Correlation Clustering with Stochastic Labellings

Nicola Rebagliati[1], Samuel Rota Bulò[2], and Marcello Pelillo[2]

[1] VTT Technical Research Centre of Finland, 02044, Finland
nicola.rebagliati@gmail.com
[2] Department of Enviromental Science, Computer Science and Statistics, Universitá
Ca' Foscari Venezia, 30121 Italy
srotabul@dsi.unive.it
pelillo@dsi.unive.it

Abstract. Correlation clustering is the problem of finding a crisp partition of the vertices of a correlation graph in such a way as to minimize the disagreements in the cluster assignments. In this paper, we discuss a relaxation to the original problem setting which allows probabilistic assignments of vertices to labels. By so doing, overlapping clusters can be captured. We also show that a known optimization heuristic can be applied to the problem formulation, but with the automatic selection of the number of classes. Additionally, we propose a simple way of building an ensemble of agreement functions sampled from a reproducing kernel Hilbert space, which allows to apply correlation clustering without the empirical estimation of pairwise correlation values.

Keywords: Correlation clustering, stochastic labelling, ensemble clustering, Baum-Eagon inequality

1 Introduction

Correlation Clustering is a recent clustering formulation, introduced in [4], which consists in partitioning vertices of a graph, whose edges are labelled as positive (similar) or negative (dissimilar). The goal is to find a partition in such a way as to minimize the number of negative intra-cluster edges and positive inter-cluster edges. Such a setting can be found, *e.g.*, in document clustering, where the number of clusters (topics) is not known in advance and a classifier is given which outputs whether two documents are similar or not. Unlike traditional partitional clustering approaches, this formulation does not need the number of clusters as a user parameter, but it is able to automatically perform a model selection.

Due to the difficulty of the problem, which is NP-complete [4], much work has been done in the direction of finding bounds and approximate solutions. In [4], the authors provide a constant time approximation for minimizing the disagreement and a polynomial time approximation scheme for maximizing the agreements. Later theoretical and practical improvements were made by [1][11][26] with insightful approximation algorithms that exploit linear programming or semidefinite programming. A spectral approach to solve correlation clustering with 2 clusters has been proposed in [10]. A learning theoretical analysis of correlation clustering is presented in [17]. Practical considerations, comparison and experimentation with different algorithms, also heuristical ones, can be found in [18][12].

An important application of correlation clustering is *consensus* clustering [25,21,14], *i.e.* a methodology for summarizing an ensemble of different partitions of the same dataset into a single partition. The partitions are typically obtained by applying different clustering algorithms with possibly different parametrizations on the dataset. Correlation clustering can be used for the consensus clustering algorithm, by noting that each partition in the ensemble provides observations of graph vertices to co-occur in a cluster. Indeed, these observations can be combined to estimate the similarity or dissimilarity among vertices in the graph.

Motivation and Contribution. The classic correlation clustering formulation leads to a hard partition of the graph vertices. This inhibits the possibility of capturing overlapping clusters, which is useful in many applications. To overcome this limitation, we discuss in this paper two alternative formulations of correlation clustering, where the requirement of having a crisp partition of the graph vertices is relaxed by allowing probabilistic assignments of vertices to clusters, which are regarded to as stochastic labellings. By so doing, vertices can be potentially assigned to multiple clusters. However, we show that the first formulation is essentially equivalent to classic correlation clustering, whereas the second one is different as it is able to capture overlapping clusters, preserving nevertheless the important property of automatic selection of the number of clusters. For each formulation an iterative scheme, based on the work of [24,23], allows to find a locally minimizing solution. In addition, we introduce a simple way of building an ensemble of agreement functions sampled from a reproducing kernel Hilbert space, without resorting on empirical estimations of the probability that two vertices will co-occur in the same class.

Previous work Our reference scheme is an adaptation of [22] to correlation clustering. In [22] they use stochastic assignments for finding overlapping communities in a social network. See also [3] for a rather different approach to the problem of finding groups from similarity matrices. However both [3,22] fix the number of classes K. By modifying the approach of [24] we have a different algorithm which automatically selects the number of classes K. In [8] they attack the problem of finding overlapping groups in correlation clustering by extending the Correlation Clustering functional with multi-labelling functions, instead of relaxing the ownership assignments.

Outline. The paper is organized as follows. Section 2 formally introduces the problem of correlation clustering within a more general setting, where we might have missing edges in the graph and noisy labels on the edges. Section 3 introduces two relaxed formulations of correlation clustering, which allow for stochastic assignments of vertices to clusters, and show some theoretical properties among which the ability of capturing overlapping clusters. We address the optimization problems related to the two proposed formulations in Section 4, where we make use of a result due to Baum and Eagon. In section 5 we introduce our ensemble of agreement functions sampled from kernel space and in Section 6 we show experiments on real and synthetic datasets. In section 7 we draw the conclusions.

2 Correlation Clustering

A *correlation graph* $G = (V, E, w)$ is an edge-weighted graph without self-loops, where $V = \{1, \dots, n\}$ is a set of vertices, $E \subseteq V \times V$ is a set of edges and $w : E \to \{0, 1\}$ is a function mapping edges $(i, j) \in E$ to 1 or 0 according to whether i and j are *correlated* or not. Hereafter, we write w_{ij} for $w(i, j)$.

