Chapter 3
Graph based clustering

The general idea of clustering is to partition a set of objects $\mathcal{O}$ into disjoint subsets of objects of "similar" nature. In the case of $k-$means the elements of $\mathcal{O}$ are characterized as points in a metric space and the distance between the points and the centroids is used as "dissimilarity" or distortion measure to group the objects. It is clear that in $k-$means it was implicitly assumed that the objects could in fact be described as points in a metric space, and that it was possible to use the mathematical notion of distance. This is one possible kind of clustering problem but it is not the only one, it may be the case that the relations among objects are encoded in non-metric properties that could in principle be negative valued, non-symmetric or violate the triangle inequality thus precluding the possibility of using metric space properties. For instance correlations between different financial institutions in Europe Fig. 3.1 and genomic data Fig. 3.2

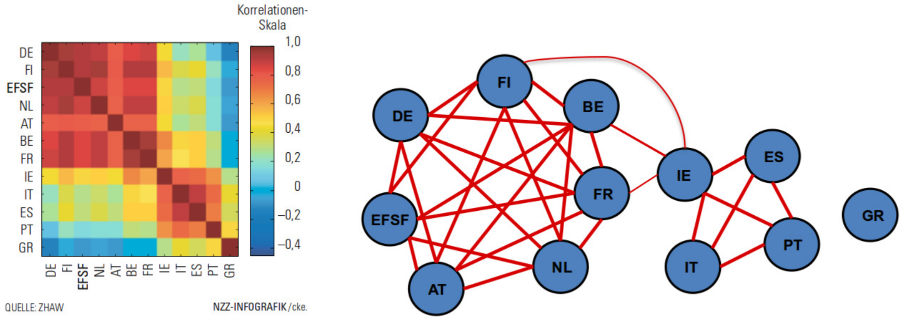

Figure 3.1 (Left) Yield change correlation of 10-years government issued bonds for different European nations. (Right) The correlations suggest two weakly connected cliques and a singleton. (Source: NZZ 21 Mar 2015, P37)

In this section the idea of graph based clustering is introduced and a number of cost functions are considered, with particular emphasis on the pairwise data clustering cost function $\mathcal{R}^{pc}$ its invariance properties and relation to $k-$means clustering.

One mathematical structure that lends itself to the clustering problem when the relations among objects are not necessarily of metric nature is that of a graph. More precisely, a graph is a pair $(\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ is the set of vertices or objects and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$, where each pair denotes an edge or link between the objects. If the pairs

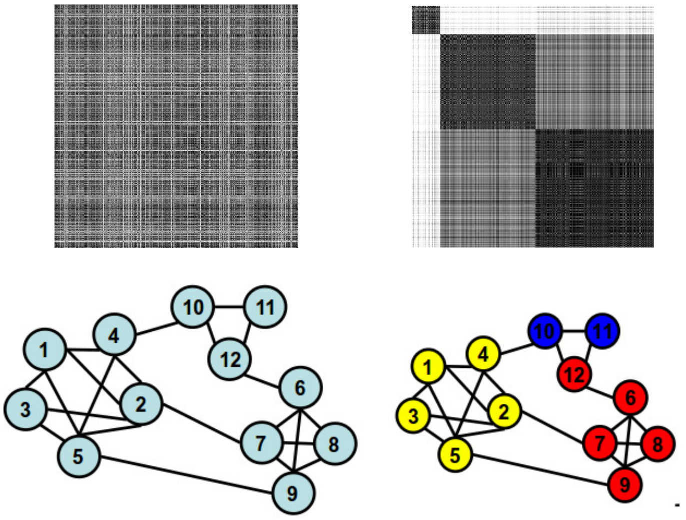

Figure 3.2 (Left) Raw data of gene correlations, and a scheme of the associated graph. (Right) Result after permuting the labels according to a clustering procedure on the vertices and scheme of the “clustered” vertices.

are assumed unordered, then the graph is said to be undirected. Weighted graphs are defined as $(\mathcal{V}, \mathcal{E}, w)$ where $w : \mathcal{E} \to \mathbb{R}$ is a function that assigns a real value to edges.

For a general clustering setting, let $\mathcal{O}$ be the set of vertices $\mathcal{V}$, the set of edges is inferred or given for each problem, and the set of (di)similarity measures between objects $\mathcal{D} = \{D_{ij}\}$ or $\mathcal{S} = \{S_{ij}\}$ be the weights. Clusters are defined as

$$\mathcal{G}_\alpha = \{\mathbf{o} \in \mathcal{O} : c(\mathbf{O}) = \alpha\}, \tag{3.1}$$

where $c(\mathbf{o})$ is the cluster assignment of object $\mathbf{o}$.

