96.

[4] Alon N Milman VD 1985 λ_1, isoperimetric inequalities for graphs and superconcentrators. *J. Comb. Theory Ser. B* **38**, 73–88.

[5] Alon N Milman VD 1987 Better expanders and superconcentrators. *J. Algorithms* **8**, 337–347.

[6] Alpert CJ and Yao S.-Z 1995 Spectral partitioning: the more eigenvectors, the better. In *Proc. 32nd ACM/IEEE International Conference on Design Automation* (Preas BT, Karger PG, Nobandegani BS and Pedram M eds), pp. 195–200. Association of Computer Machinery, New York.

[7] Anderson WN and Morley TD 1985 Eigenvalues of the Laplacian of a graph. *Linear Multilinear Algebra* **18**, 141-145.

[8] Ando T 1987 *RKHS and quadratic inequalities.* Lecture Notes, Hokkaido University, Sapporo, Japan.

[9] Aronszajn N 1950 Theory of Reproducing Kernels. *Trans. Am. Math. Soc.* **68**, 337–404.

[10] Austin T 2008 On exchangeable random variables and the statistics of large graphs and hypergraphs. *Prob. Surv.* **5**, 80–145.

[11] Azran A and Ghahramani Z 2006 Spectral methods for automatic multiscale data clustering. In *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), New York NY* (Fitzgibbon A, Taylor CJ and Lecun Y eds), pp. 190–197. IEEE Computer Society, Los Alamitos, California.

[12] Babai L and Szegedy M 1992 Local expansion of symmetric graphs. *Comb. Probab. Comput.* **1**, 1–11.

[13] Bach FR and Jordan MI 2002 Kernel Independent Component Analysis. *J. Mach. Learn. Res.* **3**, 1–48.

[14] Baker C 1977 *The numerical treatment of integral equations.* Clarendon Press.

[15] Bengio Y, Delalleau O, Le Roux N, Paiement J-F, Vincent P and Ouimet M 2004 Learning eigenfunctions links spectral embedding and kernel PCA. *Neural Comput.* **16**, 2197–2219.

[16] Biggs NL 1974 *Algebraic Graph Theory.* Cambridge Univ. Press, Cambridge.

[17] Bolla M 1993 Spectra and Euclidean representation of hypergraphs. *Discret. Math.* **117**, 19–39.

[18] Bolla M and Tusnády G 1994 Spectra and Optimal Partitions of Weighted Graphs. *Discret. Math.* **128**, 1–20.

[19] Bolla M and M.-Sáska G 2004 Optimization problems for weighted graphs and related correlation estimates. *Discrete Mathematics* **282**, 23–33.

[20] Bolla M, Friedl K and Krámli A 2010 Singular value decomposition of large random matrices (for two-way classification of microarrays). *J. Multivariate Anal.* **101**, 434–446.

[21] Bolla M 2011 Penalized versions of the Newman–Girvan modularity and their relation to multi-way cuts and k-means clustering. *Phys. Rev. E* **84**, 016108.

[22] Bolla, M., Bullins, B., Chaturapruek, S., Chen, S., Friedl, K., Spectral properties of modularity matrices, *Linear Algebra and Its Applications* **73** (2015), 359-376.

[23] Borgs C Chayes JT Lovász L T.-Sós V and Vesztergombi K 2008 Convergent Sequences of Dense Graphs I: Subgraph Frequences, Metric Properties and Testing. *Advances in Math.* **219**, 1801–1851.

[24] Breiman L and Friedman JH 1985 Estimating optimal transformations for multiple regression and correlation. *J. Am. Stat. Assoc.* **80**, 580–619.

[25] Chung F 1997 *Spectral Graph Theory*, CBMS Regional Conference Series in Mathematics **92**. American Mathematical Society, Providence RI.

[26] Cvetković DM, Doob M and Sachs H 1979 *Spectra of Graphs.* Academic Press, New York.

[27] Cvetković DM, Rowlinson P and Simić S 1997 *Eigenspaces of Graphs*, Encyclopedia of Mathematics and Its Applications Vol. 66. Lavoisier Libraire, S.A.S. France.

[28] Dhillon IS 2001 Co-clustering documents and words using bipartite spectral graph partitioning. In *Proc. ACM Int. Conf. Knowledge Disc. Data Mining (KDD 2001)* (Provost FJ and Srikant R eds), pp. 269-274. Association for Computer Machinery, New York.

[29] Diaconis P and Janson S 2008 Graph limits and exchangeable random graphs. *Rend. Mat. Appl.* (VII. Ser.) **28**, 33–61.

[30] Fiedler M 1972 Bounds for eigenvalues of doubly stochastic matrices. *Linear Algebra Appl.* **5** (98), 299–310.

[31] Fiedler M 1973 Algebraic connectivity of graphs. *Czech. Math. J.* **23** (98), 298–305.

[32] Frieze A, Kannan R and Vempala S 1998 Fast Monte-Carlo Algorithms for finding low-rank approximations. In *Proc. 39th Annual IEEE Symposium on Foundations of Computer Science (FOCS 1998), Palo Alto, California*, pp. 370–386. IEEE Computer Society, Los Alamitos.

[33] Gebelein H 1941 Das statistische Problem der Korrelation als Variations und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung. *Z. Angew. Math. Mech.* **21**, 364–379.