Let $L_k = \{1, \dots, k\}$ be a set of k labels. A (stochastic) *k-labelling*, or simply labelling if k is understood by the context, for a correlation graph $G = (V, E, w)$ is a matrix $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_n) \in \Delta_k^n$, where $\mathbf{y}_i \in \Delta_k$ is a probabilistic assignment of a label in L_k to a vertex $i \in V$, where

$$\Delta_k = \left\{ \mathbf{z} \in \mathbb{R}^k \, : \, \sum_{\ell \in L_k} z_\ell = 1 \text{ and } z_\ell \geq 0 \text{ for all } \ell \in L_k \right\}$$

is the $(k-1)$-dimensional *simplex*. We denote by $\Lambda_k = \Delta_k \cap \{0, 1\}^k$ the set of deterministic assignments of labels to vertices, *i.e.* the set of distributions with full mass on a specific label in L_k. A labelling $\mathbf{X} \in \Lambda_k^n$ is regarded as a *deterministic labelling*. Note that for stochastic as well as deterministic labellings, parameter k should be intended as the maximum number of labels assignable to vertices. This implies that some labels in L_k may not be used. Moreover, for all $k' > k$, Λ_k and Δ_k can be naturally embedded in $\Lambda_{k'}$ and $\Delta_{k'}$, respectively.

Given a deterministic labelling $\mathbf{X} \in \Lambda_k^n$ for a correlation graph $G = (V, E, w)$, we say that two vertices connected by an edge $(i, j) \in E$ *agree* if $\mathbf{x}_i^\top \mathbf{x}_j = w_{ij}$. We say that they *disagree*, in all other cases. The *total disagreement* $\phi_G(\mathbf{X})$ of a labelling $\mathbf{X} \in \Lambda_k^n$ for a correlation graph $G = (V, E, w)$ is the number of edges in G consisting of disagreeing vertices, *i.e.*

$$\phi_G(\mathbf{X}) = \sum_{(i,j) \in E} w_{ij}(1 - \mathbf{x}_i^\top \mathbf{x}_j) + (1 - w_{ij})\, \mathbf{x}_i^\top \mathbf{x}_j \, . \tag{1}$$

Similarly, the *total agreement* of a labelling $\mathbf{X} \in \Lambda_k^n$ for G is the number of edges in E consisting of agreeing vertices.

A *correlation k-clustering* of a correlation graph $G = (V, E, w)$ is a k-labelling $\mathbf{X}^* \in \Lambda_k^n$ minimizing the total disagreement, *i.e.*

$$\phi^*_{G,\Lambda_k} = \phi_G(\mathbf{X}^*) = \min \left\{ \phi_G(\mathbf{X}) \, : \, \mathbf{X} \in \Lambda_k^n \right\} \, . \tag{2}$$

A correlation n-clustering for a correlation graph G with n vertices is called simply a *correlation clustering* for G. As argued by [4], we can state the following remark.

Remark 1 (Model selection property). There is an optimal parameter value k^* such that $\phi^*_{G,\Lambda_k} \geq \phi^*_{G,\Lambda_{k^*}}$ holds for all k. Furthermore, if $k' > k$ it holds that $\phi^*_{G,\Lambda_{k'}} \leq \phi^*_{G,\Lambda_k}$. Hence, by selecting $k = n$, where n is the number of vertices of G, we are guaranteed that $\mathbf{X}^*$ is a k-labelling achieving minimum disagreement over all possible choices of k.

2.1 Clustering with Noisy Correlation Graphs

We depart from the standard correlation clustering problem, by assuming input graphs to be noisy with respect to the edge correlation values. Specifically, we are not given w_{ij} explicitly, but probabilities p_{ij} are provided of observing i and j correlated. Let $\mathcal{G} = (V, E, p)$ be a random variable generating correlation graphs (*random correlation graph variable*) with vertex set V and edge set E, where for each edge $(i, j) \in E$ the value of w_{ij} is independently drawn according to a Bernoulli distribution with parameter p_{ij}. The *expected total disagreement* of a labelling $\mathbf{X} \in \Lambda^n_k$ with respect to $\mathcal{G}$ is given by:

$$\phi_{\mathcal{G}}(\mathbf{X}) = \mathbb{E}_{\mathcal{G}}\left[\phi_{\mathcal{G}}(\mathbf{X})\right] = \sum_{(i,j)\in E} p_{ij} + \mathbf{x}_i^\top \mathbf{x}_j (1 - 2p_{ij})\,. \tag{3}$$

For notational convenience, we express total disagreement in equation (1) and expected total disagreement in equation (3) with the same symbol ϕ, but they differ in the subscript being a correlation graph in the former case and a random correlation graph variable in the latter.

In order to cope with random correlation graphs, we consider a correlation clustering formulation, where we aim at finding a labelling in such a way as to minimize the *expected* total disagreement with respect to a random correlation graph variable $\mathcal{G}$. This yields the following minimization problem

$$\phi^*_{\mathcal{G},\Lambda_k} = \phi_{\mathcal{G}}(\mathbf{X}^*) = \min\left\{\phi_{\mathcal{G}}(\mathbf{X}) \,:\, \mathbf{X} \in \Lambda^n_k\right\}\,, \tag{P}$$

where $\mathbf{X}^* \in \Lambda^n_k$ denotes a labelling achieving minimum expected disagreement. The model selection property stated in Remark 1 holds straightforwardly also for this formulation. Note that weighted versions of correlation clustering has been addressed also in [17].