3.1 Correlation Clustering

As a first example of graph clustering, assume that the proximity or similarity measure between objects i, j is of the form $S_{ij} = \{-1, 1\}$ and that the graphs should be partitioned in such a way that the agreement within a cluster and disagreement between clusters is maximized. A possible cost function for this task, first proposed for documental analysis **?**, is given by

$$\mathcal{R}^{cc}(c, \mathcal{D}) = -\sum_{\nu \le k} \sum_{(i,j)\in\mathcal{E}_{\nu\nu}} S_{ij} + \sum_{\nu \neq \mu \le k} \sum_{(i,j)\in\mathcal{E}_{\mu\nu}} S_{ij}, \tag{3.2}$$

where k denotes the number of clusters and

$$\mathcal{E}_{\alpha\beta} = \{(i,j) \in \mathcal{E} : \mathbf{o}_i \in \mathcal{G}_\alpha \wedge \mathbf{o}_j \in \mathcal{G}_\beta\}, \tag{3.3}$$

is the set of intercluster edges between clusters α and β. Observe that the first term in Eq.(3.2) favors that objects within the same cluster are in agreement whereas the second term favors that objects from different clusters disagree. It is not immediately evident but $\mathcal{R}^{cc}$ assigns more weight (is biased to) large clusters sizes. This can be seen by considering the following

$$\begin{aligned}\mathcal{R}^{cc}(c,\mathcal{D}) &= -\sum_{\nu\leq k}\sum_{(i,j)\in\mathcal{E}_{\nu\nu}} S_{ij} + \sum_{\nu\neq\mu\leq k}\sum_{(i,j)\in\mathcal{E}_{\mu\nu}} S_{ij}, \\ &= -2\sum_{\nu\leq k}\sum_{(i,j)\in\mathcal{E}_{\nu\nu}} S_{ij} + \sum_{\nu,\mu\leq k}\sum_{(i,j)\in\mathcal{E}_{\mu\nu}} S_{ij}, \\ &= -2\sum_{\nu\leq k}\sum_{(i,j)\in\mathcal{E}_{\nu\nu}} S_{ij} + \sum_{(i,j)\in\mathcal{E}} S_{ij}. \qquad (3.4)\end{aligned}$$

Notice that the second term of Eq. (3.4) is independent of the assignment c as it is a sum over all edges. Since Gibbs distribution is invariant under constant shifts of the cost function, the first term of Eq. (3.4) is the meaningful one. Because $-2\sum_{\nu\leq k}\sum_{(i,j)\in\mathcal{E}_{\nu\nu}} S_{ij}$ is a sum only over intracluster edges the larger the cluster the higher its weight, which is expected to scale as the size of the cluster squared $|\mathcal{G}|^2$. Therefore it tends to be less expensive to add a new nodes to the bigger clusters than to smaller ones.

A similar cost function to Eq. (3.2) is

$$\mathcal{R}^{gp}(c,\mathcal{D}) = \sum_{\nu\leq k}\sum_{(i,j)\in\mathcal{E}_{\nu\nu}} D_{ij} \qquad (3.5)$$

where the dissimilarities $D_{ij}\in\mathbb{R}$, this cost function has the same kind of deficiencies of Eq. (3.2) and in general is known to be bias towards very unbalanced cluster sizes i.e very large or very small.

3.2 Pairwise data clustering

A natural principle for exploratory data analysis is that of minimizing the pairwise distances between the members within each cluster. A cost function that is constructed to follow this idea is the pairwise clustering cost function

$$\mathcal{R}^{pc}(c,\mathcal{D}) = \frac{1}{2}\sum_{\nu\leq k}\left(|\mathcal{G}_\nu|\sum_{(i,j)\in\mathcal{E}_{\nu\nu}}\frac{D_{ij}}{|\mathcal{E}_{\nu\nu}|}\right), \qquad (3.6)$$

where $\mathcal{D}_{ij}$ is the dissimilarity matrix and is only assumed to have zero self-dissimilarity entries, $|\mathcal{G}_\nu|$ and $|\mathcal{E}_{\nu\nu}|$ are respectively the cardinalities of the cluster $\mathcal{G}_\nu$ and its set of edges $\mathcal{E}_{\nu\nu}$. As it will be shown below, this cost function is invariant under symmetrization of the dissimilarity matrix and also under constant additive shifts of the non diagonal elements. These two properties will be used to show that it is possible to embed a given nonmetric proximity data problem in a vector space without changing the underlying properties of the data using the "constant shift embedding" **?** procedure. To reach this result first it will be shown that the $k-$means cost function Eq. (2.49) can be written in the form of $\mathcal{R}^{pc}$, then the invariance properties of $\mathcal{R}^{pc}$ will be derived and used to establish the "constant shift embedding" procedure that is the main result of this section.

