Co-Clustering via Information-Theoretic Markov Aggregation

Clemens Blöchl, Rana Ali Amjad, *Student Member, IEEE*, and Bernhard C. Geiger, *Member, IEEE*

Abstract—We present an information-theoretic cost function for co-clustering, i.e., for simultaneous clustering of two sets based on similarities between their elements. By constructing a simple random walk on the corresponding bipartite graph, our cost function is derived from a recently proposed generalized framework for information-theoretic Markov chain aggregation. The goal of our cost function is to minimize *relevant information loss*, hence it connects to the information bottleneck formalism. Moreover, via the connection to Markov aggregation, our cost function is not *ad hoc*, but inherits its justification from the operational qualities associated with the corresponding Markov aggregation problem. We furthermore show that, for appropriate parameter settings, our cost function is identical to well-known approaches from the literature, such as "Information-Theoretic Co-Clustering" by Dhillon et al. Hence, understanding the influence of this parameter admits a deeper understanding of the relationship between previously proposed information-theoretic cost functions. We highlight some strengths and weaknesses of the cost function for different parameters. We also illustrate the performance of our cost function, optimized with a simple sequential heuristic, on several synthetic and real-world data sets, including the Newsgroup20 and the MovieLens100k data sets.

Index Terms—Co-clustering, information-theoretic cost function, clustering, Markov chains

1 Introduction and Outline

Co-clustering is the task of the simultaneous clustering of two sets, typically represented by rows and columns of a data matrix. Aside from being a clustering problem in its own right, co-clustering is also applied for clustering only one dimension of the data matrix. In these scenarios, co-clustering is an implicit method for feature clustering and provides an alternative to feature selection with, purportedly, increased robustness to noisy data [1], [2], [3].

A popular approach to co-clustering employs information-theoretic cost functions and is based on transforming the data matrix into a probabilistic description of the two sets and their relationship. For example, if the entries in the data matrix are all nonnegative, one can normalize the data matrix to obtain a joint probability distribution of two random variables taking values in the two sets. This approach has been taken by, e.g., Slonim et al. [1], Bekkerman et al. [4], El-Yaniv and Souroujon [5], and Dhillon et al. [2] (see also Section 2). A different approach to co-clustering is to identify the data matrix with the weight matrix of a bipartite graph and subsequently apply graph partitioning methods to cluster the rows and columns of the data matrix. This approach has been taken by, e.g., Dhillon [6], Labiod and Nadif [7], and Ailem et al. [8]. Other popular approaches are model-based (e.g., latent block models as in [9] and the references therein) or based on nonnegative matrix factorization (e.g., [10, Sec. 4.4]).

In this work, we combine ideas from the graph-based and the information-theoretic approaches. Specifically, we use the data matrix to define a simple random walk on a bipartite graph, i.e., a first-order, stationary Markov chain. Clustering this bipartite graph (i.e., co-clustering) thus becomes equivalent to clustering the state space of a Markov chain (i.e., Markov aggregation, cf. Section 3). This, in turn, allows us to transfer the information-theoretic cost function from the latter problem to the former. The thus presented cost function, parameterized by a single parameter β, derives its justification from the corresponding Markov aggregation problem. This justification is further inherited to other information-theoretic cost functions previously proposed in the literature [1], [2], [3], [4], [11], which we obtain as special cases for appropriate choices of β.

In several examples we discuss weaknesses inherent in the cost function for certain values (or value ranges) of β. We also present a simple sequential heuristic to optimize our cost function and analyze the influence of the choice of β on the co-clustering performance. For the synthetic data sets, we confirm that co-clustering outperforms one-sided clustering if the data matrix is noisy or if there is strong intra-cluster coupling. For the Newsgroup20 data set we observed that performance is insensitive to β as long as the number of word clusters is sufficiently large. Performance drops for few word clusters, a fact for which

- C. Blöchl is with the Rohde & Schwarz GmbH & Co. KG, Munich 81671, Germany. E-mail: clemens.bloechl@web.de.
- R. Ali Amjad is with the Institute for Communications Engineering, Technical University of Munich, Munich 80333, Germany. E-mail: ranaali.amjad@tum.de.
- B.C. Geiger is with KNOW-CENTER Gmbh, Graz A-8010, Austria. E-mail: geiger@ieee.org.