3 Relaxed Formulations with Stochastic Labellings

In this section we will relax the assumption on the labelling by allowing for stochastic assignments of vertices to labels. There is a two-fold reason why we introduce stochastic labellings. In first place it allows us to move from a discrete optimization problem to a continuous one and make use of a result known as

Baum-Eagon inequality in probability domain for finding a local solution (see Section 4). Secondly, having stochastic label assignments allows to capture overlapping clusters, by letting graph vertices to be assigned to more labels with non-zero probability.

We move from deterministic labellings to stochastic ones by replacing the variables $\mathbf{X} \in \Lambda_k^n$ with variables $\mathbf{Y} \in \Delta_k^n$ in (3):

$$\phi_{\mathcal{G}}(\mathbf{Y}) = \sum_{(i,j)\in E} p_{ij} + \mathbf{y}_i^\top \mathbf{y}_j (1 - 2p_{ij})\,. \tag{4}$$

Here, $\mathbf{y}_i^\top \mathbf{y}_j$ represents the probability of vertices i and j to occur in the same class, under independence assumption. The relaxed version of correlation k-clustering can thus be formulated as

$$\phi^*_{\mathcal{G},\Delta_k} = \phi_{\mathcal{G}}(\mathbf{Y}^*) = \min\left\{\phi_{\mathcal{G}}(\mathbf{Y}) \,:\, \mathbf{Y} \in \Delta_k^n\right\}\,, \tag{Q1}$$

where $\mathbf{Y}^* \in \Delta_k^n$ denotes an optimal stochastic k-labelling achieving minimum expected disagreement.

The relaxed formulation of correlation clustering in (Q1) is a continuous optimization problem, which turns out to be substantially equivalent to (P). Consequently, despite the stochastic label assignments, overlapping clusters are not captured. The following proposition shows that, for all choices of k, (P) and (Q1) yield the same value.

Proposition 1. *Let $\mathcal{G} = (V, E, p)$ be a random correlation graph variable. Then $\phi^*_{\mathcal{G},\Lambda_k} = \phi^*_{\mathcal{G},\Delta_k}$ for all choices of $k > 0$.*

Proof. Note that any variable $X \in \Lambda_k^n \subset \Delta_k^n$. Hence, the domain of program (P) is a strict subset of the one of (Q1), which implies $\phi^*_{\mathcal{G},\Lambda_k} \geq \phi^*_{\mathcal{G},\Delta_k}$. On the other hand, let $Y^* = (\mathbf{y}_1^*, \ldots, \mathbf{y}_n^*)$ be a solution of (Q1), let $\mathcal{X}_i \in \Lambda_k$, $1 \leq i \leq n$, be multinomial random vectors with parameters $n = 1$ and probabilities $\mathbf{y}_i^*$, and let $\mathcal{X} = (\mathcal{X}_1, \ldots \mathcal{X}_n) \in \Lambda_k^n$ be a random (deterministic) labelling generator. Then $\mathbb{E}_{\mathcal{X}}[\phi_{\mathcal{G}}(\mathcal{X})] \geq \phi^*_{\mathcal{G},\Lambda_k}$, but since $\mathbb{E}_{\mathcal{X}}[\phi_{\mathcal{G}}(\mathcal{X})] = \phi^*_{\mathcal{G},\Delta_k}$ we have that $\phi^*_{\mathcal{G},\Delta_k} \geq \phi^*_{\mathcal{G},\Lambda_k}$.

We show in Figure 1 an example of correlation clustering, where we have 3 clear overlapping clusters. In 1(a) we show the values of p_{ij} and in 1(b) we can clearly see that the solution obtained by (Q1) is a deterministic labelling $\mathbf{Y}^* \in \Lambda_k^n$ as the matrix of probabilities of co-occurrence $(\mathbf{Y}^*)^\top \mathbf{Y}^*$ contains 0s and 1s. This confirms the intuition coming from Proposition 1 and shows a clear inability of this formulation to capture overlapping clusters.

In order to overcome the limitations of (Q1) we consider a different way of computing the total disagreement of a labelling $\mathbf{X} \in \Lambda_k$ for a correlation graph $G = (V, E, w)$, which is given by

$$\varphi_G(\mathbf{X}) = \sum_{(i,j)\in E} \left(\mathbf{x}_i^\top \mathbf{x}_j - w_{ij}\right)^2\,. \tag{5}$$

In the presence of random correlation graphs generated according to $\mathcal{G} = (V, E, p)$, the corresponding expected total disagreement of a labelling $\mathbf{X}$ for $\mathcal{G}$ gives

$$\varphi_{\mathcal{G}}(\mathbf{X}) = \mathbb{E}_{\mathcal{G}}\left[\varphi_{\mathcal{G}}(\mathbf{X})\right] = \sum_{(i,j)\in E} p_{ij} + \mathbf{x}_i^\top \mathbf{x}_j (\mathbf{x}_i^\top \mathbf{x}_j - 2p_{ij})\,. \tag{6}$$

Note that $\varphi_G(\mathbf{X}) = \phi_G(\mathbf{X})$ and $\varphi_{\mathcal{G}}(\mathbf{X}) = \phi_{\mathcal{G}}(\mathbf{X})$. The relaxed version of (6), which uses a stochastic labelling $\mathbf{Y}$, is

$$\varphi_{\mathcal{G}}(\mathbf{Y}) = \sum_{(i,j)\in E} p_{ij} + \mathbf{y}_i^\top \mathbf{y}_j (\mathbf{y}_i^\top \mathbf{y}_j - 2p_{ij})\,. \tag{7}$$

Finally, the relaxed correlation k-clustering formulation related to (7) is given by

$$\varphi^*_{\mathcal{G},\Delta_k} = \varphi_{\mathcal{G}}(\mathbf{Y}^*) = \min\left\{\varphi_{\mathcal{G}}(\mathbf{Y}) \,:\, \mathbf{Y} \in \Delta_k^n\right\}\,, \tag{Q2}$$

where $\mathbf{Y}^* \in \Delta_k^n$ denotes an optimal stochastic k-labelling for the minimization.