3.3 Pairwise clustering as k-means clustering in kernel space

In order to show that $\mathcal{R}^{km}$ can be seen as a special case of $\mathcal{R}^{pc}$ it is convenient to rewrite Eq. (3.6) as

$$\mathcal{R}^{pc}(c,\mathcal{D}) = \frac{1}{2}\sum_{\nu\leq k}\frac{\sum_{i\leq n}\sum_{j\leq n} M_{i\nu}M_{j\nu}D_{ij}}{\sum_{l\leq n} M_{l\nu}}, \qquad (3.7)$$

where $M_{i\nu}$ is a $n \times k$ matrix with entries 0 or 1 depending on whether or not the object i is assigned to cluster ν. Notice that $\sum_{\nu \leq k} M_{i\nu} = 1$ meaning that a given object i can only be assigned to a single cluster and $\sum_{i \leq n} M_{i\nu} = |\mathcal{G}_\nu| = n_\nu$. Now Eq. (2.49) can be cast in the following form:

$$\mathcal{R}^{km} = \sum_{i \leq n} ||x_i - y_{c(i)}||^2 = \sum_{\nu \leq\ k} \sum_{i:c(i)=\nu} ||x_i - y_\nu||^2 = \sum_{\nu \leq\ k} \sum_{i \leq n} M_{i\nu} ||x_i - y_\nu||^2, \quad (3.8)$$

using the definition of y_ν Eq. (2.51) and the identity $x_i \cdot x_j = \frac{1}{2}x_i^2 + \frac{1}{2}x_j^2 - \frac{1}{2}||x_i - x_j||^2$ it can be shown that

$$||x_i - y_\nu||^2 = \frac{1}{n_\nu} \sum_{j \leq n} M_{j\nu} ||x_i - x_j||^2 - \frac{1}{2n_\nu^2} \sum_{j \leq n} \sum_{l \leq n} M_{j\nu} M_{l\nu} ||x_j - x_l||^2. \quad (3.9)$$

Finally observe that if Eq.(3.9) is multiplied by $M_{i\nu}$ and then summed over all objects the result is

$$\sum_{i \leq n} M_{i\nu} ||x_i - y_\nu||^2 = \frac{1}{2n_\nu} \sum_{j \leq n} \sum_{l \leq n} M_{j\nu} M_{l\nu} ||x_j - x_l||^2. \quad (3.10)$$

Substitution of Eq. (3.10) in Eq. (3.8) yields

$$\mathcal{R}^{km} = \frac{1}{2} \sum_{\nu \leq k} \sum_{j \leq n} \sum_{l \leq n} \frac{M_{j\nu} M_{l\nu} ||x_j - x_l||^2}{\sum_{i \leq n} M_{i\nu}}. \quad (3.11)$$

It is clear that if the identification $\mathcal{D}_{ij} = ||x_i - x_j||^2$ is done, then $\mathcal{R}^{km}$ can be reduced to $\mathcal{R}^{pc}$. The previous transformation can always be carried out but the reverse not necessarily, this is because the dissimilarity matrix may have negative values or sets of values that violate the triangle inequality making it impossible to interpret the data as distances in a metric space.

The invariance property of $\mathcal{R}^{pc}$ under the simmetrization of the dissimilarity matrix means that given D, the problems $\mathcal{R}^{pc}(D)$ and $\mathcal{R}^{pc}(\tilde{D})$ where $\tilde{D} = \frac{1}{2}(D + D^T)$ are the same. To see this, just notice that the expression $\sum_{i \leq n} \sum_{j \leq n} M_{i\nu} M_{j\nu} D_{ij}$ in the numerator of Eq. (3.7) always has both terms D_{ij} and D_{ji}, thus if $\tilde{D}$ is used instead of D the result is the same due to the factor 1/2 in the definition of $\tilde{D}$.

$\mathcal{R}^{pc}$ is invariant (up to a constant) under additive shifts of the non diagonal elements of D i.e. $D_{ij} \to \tilde{D}_{ij} = D_{ij} + d(1 - \delta_{ij})$, this can be seen as follows

$$\begin{aligned}
\mathcal{R}^{pc}(\tilde{D}) &= \frac{1}{2} \sum_{\nu \leq k} \frac{\sum_{i \leq n} \sum_{j \leq n} M_{i\nu} M_{j\nu} \{D_{ij} + d(1 - \delta_{ij})\}}{\sum_{l \leq n} M_{l\nu}} \\
&= \mathcal{R}^{pc}(D) + \frac{1}{2} \sum_{\nu \leq k} \frac{\sum_{i \leq n} \sum_{j \leq n} M_{i\nu} M_{j\nu} d(1 - \delta_{ij})}{\sum_{l \leq n} M_{l\nu}} \\
&= \mathcal{R}^{pc}(D) + \frac{d}{2} \sum_{\nu \leq k} (n_\nu - 1) \\
&= \mathcal{R}^{pc}(D) + \frac{d}{2}(n - k) \quad (3.12)
\end{aligned}$$

Since the minimization problem is insensitive to overall additive constants $\mathcal{R}^{pc}$ and $\mathcal{R}^{pc}(\tilde{D})$ are indeed equivalent. The last idea required to build the constant shift embedding is that of the centralized version of a matrix. Let $U_{ij}(n) = 1$ be a $n \times n$ matrix with all its entries equal to 1, I_n the $(n \times n)$ identity matrix and $Q = I_n - \frac{1}{n}U(n)$. The centralized version of a matrix P is defined as $P^c = QPQ$, in components it reads:

$$P^c_{ij} = P_{ij} - \frac{1}{n}\sum_{k\leq n} P_{ik} - \frac{1}{n}\sum_{k\leq n} P_{kj} + \frac{1}{n^2}\sum_{k,l\leq n} P_{kl}. \tag{3.13}$$

Given the symmetrization invariance of $\mathcal{R}^{pc}$ there is no loss of generality if the dissimilarity matrix is assumed symmetric from now on. Let S be a symmetric matrix such that

$$D_{ij} = S_{ii} + S_{jj} - 2S_{ij}. \tag{3.14}$$

By construction D is symmetric and its diagonal elements are all zero, but S is not unique because there are $(n^2 - n)/2$ equations and $(n^2 + n)/2$ variables. Even though S is not unique, it is the case that all S have the same centralized version S^c, and that S^c is also a valid decomposition. These statements are proven in Lemma 3. and Lemma 4.