Manuscript received 29 Dec. 2017; revised 30 May 2018; accepted 31 May 2018. Date of publication 0 . 0000; date of current version 0 . 0000.
(Corresponding author: Bernhard C. Geiger.)
Recommended for acceptance by E. Keogh.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2018.2846252

we provide a theoretical explanation. The parameter β has a somewhat stronger influence on the performance on the MovieLens100k data set, for which we obtained movie clusters largely consistent with genres. Finally, for the Southern Women Event Participation Dataset, our results are remarkably similar to the reference co-clusterings from [12], [13].

In summary, our contribution is threefold:

(1) We provide a generalized framework for information-theoretic co-clustering via connecting it with Markov aggregation. The cost function, parameterized with a single parameter and connected with the information bottleneck formalism, is justified by well-defined operational goals of the Markov aggregation problem (Sections 3 an 4).
(2) Our generalized framework contains previously proposed information-theoretic cost functions as special cases (Section 5). Since the parameter of our cost function has an intuitive meaning, our framework leads to a deeper understanding of the previously proposed approaches. This understanding is further developed by pointing at the strengths and limitations of information-theoretic cost functions for co-clustering with the help of examples and experiments on synthetic datasets (Section 6). We also discuss the influence of the single parameter on the co-clustering results and present general guidelines for setting this parameter depending on the characteristics of the dataset.
(3) We perform experiments (Section 7) with real-world datasets. Varying the parameter allows us to compare our results to those obtained via cost functions previously proposed in the literature.

We do not address the important issues of choosing the number of clusters, nor do we design sophisticated optimization heuristics and/or initialization procedures; essentially, most heuristics proposed for previous cost functions such as in [2], [11] can be adapted to our framework.

The fact that our cost function contains previously proposed cost functions as special cases allows us to compare them *fairly*, i.e., with the same initialization steps and the same optimization heuristic. For example, the insensitivity to β in our experiments with the Newsgroup20 datasets provides a new perspective on the differences reported in [1], [2], [3], [4], suggesting that they are due to differences in optimization heuristics, preprocessing steps, or choice of data subsets rather than due to differences in the cost function.

Notation. Random variables (RVs) are denoted by upper case letters (Z), lower case letters (z) are reserved for realizations and constants, and calligraphic letters ($\mathcal{Z}$) are used for sets. We use bold upper case letters ($\mathbf{Z}$) to denote matrices. We assume that the reader is familiar with information-theoretic quantities. Specifically, the *mutual information* between two RVs Z and S with finite alphabet and joint distribution $P_{Z,S}$ is denoted as $I(Z;S)$ [14, eq. (2.28)]. Note further that $I(Z;S) = H(S) - H(S|Z)$, where $H(S)$ is the entropy of S and where $H(S|Z)$ is the conditional entropy of S given Z.

2 Related Work

2.1 Information-Theoretic Co-Clustering Approaches

Information-theoretic approaches to co-clustering require a probability distribution over the sets to be clustered, which we will denote as $\mathcal{X}$ and $\mathcal{Y}$. For example, if the data matrix $\mathbf{W}$ is nonnegative, then one can normalize it such that its entries sum to one. One can thus define RVs X and Y over the sets $\mathcal{X}$ and $\mathcal{Y}$ that have a joint distribution $P_{X,Y} \propto \mathbf{W}$.