Let $d_{\mathcal{G}}(\mathbf{Y})$ be the following function

$$d_{\mathcal{G}}(\mathbf{Y}) = \sum_{(i,j)\in E} \mathbf{y}_i^\top \mathbf{y}_j (1 - \mathbf{y}_i^\top \mathbf{y}_j)$$

which measures the uncertainty of the stochastic labelling $\mathbf{Y}$. Indeed, $d_{\mathcal{G}}(\mathbf{X}) = 0$ for all $\mathbf{X} \in \Lambda_k^n$, while it is strictly positive in general.

The next result, which is close in spirit to Proposition 1, relates the correlation clustering formulations (Q2) and (P). Specifically it provides a lower and upper bound for (P) in terms of (Q2) and $d_{\mathcal{G}}(\cdot)$ for all choices of k.

Proposition 2. *Let $\mathcal{G} = (V, E, p)$ be a random correlation graph variable. Then*

$$\varphi^*_{\mathcal{G},\Delta_k} \le \phi^*_{\mathcal{G},\Lambda_k} \le \varphi^*_{\mathcal{G},\Delta_k} + d_{\mathcal{G}}(\mathbf{Y}^*)$$

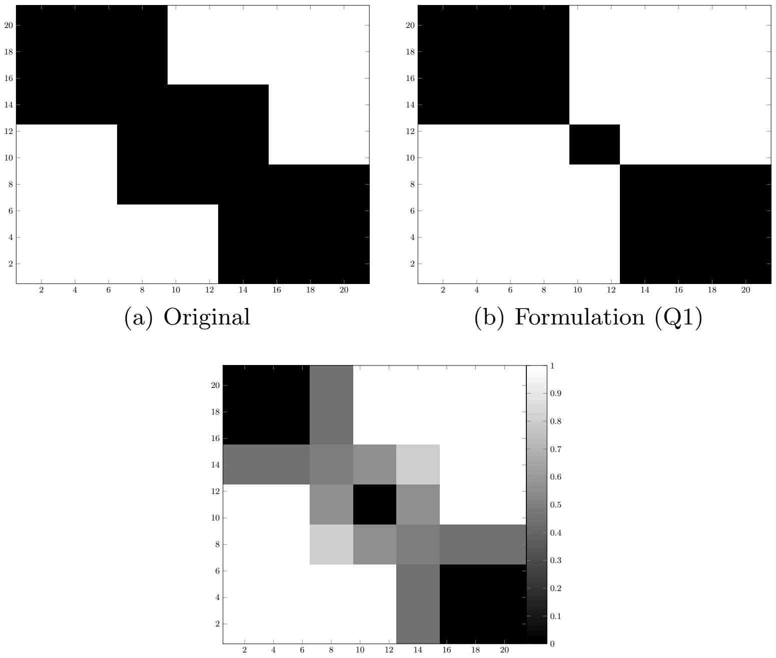
for all choices of $k > 0$, where $\mathbf{Y}^* \in \Delta_k^n$ is a solution of (Q2).

Proof. The first inequality $\varphi^*_{\mathcal{G},\Delta_k} \le \phi^*_{\mathcal{G},\Lambda_k}$ trivially holds because $\Lambda_k \subset \Delta_k$. The second inequality follows by noting that $\phi_{\mathcal{G}}(\mathbf{Y}) = \varphi_{\mathcal{G}}(\mathbf{Y}) + d_{\mathcal{G}}(\mathbf{Y})$, which implies $\phi^*_{\mathcal{G},\Delta_k} \le \varphi^*_{\mathcal{G},\Delta_k} + d_{\mathcal{G}}(\mathbf{Y}^*)$. By Proposition 1 the result derives.

From Proposition (2) we can see that if the solution of (Q2) is deterministic, then it is also a solution of (P). Otherwise, the higher the distance from a deterministic labelling, the larger the gap between $\varphi^*_{\mathcal{G},\Delta_k}$ and $\phi^*_{\mathcal{G},\Lambda_k}$ might be.

In Figure 1(c) we show the behaviour of formulation (Q2) with the toy example with 3 overlapping clusters, which has been previously introduced. We note that as opposed to (Q1), this formulation is indeed able to assign vertices to multiple classes, obtaining thereby a solution which reflects to the desired clustering.

Also for formulations (Q1) and (Q2) the model selection property of Remark 1 holds, clearly on the respective objective functions. It is worth mentioning that a formulation, which is equivalent to (Q2), has been used in [22] for communities



(a) Original

(b) Formulation (Q1)

(c) Formulation (Q2)

Fig. 1. Example of correlation clustering with 3 clear overlapping clusters. Left to right: Original correlation graph; $(\mathbf{Y}^*)^\top \mathbf{Y}^*$ with $\mathbf{Y}^*$ solution of (Q1); $(\mathbf{Y}^*)^\top \mathbf{Y}^*$ with $\mathbf{Y}^*$ solution of (Q2).

detection. However, the authors were not aware of the relation with correlation clustering and, thus, the automatic selection of the number of clusters.