Lemma 3 ($S^c = -\frac{1}{2}D^c$ for any S that decompose D as in Eq.(3.14))**.** *Let S be a symmetric matrix that satisfies the decomposition Eq.(3.14) for a given dissimilarity matrix D. Using both Eq.(3.13) and Eq. (3.14), the centralized version of S is calculated as follows*

$$\begin{aligned}
S^c_{ij} = -\frac{1}{2}\Bigg[&(D_{ij} - S_{ii} - S_{jj}) - \frac{1}{n}\sum_{k\leq n}(D_{ik} - S_{ii} - S_{kk}) \\
&- \frac{1}{n}\sum_{k\leq n}(D_{kj} - S_{kk} - S_{jj}) + \frac{1}{n^2}\sum_{k,l\leq n}(D_{kl} - S_{kk} - S_{ll})\Bigg] \\
= -\frac{1}{2}\Bigg[&D_{ij} - \frac{1}{n}\sum_{k\leq n} D_{ik} - \frac{1}{n}\sum_{k\leq n} D_{kj} + \frac{1}{n^2}\sum_{k,l\leq 1} D_{kl}\Bigg] = -\frac{1}{2}D^c_{ij}.
\end{aligned} \tag{3.15}$$

Lemma 4 (If S satisfies Eq.(3.14) so does S^c)**.** *Assume that S is a matrix that satisfies Eq.(3.14). Direct substitution of S^c in the right hand of Eq.(3.14) yields*

$$\begin{aligned}
S^c_{ii} + S^c_{jj} - 2S^c_{ij} &= S_{ii} + S_{jj} - 2S_{ij} \\
&- \frac{1}{n}\sum_{k\leq n} S_{ik} - \frac{1}{n}\sum_{k\leq n} S_{ki} + \frac{1}{n^2}\sum_{k,l\leq n} S_{kl} \\
&- \frac{1}{n}\sum_{k\leq n} S_{jk} - \frac{1}{n}\sum_{k\leq n} S_{kj} + \frac{1}{n^2}\sum_{k,l\leq n} S_{kl} \\
&+ 2\left(+\frac{1}{n}\sum_{k\leq n} S_{ik} + \frac{1}{n}\sum_{k\leq n} S_{kj} - \frac{1}{n^2}\sum_{k,l\leq n} S_{kl}.\right) \\
&= D_{ij},
\end{aligned} \tag{3.16}$$

where the summation terms (middle three lines) cancel completely due to the symmetry of S.

The importance of Lemma 3 strives in the result that if S^c is positive semidefinite, then the elements of D are given by squared euclidean distances $D_{ij} = ||x_i - x_j||^2$, see **??**. In general S^c is not positive semidefinite, nevertheless it is possible to construct a matrix $\tilde{S}$ that is positive semidefinite and that does not change the original problem. Notice that the $(n \times n)$ matrix $\tilde{A} = A - \lambda(A)_{min} I_n$, where $\lambda(A)_{min}$ denotes the smallest eigenvalue of A, is positive definite (consider what happens when $\tilde{A}$ acts on the eigenvectors of A). Using this property construct the $\tilde{S}$ matrix, in components $\tilde{S}_{ij} = S^c_{ij} - \lambda(S^c)_{min}\delta_{ij}$ and use Eq. (3.14) to obtain the matrix $\tilde{D}$ given by

$$\begin{aligned} \tilde{D}_{ij} &= \tilde{S}_{ii} + \tilde{S}_{jj} - 2\tilde{S}_{ij} \\ &= S^c_{ii} + S^c_{jj} - 2S^c_{ij} - 2\lambda(S^c)_{min}(1 - \delta_{ij}) \\ &= D_{ij} - 2\lambda(S^c)_{min}(1 - \delta_{ij}). \end{aligned} \tag{3.17}$$

Observe that $\tilde{D}$ and D are related by an off diagonal constant shift thus they correspond to the same problem, furthermore since $\tilde{S}$ is positive semidefinite, $\tilde{D}_{ij} = ||x_i - x_j||^2$, the actual vectors can be obtained using the Kernel PCA method by **?** and the equivalence between $\mathcal{R}^{km}$ and $\mathcal{R}^{pc}$ can be used. Fig. 3.3 summarizes the steps that have been made up to this point.