One-sided clustering, i.e., clustering only the RV X with a clustering function Φ such that information about Y is preserved, was one of the main motivations behind the information bottleneck (IB) method [15]. Several algorithmic approaches have been proposed, including agglomerative [16] and sequential [11] methods and a method reminiscent of k-means [17] (the latter being equivalent to the fixed-point iterations in the original paper [15]).

An early information-theoretic approach to co-clustering was proposed by Slonim and Tishby [1] and is based on the IB method [15]. There, the authors proposed first finding the clustering function Φ maximizing $I(\Phi(X);Y)$, and then, after fixing Φ, finding the clustering function Ψ that maximizes $I(\Phi(X);\Psi(Y))$. Their approach was improved later by El-Yaniv and Souroujon, who suggested iterating this procedure multiple times [5]. Also based on the IB method is the work of Wang et al. [3]. They used a multivariate extension of mutual information to compress "input information"—captured by the mutual information terms $I(X;Y)$, $I(X;\Phi(X))$, and $I(Y;\Psi(Y))$—while preserving relevant information—captured by the information shared between the clusters, $I(\Phi(X);\Psi(Y))$, and the predictive power of the clusters, $I(\Phi(X);Y)$ and $I(X;\Psi(Y))$.

In 2003, Dhillon et al. proposed a co-clustering algorithm simultaneously determining clustering functions Φ and Ψ with the goal to maximize $I(\Phi(X);\Psi(Y))$ [2]. They showed that the problem is equivalent to a constrained nonnegative matrix tri-factorization problem [2, Lemma 2.1] with Kullback-Leibler divergence as cost function. (An iterative update rule for the entries of the three matrices is provided in [10, Sec. 4.4].) The work in [2] was generalized into various directions. On the one hand, Bekkerman et al. investigated simultaneous clustering of more than two sets in [4]. Rather than maximizing one of the multivariate extension of mutual information, the authors suggested maximizing the sum of mutual information terms between pairs of clusters; the pairs of clusters considered in the sum are determined by an undirected graph that has to be provided by the user. On the other hand, Banerjee et al. viewed co-clustering as a matrix approximation problem [18], of which the nonnegative matrix tri-factorization problem of [2, Lemma 2.1] is a special case. Their generalized framework admits any Bregman divergence (e.g., Kullback-Leibler divergence or squared euclidean distance) as cost function and several co-clustering schemes characterized by the type of summary statistic used to approximate the matrix.

Finally, Laclau et al. formulate the co-clustering problem as an optimal transport problem with entropic regularization [19]. Their formulation also turns into a probability matrix approximation problem with Kullback-Leibler divergence as cost function, but 1) the order of original and approximate distribution is swapped compared to [2, Lemma 2.1], and 2) the approximate distribution is obtained differently. They proposed solving the co-clustering problem with the Sinkhorn-Knopp algorithm and suggested a heuristic to determine the number of clusters.

2.2 Markov Aggregation and Lumpability

Markov aggregation is the task of replacing a Markov chain $\{Z_t\colon t=1,2,\ldots\}$ with a alphabet $\mathcal{Z}$ by a Markov chain with a smaller alphabet $\overline{\mathcal{Z}}$, sacrificing model accuracy for a reduction in model complexity. Aggregation is usually performed by partitioning (i.e., clustering) the alphabet $\mathcal{Z}$ and defining a Markov chain on the partitioned alphabet $\overline{\mathcal{Z}}$. Information-theoretic cost functions for Markov aggregation had been proposed in, e.g., [20], [21], [22] and were recently unified in [23]. More generally, aggregations of dynamical systems that are not necessarily Markov were discussed in [24]. In contrast to [20], [21], [22], [23], the cost functions proposed by [24] are task-specific in the sense that they aim to predict an observation based on Z_t from the aggregated process.