Formulation (Q2) is, unfortunately, an highly non-convex minimization problem which is very difficult to attack with an exact algorithm working in a reasonable computational time. In the next section we propose to use non-exact algorithms based on two iterative formulations, for both (Q1) and (Q2), which ensure to return a locally minimizing solution.

4 Optimization using the Baum-Eagon Inequality

In order to solve our optimization problem we shall use the following important result which is generally known as the Baum-Eagon inequality [5].

Theorem 1 (Baum-Eagon). *Let $\mathbf{Y} \in \Delta_k^n$ and $Q(\mathbf{Y})$ be a homogeneous polynomial in the variables $y_{i\ell}$ with nonnegative coefficients. Define the mapping $\mathbf{Z} = \mathcal{M}(\mathbf{Y}) \in \Delta_k^n$ as follows:*

$$z_{i\ell} = y_{i\ell} \frac{\partial Q(\mathbf{Y})}{\partial y_{i\ell}} \bigg/ \sum_{\ell' \in L_k} y_{i\ell'} \frac{\partial Q(\mathbf{Y})}{\partial y_{i\ell'}} \,, \tag{8}$$

for all $i = 1 \dots n$ and $\ell \in L_k$. Then $Q(\mathcal{M}(\mathbf{Y})) > Q(\mathbf{Y})$, unless $\mathcal{M}(\mathbf{Y}) = \mathbf{Y}$. In other words $\mathcal{M}$ is a growth transformation for the polynomial Q.

Although the original theorem applies to homogeneous polynomials only, the result has been generalized later by Baum and Sell [7] who proved that Theorem 1 still holds in the case of arbitrary polynomials with nonnegative coefficients, and further extended the result by proving that $\mathcal{M}$ increases Q *homotopically*, which means that for all $0 \leq \eta \leq 1$, $Q(\eta\mathcal{M}(\mathbf{Y})+(1-\eta)\mathbf{Y}) \geq Q(\mathbf{Y})$ with equality if and only if $\mathcal{M}(\mathbf{Y}) = \mathbf{Y}$.

The Baum-Eagon inequality provides an effective iterative means for maximizing polynomial functions in probability domains, and in fact it has served as the basis for various statistical estimation techniques developed within the theory of probabilistic functions of Markov chains [6]. As pointed out in [7], we remark that the mapping $\mathcal{M}$ defined in Theorem 1 makes use of the first derivative only and yet is able to take finite steps while increasing Q. This contrasts sharply with classical gradient methods, for which an increase in the objective function is guaranteed only when infinitesimal steps are taken, and determining the optimal step size entails computing higher-order derivatives.

It is not difficult to show that, by starting from the interior of the simplex, the fixed points of the Baum-Eagon dynamics satisfy the first-order Karush-Kuhn-Tucker necessary conditions for local maxima and that strict local solutions are in correspondence to asymptotically stable points.

4.1 Algorithms for Correlation Clustering with Stochastic Labellings

We show now how the Baum-Eagon inequality can be used in order to optimize the relaxed formulations of correlation k-clustering introduced in Section 3. The theorem, however, cannot be applied directly as its hypothesis are not fulfilled. Indeed, the polynomials with variables in probability domain of (Q1) and (Q2) need to be minimized and not maximized, and they do not have in general nonnegative coefficients. Nevertheless, by exploiting the simplex constraints, we can transform the aforementioned formulations into equivalent ones, which can then be tackled by using the Baum-Eagon theorem. Hereafter, we denote with $\mathbf{E}$ a $k \times k$ matrix of all 1's, and with $\mathbf{I}$ the $k \times k$ identity matrix.

As for (Q1), by observing that $\mathbf{y}_i^\top \mathbf{E} \mathbf{y}_j = 1$ for all $(i, j) \in E$ and $\mathbf{Y} \in \Delta_k^n$, it is straightforward to rewrite $-\phi_{\mathcal{G}}(\mathbf{Y})$ as

$$-\phi_{\mathcal{G}}(\mathbf{Y}) = -|E| + \sum_{(i,j)\in E} \mathbf{y}_i^\top [\mathbf{E} + (2p_{ij} - 1)\mathbf{I}]\mathbf{y}_j - p_{ij}$$

which is a homogeneous polynomial with nonnegative coefficients (constant terms can be dropped), in probability domain Δ_k^n. This equivalence allows us to find a local solution of (Q1) by maximizing $-\phi_{\mathcal{G}}$. Hence, we can apply the Baum-Eagon theorem by using (8) with $Q = -\phi_{\mathcal{G}}$. This yields the following update rule for $\mathbf{Y} = (y_{i\ell})$:

$$y_{i\ell}^{(t+1)} = y_{i\ell}^{(t)} \frac{\left[\sum_{j\in E_i} 1-(1-2p_{ij})y_{j\ell}^{(t)}\right]}{\sum_{\ell\in L_k} y_{i\ell}^{(t)}\left[\sum_{j\in E_i} 1-(1-2p_{ij})y_{j\ell}^{(t)}\right]}\,, \tag{Alg-Q1}$$

where $E_i = \{j \mid (i,j) \in E\}$ and the starting labelling $\mathbf{Y}^{(0)}$ might be any point in the interior of Δ_k^n.