$$\begin{tikzcd}
D \arrow[r, "\text{Decomposition}"] & D_{ij} = S_{ii}+S_{jj}-2S_{ij} \arrow[r, "S \to S^c = -\frac{1}{2}D^c", "\tilde{S}_{ij}=S^c_{ij}-\lambda(S^c)_{min}\delta_{ij}"'] & \tilde{D}_{ij} = D_{ij}-2\lambda(S^c)_{min}(1-\delta_{ij}) \arrow[d, "\tilde{D}_{ij}=||x_i-x_j||^2"] \\
\mathcal{R}^{pc}(D) \arrow[r, "\mathcal{R}^{pc}(D)=\mathcal{R}^{pc}\left(\frac{1}{2}(D+D^T)\right)", "\mathcal{R}^{pc}(D)=\mathcal{R}^{pc}(D+d(I_n-U(n)))"'] & \mathcal{R}^{pc}(D) = \mathcal{R}^{pc}(\tilde{D}) \arrow[r] & \underbrace{\mathcal{R}^{pc}(\tilde{D}) \sim \mathcal{R}^{km}(\tilde{D})}_{\text{Identical assigments}}
\end{tikzcd}$$

Figure 3.3 Scheme for embedding a pairwise clustering problem in a vector space, and its relation with $k-$means clustering

At this stage some comments are in order. First of all, the fact that the original pairwise clustering problem can be embedded in a vector space implies not only its equivalence to a $k-$means problem, but also that the ideas of centroids and cluster representatives can be used, for instance, for prediction rules. Second, in general it is hard to denoise pairwise data while for vectorial data this is possible, thus the constant shift embedding procedure allows to use standard techniques for preprocessing and denoising vectorial data on pairwise data. Third, the minimization processes of either the pairwise cost function or the $k-$means problems, is $\mathcal{NP}-$hard and algorithms such as deterministic annealing are required. In order to use deterministic annealing, the mean field approximation is needed, and it is the case that for $\mathcal{R}^{km}$ this is an exact procedure whereas for $\mathcal{R}^{pc}$ it remains an approximation. Finally, the reader is directed to **?** for example applications of the "constant shift embedding" procedure, as well for its relation with other pairwise cost functions such as Average association and Average cut.

Chapter 4
Mean field approximation

The explicit calculation of the Gibbs distribution given a general cost function $\mathcal{R}$ is not always a simple task as in the case of $k-$means. The problem is usually the evaluation of the partition function that involves exponentially many terms and is analytically intractable. This often happens when the cost functions is not linear in the individual costs, or includes non trivial structures terms that make that the cost are not independent from one other. For example a local smoothness constrain term such as $\sum_{i\leq n}\sum_{j\in\mathcal{N}(i)}\mathbb{I}_{c(i)\neq c(j)}$ where $\mathcal{N}(j)$ is the neighborhood of object j.

In order to proceed, a criteria to assess the "goodness" of an approximation $\mathbf{Q}$ to a given distribution $\mathbf{P}$ is required as well as the fact that the approximation $\mathbf{Q}$ can be efficiently calculated compared to the original distribution. The program to build $\mathbf{Q}$ is then to assume that it is of a factorial form and then minimize the Kullback-Leibler divergence or relative entropy (see below) between $\mathbf{P}$ and $\mathbf{Q}$ to obtain the best factorial approximation $\mathbf{Q}$ of $\mathbf{P}$. To carry out this program this section is organized as follows: first the Kullback-Leibler divergence is defined and its positivity is demonstrated, afterwards the factorial approximation $\mathbf{Q}$ is introduced and then the minimization of the relative entropy between $\mathbf{Q}$ and $\mathbf{P}$ is explicitly carried out.

Definition 3 (Kullback-Leibler divergence or Relative entropy)**.** *Let $\mathbf{P}(x)$ and $\mathbf{Q}(x)$ be two probability distributions defined over the same sample space Ω. The Kullback-Leibler divergence or Relative entropy is given by*

$$D(\mathbf{Q}||\mathbf{P}) = \sum_{x\in\Omega} p(x)\log\frac{p(x)}{q(x)}, \tag{4.1}$$

in the discrete case and

$$D(\mathbf{Q}||\mathbf{P}) = \int_{\Omega} p(x)\log\frac{p(x)}{q(x)}dx, \tag{4.2}$$

in the continuous case.

Observe that $D(\mathbf{Q}||\mathbf{P})$ is identically zero if $\mathbf{P}=\mathbf{Q}$[1], notice also that the relative entropy is not symmetric in its arguments and that it is positive semidefinite as show in theorem 3. An intuitive but important interpretation

of the Kullback-Leibler divergence is that it quantifies the coding cost of describing data with a probability distribution $\mathbf{Q}$ when the true distribution is $\mathbf{P}$?. Therefore its use as a comparing criteria between distributions.