[34] Girosi F, Jones M and Poggio T 1995 Regularization theory and neural networks architectures. *Neural Comput.* **7**, 219–269.

[35] Hagen L and Kahng AB 1992 New spectral methods for ratio cut partitioning and clustering. *IEEE Trans. Comput. Aided Des.* **11**, 1074–1085.

[36] Hoffman AL 1969 The change in the least eigenvalue of the adjacency matrix of a graph under imbedding. *SIAM J. Appl. Math.* **17**, 664–671.

[37] Juhász F and Mályusz K 1980 Problems of cluster analysis from the viewpoint of numerical analysis. In *Numerical Methods, Coll. Math. Soc. J. Bolyai* (ed. Rózsa P), Vol 22, pp. 405–415. North-Holland, Amsterdam.

[38] Kelmans AK 1967 Properties of the charasteristic polynomial of a graph. *Cibernetics in the Science of Communication* **4**, Energija, Moskva–Leningrad, 27–41 (in Russian).

[39] Kleinberg J 1997 Authoritative sources in hyperlinked environment. IBM Research Report RJ 10076 (91892).

[40] Lovász L 1993 *Combinatorial Problems and Exercises.* Akadémiai Kiadó–North Holland, Budapest–Amsterdam.

[41] Liotta G ed. 2004 *Graph drawing.* Lecture Notes in Computer Science **2912**. Springer.

[42] Maas C 1987 Transportation in graphs and the admittance spectrum. *Discret. Appl. Math.* **16**, 31–49.

[43] Merris R 1987 Characteristic vertices of trees. *Linear Multilinear Algebra* **22**, 115–131.

[44] Mohar B 1988 Isoperimetric inequalities, growth and the spectrum of graphs. *Linear Algebra Appl.* **103**, 119–131.

[45] Mohar B 1991 The Laplacian spectrum of graphs. In *Graph Theory, Combinatorics, and Applications* Vol 2 (Alavi Y, Chartrand G, Oellermann OR and Schwenk AJ eds), pp. 871–898. Wiley.

[46] Newman MEJ and Girvan M 2004 Finding and evaluating community structure in networks. *Phys. Rev. E* **69**, 026113.

[47] Newman MEJ 2004 Detecting community structure in networks. *Eur. Phys. J. B* **38**, 321–330.

[48] Newman MEJ 2010 *Networks, An Introduction.* Oxford University Press.

[49] Ng AY Jordan MI and Weiss Y 2001 On spectral clustering: analysis and an algorithm. In *Proceedings of the 14th Neural Information Processing Systems Conference, NIPS 2001* (Dietterich TG, Becker S and Ghahramani Z eds), pp. 849–856. MIT Press, Cambridge, USA.

[50] Parzen E 1963 Probability density functionals and reproducing kernel Hilbert spaces. In *Proceedings of the Symposium on Time Series Analysis* (ed. Rosenblatt M), pp. 155–169. Brown University, Providence, RI, USA.

[51] Pisanski T and Shawe-Taylor J 2000 Characterizing graph drawing with eigenvectors. *J. Chem. Inf. Comput. Sci.* **40**, 567–571.

[52] Reisen K and Bunke H 2010 *Graph Classification and Clustering Based on Vector Space Embedding.* World Scientific.

[53] Rényi A 1959 On measures of dependence. *Acta Math. Acad. Sci. Hung.* **10**, 441–451.

[54] Rényi A 1959 New version of the probabilistic generalization of the large sieve. *Acta Math. Acad. Sci. Hung.* **10**, 218–226.

[55] Riesz F 1907 Sur une espéce de géométrie analytique des systémes de fonctions sommables. *C. R. Acad. Sci. Paris* **144**, 1409–1411.

[56] Riesz F 1909 Sur les opérations fonctionnelles linéaires. *C. R. Acad. Sci. Paris* **149**, 974–977.

[57] Riesz F and Sz.-Nagy B 1952 *Leçons d'analyse fonctionnelle.* Academic Publishing House, Budapest.

[58] Shawe-Taylor J and Cristianini N 2004 *Kernel Methods for Pattern Analysis.* Cambridge Univ. Press, Cambridge.

[59] Shi J and Malik J 2000 Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **22** (8), 888–905.

[60] Schölkopf B Smola AJ and Müller K-R 1998 Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.* **10**, 1299–1319.

[61] Schölkopf B and Smola AJ 2002 *Learning with Kernels.* MIT Press, Cambridge, USA.

[62] Von Luxburg U 2006 A tutorial on spectral clustering. *Stat. Comput.* **17**, 395–416.

[63] Von Luxburg U Belkin M and Bousquet O 2008 Consistency of spectral clustering. *Ann. Stat.* **36**, 555–586.

[64] White S and Smyth P 2005 A spectral clustering approach to find communities in graphs. In *Proc. SIAM International Conference on Data Mining* (Kargupta H, Srivastava J, Kamath Ch and Goodman A), pp. 76–84. SIAM, Newport Beach.

[65] Yan S, Xu D, Zhang B, Zhang H-J, Yang Q and Lin S 2007 Graph embedding and extensions: a general framework for dimensionality reduction. *IEEE Trans. Pattern Anal. Mach. Intell.* **29** (1), 40–48.