Similarly for (Q2), we can rewrite $-\varphi_{\mathcal{G}}(\mathbf{Y})$ as the following homogeneous polynomial with nonnegative coefficients:

$$-\varphi_{\mathcal{G}}(\mathbf{Y}) = -|E| + \sum_{(i,j)\in E} \left[\mathbf{y}_i^\top(\mathbf{E}-\mathbf{I})\mathbf{y}_j\right]^2 - p_{ij}$$

which can be locally maximized by means of the Baum-Eagon result obtaining a local solution of (Q2). This yields the following update rule:

$$y_{i\ell}^{(t+1)} = y_{i\ell}^{(t)} \frac{\displaystyle\sum_{j\in E_i}\left(1-y_{j\ell}^{(t)}\right)(1-\mathbf{y}_i^\top\mathbf{y}_j)+2p_{ij}y_{j\ell}^{(t)}}{\displaystyle\sum_{\ell\in L_k} y_{i\ell}^{(t)} \sum_{j\in E_i}\left(1-y_{j\ell}^{(t)}\right)(1-\mathbf{y}_i^\top\mathbf{y}_j)+2p_{ij}y_{j\ell}^{(t)}}\,, \tag{Alg-Q2}$$

where the starting labelling $\mathbf{Y}^{(0)}$ might be any point in the interior of Δ_k^n.

Both update rules (Alg-Q1) and (Alg-Q2) satisfy the invariant property $\mathbf{Y}^{(t)} \in \Delta_k^n$ for all $t > 0$ if $\mathbf{Y}^{(0)} \in \Delta_k^n$ and lead to a local solution of the respective correlation clustering formulations.

5 Ensemble of Random Functions Sampled from Kernel Space

In this section we show how to construct a simple ensemble of agreement functions sampled from a reproducing kernel Hilbert space, which allows to obtain a random correlation graph variable for our algorithm from an arbitrary clustering dataset, without resorting on empirical estimations of the probability that two vertices will co-occur in the same class. This is an alternative approach to in [13].

A *kernel* is a symmetric function $K : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ such that for any dataset $(\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{X}^n$ the comparison matrix $\mathbf{K}$ with entries $k_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$, regarded to as *Gram matrix*, is positive semidefinite, i.e. all its eigenvalues are nonnegative. A kernel uniquely determines a *reproducing kernel Hilbert space* [2]. This is a vector space $\mathcal{H}$ of functions $f : \mathbb{X} \to \mathbb{R}$ with the following properties:

- $f(\mathbf{x}) = \langle f, K(\mathbf{x}, \cdot)\rangle_{\mathcal{H}}$
- $\forall \mathbf{x} \in \mathbb{X}.\, K(\mathbf{x}, \cdot) \in \mathcal{H}$

where $\mathcal{H} = \overline{span\{K(\mathbf{x}, \cdot) : \mathbf{x} \in \mathcal{H}\}}$.

A *feature map* is a function $\Phi : \mathbb{X} \to \mathcal{H}$ associated to a kernel K such that $k_{ij} = \langle \Phi_i, \Phi_j \rangle_{\mathcal{H}}$, where $\Phi_i = \Phi(\mathbf{x}_i)$. By the reproducing kernel property, we can associate each function $f \in \mathcal{H}$ with an evaluating hyperplane w_f, such that $f(\mathbf{x}) = \langle w_f, \Phi(\mathbf{x}) \rangle_{\mathcal{H}}$. At the same time, a function $f \in \mathcal{H}$, can be regarded as a 2-class classifier mapping $\mathbf{x} \in \mathbb{X}$ to a label according to the sign of $f(\mathbf{x})$.

The probability p_{ij} that a randomly drawn function from $f \in \mathcal{H}$ with uniform distribution will put two data points $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{X}$ in the same class can be computed as a function of the angle θ_{ij} between Φ_i and Φ_j [16]:

$$p_{ij} = 1 - \frac{\theta_{ij}}{\pi} \,.$$

The angle here is given by $\theta_{ij} = \arccos(k_{ij}/\sqrt{k_{ii}k_{jj}})$. The matrix $\mathbf{P} = (p_{ij})$ of the probabilities can thus be computed as the following function of the kernel matrix $\mathbf{K}$:

$$\mathbf{P} = \mathbf{E} - \frac{1}{\pi} \arccos(\mathbf{D}_{\mathbf{K}}^{-\frac{1}{2}} \mathbf{K} \mathbf{D}_{\mathbf{K}}^{-\frac{1}{2}}) \,,$$

where $\mathbf{D}_{\mathbf{K}}$ is the diagonal of $\mathbf{K}$. Note that for the Gaussian kernel the formula is simpler because the features have norm 1 as $K(\mathbf{x}, \mathbf{x}) = 1$. In this case indeed we obtain

$$\mathbf{P}_{\text{rbf}} = \mathbf{E} - \frac{1}{\pi} \arccos(\mathbf{K}_{\text{rbf}}) \,.$$

Matrix $\mathbf{P}$ can be used to obtain a random correlation graph variable $\mathcal{G}$ representing the data to cluster by means of one of the correlation clustering approaches, previously described.

In order to account for classifiers with a larger number of classes, we consider the possibility of specifying the number of functions f that should be drawn from $\mathcal{H}$ for the classification. Under independence assumption the probability that two sample points $\mathbf{x}_i$ and $\mathbf{x}_j$ will be given the same class by each of the sampled functions, say d, is simply p_{ij}^d.