[1] In the continuous case this holds if both distributions differ only over a set of zero measure

Theorem 3 ($D(\mathbf{P}||\mathbf{Q})$ is positive semidefinite)**.** *Let $\mathbf{P}(x)$ and $\mathbf{Q}(x)$ be two probability distributions defined over the same sample space Ω, then*

$$\begin{aligned} -D(\mathbf{P}||\mathbf{Q}) &= -\sum_{x\in\Omega} p(x)\log\frac{p(x)}{q(x)} && (4.3)\\ &= \sum_{x\in\Omega} p(x)\log\frac{q(x)}{p(x)} && (4.4)\\ &\leq \sum_{x\in\Omega} p(x)\left(\frac{q(x)}{p(x)}-1\right) && (4.5)\\ &= \sum_{x\in\Omega}(q(x)-p(x)) = 0. && (4.6)\end{aligned}$$

Let $\mathbf{Q}$ be a factorial approximation to $\mathbf{P}$ given by

$$\mathbf{Q}(c,\theta) = \prod_{i\leq n} q_i(c(i)), \tag{4.7}$$

where θ is a set of free parameters, $q_i(\nu) \in [0,1]$ and $\sum_{\nu\leq k} q_i(\nu) = 1\ \forall i$. Notice that the probability of a configuration c is given by the probability of the individual assignments $q_i(c(i))$ and that there are no correlations between the assignments of different objects.

For convenience during the minimization process the explicit dependence of the probability distribution and cost function on the data is omitted. The Gibbs distribution for a general cost function $\mathcal{R}(C)$ can be written as

$$\mathbf{P}(\mathbf{c}) = \frac{e^{-\beta\mathcal{R}(c)}}{\sum_{c'\in\mathcal{C}} e^{-\beta\mathcal{R}(c')}} = \frac{e^{-\beta\mathcal{R}(c)}}{Z} = \exp\left(-\beta(\mathcal{R}(c)-\mathcal{F})\right), \tag{4.8}$$

where $\mathcal{F} = -\frac{1}{\beta}\log Z$ is known as "free energy" by analogy with statistical physics. As explained at the beginning of this section, the idea is to minimize the quantity $D(\mathbf{Q}||\mathbf{P})$ assuming that $\mathbf{Q}$ is of the form Eq. 4.7. By theorem 3 it holds that

$$0 \leq D(\mathbf{Q}||\mathbf{P}) \tag{4.9}$$

Introducing the explicit forms of the distributions Eq. (4.7) and Eq. (4.8) the inequality Eq. (4.9) can be written as

$$\begin{aligned} 0 &\leq \sum_{c\in\mathcal{C}}\prod_{i\leq n} q_i(c(i))\left(\sum_{s\leq n}\log q_s(c(s)) + \beta\mathcal{R}(c) - \beta\mathcal{F}\right)\\ &\leq \sum_{s\leq n}\left(\sum_{c\in\mathcal{C}}\prod_{i\leq n} q_i(c(i))\log q_s(c(s))\right) + \beta\mathbb{E}_{\mathbf{Q}}(\mathcal{R}(c)) - \beta\mathcal{F}, && (4.10)\end{aligned}$$

where the last two terms in the r.h.s of Eq. (4.10) follow from the definition of expectation value and the normalization of the distribution $\mathbf{Q}$. In order to simplify the first term of the r.h.s of Eq. (4.10) observe that the sum over all possible configurations $\sum_{c\in\mathcal{C}}$ can be decomposed as follows:

$$\sum_{c\in\mathcal{C}} = \sum_{r_1=1}^{k}\sum_{c\in\mathcal{C}^1_{r_1}}, \tag{4.11}$$

where $\mathcal{C}^1_{r_1}$ denotes the set of mappings that specifically map the first object to the r_1 cluster. $\sum_{c\in\mathcal{C}^1_{r_1}}$ can be calculated in a similar way:

$$\sum_{c\in\mathcal{C}^1_{r_1}} = \sum_{r_2=1}^{k} \sum_{c\in\mathcal{C}^{1,2}_{r_1,r_2}} , \tag{4.12}$$

where $\mathcal{C}^{1,2}_{r_1,r_2}$ is the set of all mappings that assign the first object to the r_1 cluster and the second to the r_2 cluster. After n iterations the result is then

$$\sum_{c\in\mathcal{C}} = \sum_{r_1=1}^{k}\sum_{r_2=1}^{k}\cdots\sum_{r_n=1}^{k} \sum_{c\in\mathcal{C}^{1,2,\ldots,n}_{r_1,r_2,\ldots,r_n}} , \tag{4.13}$$

where $\mathcal{C}^{1,2,\ldots,n}_{r_1,r_2,\ldots,r_n}$ is the set that contains the single mapping $(1,2,\ldots n)\mapsto(r_1,r_2,\ldots r_n)$. Using Eq. (4.13) the first term of the r.h.s of Eq. (4.10) can be written as

$$\begin{aligned}
&\sum_{s\leq n}\left(\sum_{c\in\mathcal{C}}\prod_{i\leq n} q_i(c(i))\log q_s(c(s))\right) = \\
&\sum_{s\leq n}\sum_{r_1=1}^{k}\sum_{r_2=1}^{k}\cdots\sum_{r_n=1}^{k}\sum_{c\in\mathcal{C}^{1,2,\ldots,n}_{r_1,r_2,\ldots,r_n}} q_i(c(i))\log q_s(c(s)) = \\
&\sum_{s\leq n}\sum_{r_1=1}^{k} q_i(r_1)\sum_{r_2=1}^{k} q_2(r_2)\cdots\sum_{r_n=1}^{k} q_n(r_n)\log q_s(r_s) = \\
&\sum_{s\leq n}\sum_{r_s=1}^{k} q_s(r_s)\log q_s(r_s)
\end{aligned} \tag{4.14}$$