Closely related to Markov aggregation is the topic of *lumpability*, i.e., the question whether a non-injective function of a Markov chain is Markov. Initial research in this area has performed by Kemeny and Snell (strong and weak lumpability, [25, Sections 6.3-6.4]), Rosenblatt (lumpability of continuous-valued Markov processes [26]), and Buchholz (exact lumpability [27]). Gurvits and Ledoux discovered linear-algebraic conditions on the transition probability matrix of $\{Z_t\colon t=1,2,\ldots\}$ and the aggregation function for weak and strong lumpability [28]. An equivalent characterization of strong lumpability in information-theoretic terms has been presented by Geiger and Temmel and Pfante et al. in [29] and [30], respectively. This information-theoretic characterization was used in a cost function for Markov aggregation in [21].

3 Generalized Information-Theoretic Markov Aggregation

Suppose $\{Z_t\colon t=1,2,\ldots\}$ is a discrete-time, first-order, stationary Markov chain with finite alphabet $\mathcal{Z}$ and state transition matrix $\mathbf{A}=[A_{ij}]$, where

$$\forall i,j\in\mathcal{Z}, t>1\colon\quad A_{ij}:=\Pr(Z_t=j|Z_{t-1}=i). \tag{1}$$

Throughout this work we assume that $\mathbf{A}$ is irreducible. The Markov aggregation problem is concerned with finding a function $\zeta\colon\mathcal{Z}\to\overline{\mathcal{Z}}$, where typically $|\mathcal{Z}|\gg|\overline{\mathcal{Z}}|$, such that the reduced model captures *relevant* aspects of the original model. Specifically, the authors of [23] suggest trading between two different objectives: The objective to make the process $\{\zeta(Z_t)\}$ as close to a Markov chain as possible, and the objective that $\{\zeta(Z_t)\}$ preserves the temporal dependence structure of the original Markov chain $\{Z_t\}$. They propose the following information-theoretic cost function for Markov aggregation:

Definition 1 (Generalized Markov Aggregation [23]). *Let $\{Z_t\}$ be a discrete-time, stationary Markov chain with alphabet $\mathcal{Z}$ and state transition matrix $\mathbf{A}$, and suppose the set $\overline{\mathcal{Z}}$ is given. Let $\beta\in[0,1]$. The generalized information-theoretic Markov aggregation problem concerns finding a minimizer $\hat{\zeta}$ of*

$$\min_{\zeta\colon\mathcal{Z}\to\overline{\mathcal{Z}}}\mathcal{L}_\beta(\zeta), \tag{2}$$

where the minimization is over all functions $\zeta\colon\mathcal{Z}\to\overline{\mathcal{Z}}$ and where, with $\overline{Z}_t:=\zeta(Z_t)$ for every $t\geq 1$,

$$\begin{aligned}\mathcal{L}_\beta(\zeta):=\beta I(Z_1;Z_2)+(1-2\beta)I(Z_1;\overline{Z}_2)\\ -(1-\beta)I(\overline{Z}_1;\overline{Z}_2).\end{aligned} \tag{3}$$

For $\beta=1$, the cost function is reminiscent of the IB functional [15], where compression is enforced by limiting the alphabet size of the compressed variable. For $\beta=0$, the cost function is linked to the phenomenon of lumpability and ζ is chosen such that the process $\{\overline{Z}_t\}$ is "as Markov as possible"; indeed, if $\mathcal{L}_0(\zeta)=0$, then $\{\overline{Z}_t\}$ is a Markov chain [21, Th. 1]. Finally, it can be shown that minimizing $\mathcal{L}_{\frac{1}{2}}(\zeta)$ is equivalent to maximizing $I(\overline{Z}_1;\overline{Z}_2)$; essentially, this means that one wants to predict $\overline{Z}_2$ from $\overline{Z}_1$ with high accuracy, i.e., the temporal dependence structure should be preserved. This cost function was considered in [20] and was shown to be related to spectral clustering.