6 Experiments

In this section we assess the effectiveness of the relaxed formulations introduced in section 3 on both real and synthetic datasets.

For the experiments we considered the heuristics we introduced in Section 4, namely Alg-Q1 and Alg-Q2, which provide solutions to (Q1) and (Q2), respectively. We compared our algorithms against two heuristics for (P). The first is a randomized heuristic, called CC-Pivot, yielding a 11/7 approximation, which has been introduced in [1]. The second one is a local search heuristic, called Best One Element Move (BOEM), introduced in [15]. All four heuristics are repeated with 25 different random initializations and best results are returned.

We evaluated the algorithms on two real datasets from the UCI Machine Learning Repository: Iris and House-Votes. Iris consists of 150 data points in 4-dimensional space divided uniformly into 3 classes. House-Votes consists of 435 data points in 17-dimensional space divided into 2 classes (267/168). We also considered a synthetic dataset “4NG” composed by four overlapping gaussians with 50 points each and 50 points uniformly sampled in the hyperbox containing the data as outliers.

For each dataset we created a random correlation graph variable according to the method described in Section 5 in conjunction with a RBF kernel with manually tuned scale parameter, and with $d = 3$ sampled functions.

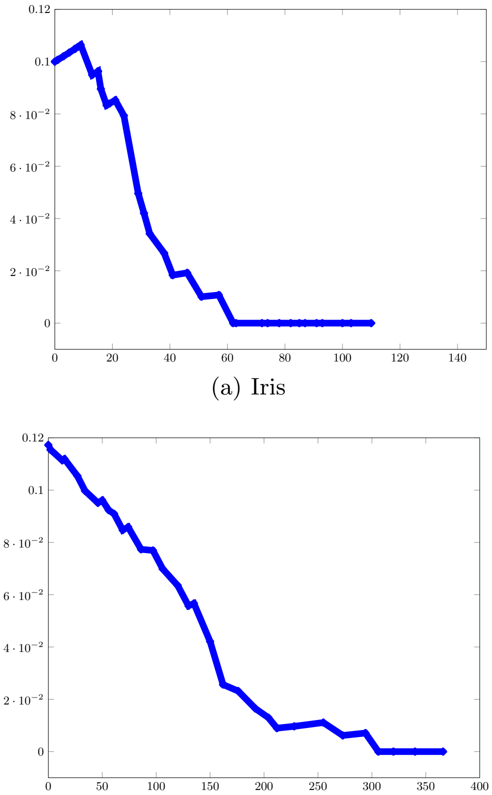
Note that as for our algorithms, we run them with a maximum number of classes $k = 20$, which was larger than the number of classes found in the datasets. By so doing, the algorithms were able to automatically find the number of clusters. The running time of Alg-Q1 and Alg-Q2 is comparable to other methods and take few minutes (< 15) with Matlab 7.8.0 [19] for Windows 7 ©Intel ®Core ™Duo CPU T6600 2.20GHz, 4GB RAM.

We assessed the quality of the clusterings obtained from the algorithms by computing the *confusion error* [20]. Since confusion error does not penalize the selection of a number of clusters larger than the ground truth we report also the associated number of clusters.

In Table 1 we report best results obtained by all the methods on all datasets. Beside the name of each dataset, we show the optimal number of classes. For each combination of dataset and algorithm we provide the number of classes obtained and the associated confusion error. As we can see among the four approaches, Alg-Q2 is the one achieving the best compromise between the automatic selection of the number of classes, and the confusion error, while the other approaches tend to overestimate the number of actual clusters in the data. Note that an advantage of having stochastic labelling is that we can measure the uncertainty in a label assignment. Since our algorithm is the only one which is able to capture such information, we report in Figure 2 the effect on the confusion error of the removal of points with the most uncertain label assignments obtained by it. As we can see, the error nicely decreases to zero. This indicates that the points where the algorithm exhibits uncertain label assignments are those leading to misclassification.

Dataset (K)	σ	BOEM	CC-Pivot	Alg-Q1	Alg-Q2
Iris (3)	0.4	(31, 0.08)	(10, 0.10)	(11, 0.13)	(3, 0.10)
House-Votes (2)	0.8	(8, 0.11)	(5, 0.14)	(20, 0.37)	(2, 0.12)
Ten-Digits (10)	0.05	*	*	(20, 0.21)	(15, 0.17)
4NG (4)	0.1	(42, 0.13)	(56, 0.10)	(19, 0.13)	(7, 0.16)

Table 1. Results obtained on the datasets. We report for each combination of dataset and algorithm the number of clusters found by the algorithm and the confusion error of the solution found. We also report the optimal value of σ used for the experiment. For the Ten-Digits dataset both BOEM and CC-Pivot returned an high number of classes and their result are not significant.



(a) Iris

(b) House-Votes

Fig. 2. Plot of the confusion error obtained by (Alg-Q2) by iteratively removing vertices with uncertain labels. On the x-axis we report the number of vertices removed from the dataset.

We also compared our method with the algorithm Left-Stochastic Decomposition (LSD) of [3] on datasets from [9] using the Misclassification Error [20]. As we can see from Table 2, both approaches perform comparably well, although our method achieves the best scores on most of the datasets that have been taken into account.