Thus Eq. (4.10) can be cast in the following form

$$\mathcal{F} \leq \frac{1}{\beta}\sum_{s\leq n}\sum_{\nu\leq k} q_s(\nu)\log q_s(\nu) + \mathbb{E}_{\mathbf{Q}}(\mathcal{R}(c)) \equiv \mathcal{B}(\{q_i(\nu)\}), \tag{4.15}$$

The first term in the r.h.s of Eq. 4.15 corresponds to the scaled negative entropy of the approximate distribution, that term together with the expected cost define the free energy of the approximate model.[2] The bound $\mathcal{B}(\{q_i(\nu)\})$ is then the free energy of $\mathbf{Q}$ and it has to be minimized with respect to the $\{q_i(\nu)\}$ subject only to the normalization constrains.

The extremality conditions of $\mathcal{B}(\{q_i(\nu)\})$ are given by

$$0 = \frac{\partial}{\partial q_u(\alpha)}\left\{\mathcal{B}(\{q_i(\nu)\}) + \sum_{i\leq n}\lambda_i\Big(\sum_{\nu\leq k} q_i(\nu) - 1\Big)\right\}, \tag{4.16}$$

[2] Starting from the definition of the entropy of a distribution $S(\mathbf{P}) = -\sum_{x\in\Omega} p(x)\log p(x)$ and using the form of the Gibbs distribution Eq. (4.8) it can be shown that $\mathcal{F} = -\frac{1}{\beta}S(\mathbf{P}) + \mathbb{E}_{\mathbf{P}}\mathcal{R}$ which can be taken as a definition of free energy for any distribution.

where the λ_i are the Lagrange coefficients that are fixed by the normalization constrains. Using the explicit form of $\mathcal{B}(\{q_i(\nu)\})$ Eq. (4.16) takes the form

$$\begin{aligned} 0 =& \frac{\partial}{\partial q_u(\alpha)} \frac{1}{\beta} \sum_{s \le n} \sum_{\nu \le k} q_s(\nu) \log q_s(\nu) + \frac{\partial}{\partial q_u(\alpha)} \mathbb{E}_{\mathbf{Q}}(\mathcal{R}(c)) + \lambda_u \\ =& \frac{1}{\beta}(1 + \log q_u(\alpha)) + \frac{\partial}{\partial q_u(\alpha)} \sum_{c \in \mathcal{C}} \prod_{i \le n} q_i(c(i)) \mathcal{R}(c) + \lambda_u \\ =& \frac{1}{\beta}(1 + \log q_u(\alpha)) + \sum_{c \in \mathcal{C}} \prod_{i \ne u \le n} q_i(c(i)) \mathbb{I}_{\{c(u)=\alpha\}} \mathcal{R}(c) + \lambda_u. \end{aligned} \tag{4.17}$$

The second term in the r.h.s of Eq. (4.17) is called the mean field $h_{u\alpha}$ and it is the expected cost of $\mathcal{R}(c)$ under the constrain that object u is assigned to cluster α that is,

$$h_{u\alpha} = \mathbb{E}_{\mathbf{Q}_{u \to \alpha}}\{\mathcal{R}(c)\} = \sum_{c \in \mathcal{C}} \prod_{i \ne u \le n} q_i(c(i)) \mathbb{I}_{\{c(u)=\alpha\}} \mathcal{R}(c). \tag{4.18}$$

Using the mean field notation, the assignment probabilities can be obtained from Eq. (4.17) as

$$q_u(\alpha) = \exp\left(-1 - \beta(h_{u\alpha} + \lambda_u)\right),$$

and upon normalization

$$q_u(\alpha) = \frac{e^{-\beta(h_{u\alpha})}}{\sum_{\nu \le k} e^{-\beta(h_{u\nu})}}. \tag{4.19}$$

Having established the extremality conditions Eq. (4.17), it remains to verify that the second variations are positive namely

$$\frac{\partial^2}{\partial^2 q_u(\alpha)^2} \mathcal{B}(\{q_i(\nu)\}) = (\beta q_u(\alpha))^{-1} > 0. \tag{4.20}$$

Eq. (4.20) implies then that an asynchronous updating scheme of the form

$$\begin{aligned} q_{s(t)}^{\text{New}} &= \frac{e^{-\beta(h_{s(t)\alpha})}}{\sum_{\nu \le k} e^{-\beta(h_{s(t)\nu})}} \text{ Where} \\ h_{s(t)\alpha} &= \mathbb{E}_{\mathbf{Q}^{\text{Old}}_{s(t) \to \alpha}}\{\mathcal{R}(c)\} \end{aligned} \tag{4.21}$$

where $s(t)$ is an arbitrary site visitation scheme will converge to a local minimum of the approximate free energy Eq. (4.15) in the space of factorial distributions.

Some remarks are in order regarding the mean field approach. First of all, the factorial form of $\mathbf{Q}$ simplifies the summation over exponentially many assignment configurations k^n of n objects into k clusters, to kn summations and an optimization problem that can be tackled with an Expectation-Minimization approach. Second, the factorial form of $\mathbf{Q}$ is sufficient and convenient to perform a mean field treatment of a problem, but it is not necessary, it is sometimes possible to use distributions that contain correlations,tree like structures for example, that also allows for efficient computations.

Examples

Example 4 (Smooth $k-$means clustering)**.** *Consider the $k-$means cost function (Eq.(2.49)) to which a regularizer, that favors homogeneous assignments in the local neighborhood $\mathcal{N}(i)$ of each object i, is added*

$$\mathcal{R}^{skm}(c) = \sum_{i \leq n} ||x_i - y_{c(i)}||^2 + \frac{\lambda}{2} \sum_{i \leq n} \sum_{j \in \mathcal{N}(i)} \mathbb{I}_{\{c(i) \neq c(j)\}}.$$

Observe that if objects i and j are neighbors and the assignment c allocates both of them to different clusters then an extra cost of λ must be payed. In order to calculate $h_{u\alpha} = \mathbb{E}_{\mathbf{Q}_{u \to \alpha}}\{\mathcal{R}(c)\}$ it is useful to split $\mathcal{R}^{skm}(c)$ in to a term that contains the object u and one that does not:

$$\mathcal{R}^{skm}(c) = ||x_u - y_{c(u)}||^2 + \lambda \sum_{j \in \mathcal{N}(u)} \mathbb{I}_{\{c(u) \neq c(j)\}} + \mathcal{R}(c|u),$$

where $\mathcal{R}(c|u)$ are costs independent of the object u. $h_{u\alpha}$ is then given by

$$h_{u\alpha} = \mathbb{E}_{\mathbf{Q}_{u \to \alpha}}\{||x_u - y_{c(u)}||^2\} + \mathbb{E}_{\mathbf{Q}_{u \to \alpha}}\{\lambda \sum_{j \in \mathcal{N}(u)} \mathbb{I}_{\{c(u) \neq c(j)\}}\} + \mathbb{E}_{\mathbf{Q}_{u \to \alpha}}\{\mathcal{R}(c|u)\}$$

The first term in the r.h.s of the previous equation can be evaluated to $||x_u - y_\alpha||^2$, the last term is constant and independent of u so its value does not change if the cluster assignment of u, $c(u)$, does. The second term is evaluated as follows

$$\begin{aligned}
\mathbb{E}_{\mathbf{Q}_{u \to \alpha}}\{\lambda \sum_{j \in \mathcal{N}(u)} \mathbb{I}_{\{c(u) \neq c(j)\}}\} &= \lambda \sum_{j \in \mathcal{N}(u)} \mathbb{E}_{\mathbf{Q}_{u \to \alpha}} \mathbb{I}_{\{c(u) \neq c(j)\}} \\
&= \lambda \sum_{j \in \mathcal{N}(u)} \mathbb{E}_{\mathbf{Q}_{u \to \alpha}} \mathbb{I}_{\{\alpha \neq c(j)\}} \\
&= \lambda \sum_{j \in \mathcal{N}(u)} \mathbb{E}_{\mathbf{Q}_{u \to \alpha}} \sum_{\nu \neq \alpha \leq k} \mathbb{I}_{\{\nu = c(j)\}} \\
&= \lambda \sum_{j \in \mathcal{N}(u)} \sum_{\nu \neq \alpha \leq k} \mathbb{E}_{\mathbf{Q}_{u \to \alpha}} \mathbb{I}_{\{\nu = c(j)\}} \qquad (*)
\end{aligned}$$

In $(*)$, $\mathbb{E}_{\mathbf{Q}_{u \to \alpha}} \mathbb{I}_{\{\nu = c(j)\}}$ is evaluated as

$$\begin{aligned}
\mathbb{E}_{\mathbf{Q}_{u \to \alpha}} \mathbb{I}_{\{\nu = c(j)\}} &= \sum_{c \in \mathcal{C}} \prod_{i \neq u \leq n} q_i(c(i)) \mathbb{I}_{\{c(u) = \alpha\}} \mathbb{I}_{\{\nu = c(j)\}}. \\
&= q_j(\nu) \sum_{c \in \mathcal{C}} \prod_{i \neq \{u,j\} \leq n} q_i(c(i)) \mathbb{I}_{\{c(u) = \alpha\}} \mathbb{I}_{\{\nu = c(j)\}}. \\
&= q_j(\nu) \qquad (**)
\end{aligned}$$

In the previous equation, $(**)$ was obtained by using Eq.(4.13). Inserting $(**)$ in $(*)$ and collecting results the final form of $h_{u\alpha}$ is

$$h_{u\alpha} = ||x_u - y_\alpha||^2 + \lambda \sum_{j \in \mathcal{N}(u)} \sum_{\nu \neq \alpha \leq k} q_j(\nu) + \mathbb{E}_{\mathbf{Q}_{u \to \alpha}}\{\mathcal{R}(c|u)\} \qquad (4.22)$$