7 Conclusions

The aim of this work is showing the relationship between classical Correlation clustering and a relaxed version which allows for stochastic labellings instead of hard ones. In proposition 1 we show that this relaxation is necessary, because Correlation clustering by itself cannot capture stochastic labellings. In proposition 2 the two functionals are put in relation. Moreover, we argue that the relaxation still preserves the property of model selection peculiar of Correlation

Dataset (K)	Alg-Q2	LSD
Amazon Binary (2)	**.354**	.390
Aural Sonar (2)	**.120**	.140
Patrol (8)	**.253**	.440
Protein (4)	.347	**.200**
Voting (2)	**.094**	.100
Yeast Pfam 7-12 (2)	.380	**.360**
Yeast SW 5-7 (2)	.295	**.28**
Yeast SW 5-12 (2)	**.090**	**.090**
Yeast SW 7-12 (2)	**.095**	.100

Table 2. A comparison with [3] on datasets of [9]. Number of used clusters in parenthesis.

Clustering. For both formulations we provide how to apply the Baum-Eagon inequality in order to obtain converging algorithms. As a further contribution, we show how we can practically build a simple ensemble of agreement functions sampled from a reproducing kernel Hilbert space of functions. In the experiments we obtain promising results compared to other, state-of-the-art, methods.

Acknowledgments We acknowledge financial support from the FET programme within the EU FP7, under the SIMBAD project (contract 213250) and a tenure of an ERCIM "Alain Bensoussan" Fellowship Programme under the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement 246016.

References

1. Ailon, N., Charikar, M., Newman, A.: Aggregating inconsistent information: ranking and clustering. In: STOC. (2005) 684–693
2. Aronszajn, N.: Theory of reproducing kernels. Trans. Amer. Math. Soc. **68** (1950) 337–404
3. Arora, R., Gupta, M., Kapila, A., Fazel, M.: Clustering by left-stochastic matrix factorization. In: ICML. (2011) 761–768
4. Bansal, N., Blum, A., Chawla, S.: Correlation clustering. Machine Learning **56**(1-3) (2004) 89–113
5. Baum, L.E., Eagon, J.A.: An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. Bull. Amer. Math. Soc. **73** (1967) 360–363
6. Baum, L.E., Petrie, T., Soules, G., Weiss, N.: A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Annals of Math. Statistics **41** (1970) 164–171
7. Baum, L.E., Sell, G.R.: Growth transformations for functions on manifolds. Pac. J. Math **27** (1968) 221–227
8. Bonchi, F., Gionis, A., Ukkonen, A.: Overlapping correlation clustering. In: ICDM. (2011) 51–60
9. Chen, Y., Garcia, E.K., Gupta, M.R., Rahimi, A., Cazzanti, L.: Similarity-based Classification: Concepts and Algorithms. Journal of Machine Learning Research **10** (March 2009) 747–776

10. Coleman, T., Saunderson, J., Wirth, A.: Spectral clustering with inconsistent advice. In: ICML. (2008) 152–159
11. Demaine, E.D., Emanuel, D., Fiat, A., Immorlica, N.: Correlation clustering in general weighted graphs. Theor. Comput. Sci. **361**(2-3) (2006) 172–187
12. Downing, N., Stuckey, P.J., Wirth, A.: Improved consensus clustering via linear programming. In: ACSC. (2010) 61–70
13. Fern, X.Z., Brodley, C.E.: Random projection for high dimensional data clustering: A cluster ensemble approach. In Fawcett, T., Mishra, N., eds.: ICML, AAAI Press (2003) 186–193
14. Fred, A.L.N., Jain, A.K.: Combining multiple clusterings using evidence accumulation. IEEE Trans. Pattern Anal. Mach. Intell. **27**(6) (2005) 835–850
15. Gionis, A., Mannila, H., Tsaparas, P.: Clustering aggregation. In: In Proceedings of the 21st International Conference on Data Engineering (ICDE). (2005) 341–352
16. Goemans, M.X., Williamson, D.P.: Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. J. ACM **42**(6) (1995) 1115–1145
17. Joachims, T., Hopcroft, J.E.: Error bounds for correlation clustering. In: ICML. (2005) 385–392
18. Mathieu, C., Schudy, W.: Bounding and comparing methods for correlation clustering beyond ilp. In: ILP-NLP. (2009)
19. MATLAB: version 7.8.0 (R2009a). The MathWorks Inc., Natick, Massachusetts (2009)
20. Meila, M.: Comparing Clusterings by the Variation of Information. (2003) 173–187
21. Monti, S., Tamayo, P., Mesirov, J.P., Golub, T.R.: Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning **52**(1-2) (2003) 91–118
22. Nepusz, T., Petróczi, A., Négyessy, L., Bazsó, F.: Fuzzy communities and the concept of bridgeness in complex networks. Physical Review E **77**(1) (January 2008) 016107+
23. Rota Bulò, S., Lourenço, A., Fred, A.L.N., Pelillo, M.: Pairwise probabilistic clustering using evidence accumulation. In: SSPR/SPR. (2010) 395–404
24. Rota Bulò, S., Pelillo, M.: Probabilistic clustering using the baum-eagon inequality. In: ICPR. (2010) 1429–1432
25. Strehl, A., Ghosh, J.: Cluster ensembles — a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research **3** (2002) 583–617
26. Swamy, C.: Correlation clustering: maximizing agreements via semidefinite programming. In: SODA. (2004) 526–527