In the spirit of the IB formalism, mutual information can be used to measure relevance. Relevant information loss measures the information about some relevant RV S that is lost by processing a statistically related RV Z in a deterministic function ζ. The quantity was introduced by Plumbley in the context of unsupervised neural networks [31]:

Definition 2 (Relevant Information Loss). *Let S and Z be RVs with finite alphabet, and let ζ be a function defined on the alphabet $\mathcal{Z}$ of Z. Then, the relevant information loss w.r.t. S that is induced by ζ is*

$$L_S(Z\to\zeta(Z)):=I(S;Z)-I(S;\zeta(Z))=I(S;Z|\zeta(Z))\geq 0. \tag{4}$$

With this definition, we can rewrite the cost function for Markov aggregation in terms of relevant information loss:

Lemma 1. *In the setting of Definition 1 we have*

$$\mathcal{L}_\beta(\zeta)=\beta L_{Z_1}(Z_2\to\overline{Z}_2)+(1-\beta)L_{\overline{Z}_2}(Z_1\to\overline{Z}_1). \tag{5}$$

The function ζ partitions the alphabet $\mathcal{Z}$ into clusters. Hence, the first term captures how much information is lost about Z_1 if Z_2 is clustered via ζ, while the second term captures how much information is lost about the *cluster* $\overline{Z}_2$ if Z_1 is clustered via ζ. This formulation will be our starting point for developing an information-theoretic cost function for co-clustering.

4 Information-Theoretic Co-Clustering via Markov Aggregation

We now turn to the co-clustering problem. Suppose we have two disjoint finite sets $\mathcal{X}$ and $\mathcal{Y}$ and a $|\mathcal{X}|\times|\mathcal{Y}|$ matrix $\mathbf{W}$ containing, e.g., similarities, the number of co-occurrences, or correlations between elements of these two sets. As an example, if $\mathcal{X}$ is a set of documents and $\mathcal{Y}$ a set of words, then the (i,j)th entry of $\mathbf{W}$ could be the number of times the word j appeared in document i. Co-clustering is concerned with finding partitions of $\mathcal{X}$ and $\mathcal{Y}$ (document and word clusters in this example), sacrificing information about the individual data elements to make the group characteristics more prominent and accessible.

4.1 Adapting the Cost Function

If the matrix $\mathbf{W}$ is nonnegative, we can interpret it as the weight matrix of an undirected, weighted, bipartite graph, cf. [6]. Throughout this work we will assume that $\mathbf{W}$ is such that the bipartite graph is irreducible. On this graph, one can then define a simple random walk, i.e., a Markov chain $\{Z_t\}$ with alphabet $\mathcal{X}\cup\mathcal{Y}$ and state transition matrix

$$\mathbf{A} = \mathbf{D}^{-1}\begin{bmatrix} 0 & \mathbf{W} \\ \mathbf{W}^T & 0 \end{bmatrix}, \tag{6}$$

where $\mathbf{D}$ is a diagonal matrix collecting sums of all connected edge weights of respective nodes. The matrix $\mathbf{D}$ normalizes each row of $\mathbf{A}$ to make it a probability distribution. Since the graph is bipartite and undirected, the Markov chain $\{Z_t\}$ is 2-periodic and reversible.

We now apply the Markov aggregation framework from Definition 1 and Lemma 1 to the co-clustering problem. To this end, we add the constraint that the function ζ from Definition 1 does not put elements of $\mathcal{X}$ and $\mathcal{Y}$ in the same cluster. This mutual exclusivity constraint guarantees that there exist functions Φ and Ψ such that

$$\forall i\in\mathcal{X}\cup\mathcal{Y}:\quad \zeta(i) = \begin{cases} \Phi(i), & i\in\mathcal{X} \\ \Psi(i), & i\in\mathcal{Y}. \end{cases} \tag{7}$$

The following proposition transfers the cost function from Lemma 1 to the co-clustering setting:

Proposition 1. *Suppose two disjoint finite sets $\mathcal{X}$ and $\mathcal{Y}$ and a nonnegative $|\mathcal{X}|\times|\mathcal{Y}|$ matrix $\mathbf{W}$ containing similarities between elements of these two sets are given. Define two discrete RVs X and Y over these sets, where the joint distribution $P_{X,Y}$ is obtained by normalizing $\mathbf{W}$. Let $\{Z_t\}$ be a stationary Markov chain with alphabet $\mathcal{X}\cup\mathcal{Y}$ and state transition matrix $\mathbf{A}$ given in (6). Let $\beta\in[0,1]$ and suppose the sets $\overline{\mathcal{X}}$ and $\overline{\mathcal{Y}}$ are given.*

For every function $\zeta\colon \mathcal{X}\cup\mathcal{Y}\to\overline{\mathcal{X}}\cup\overline{\mathcal{Y}}$ satisfying the mutual exclusivity constraint (7), we have

$$\begin{aligned} 2\cdot\mathcal{L}_\beta(\zeta) &= \beta(L_X(Y\to\overline{Y}) + L_Y(X\to\overline{X})) \\ &\quad + (1-\beta)(L_{\overline{X}}(Y\to\overline{Y}) + L_{\overline{Y}}(X\to\overline{X})) =: \mathcal{L}_\beta(\Phi,\Psi) \end{aligned} \tag{8}$$

where $\overline{X} := \Phi(X)$ and $\overline{Y} := \Psi(Y)$.

Proof. Suppose that $\{Z_t\}$ is a Markov chain with state space $\mathcal{X}\cup\mathcal{Y}$ and state transition matrix $\mathbf{A}$ as in (6), with $\mathbf{D}$ given by

$$\mathbf{D} = \operatorname{diag}\left(\begin{bmatrix} 0 & \mathbf{W} \\ \mathbf{W}^T & 0 \end{bmatrix}\mathbf{1}\right), \tag{9}$$

where $\mathbf{1}$ is a vector of ones of appropriate length. Suppose $\boldsymbol{\mu} = [\mu_i]$ is the invariant distribution of $\mathbf{A}$, i.e., $\boldsymbol{\mu}^T = \boldsymbol{\mu}^T\mathbf{A}$. It follows that $\operatorname{diag}(\boldsymbol{\mu})\propto\mathbf{D}$. Suppose further that $P_{X,Y}$ is the joint distribution obtained by normalizing $\mathbf{W}$. Then, the marginal distributions for X and Y are $P_X = \sum_{y\in\mathcal{Y}} P_{X,Y}(\cdot,y)\propto\mathbf{W1}$ and $P_Y^T = \sum_{x\in\mathcal{X}} P_{X,Y}(x,\cdot)\propto \mathbf{1}^T\mathbf{W}$, respectively. From the 2-periodicity of $\{Z_t\}$ thus follows that

$$\mu_i = \frac{1}{2}\begin{cases} P_X(i), & i\in\mathcal{X} \\ P_Y(i), & i\in\mathcal{Y}. \end{cases} \tag{10}$$

Now assume that the Markov chain $\{Z_t\}$ is stationary, i.e., the distribution of Z_1 coincides with the invariant distribution $\boldsymbol{\mu}$. Let U be a RV that indicates whether Z_1 was drawn from $\mathcal{X}$ or $\mathcal{Y}$, i.e.,

$$U := \begin{cases} 1, & Z_1\in\mathcal{X} \\ 0, & Z_1\in\mathcal{Y}. \end{cases} \tag{11}$$

Note that U is a function not only of Z_1 but, by periodicity, of Z_t for every t. The RV U thus connects P_{Z_t} with P_X or P_Y; e.g., if $U=1$, then $P_{Z_3} = P_X$. It follows from (10) that $\Pr(U=1) = \Pr(U=0) = \frac{1}{2}$.

Finally, suppose that ζ satisfies the mutual exclusivity constraint (7); hence $\Phi(\mathcal{X}) = \overline{\mathcal{X}}$, $\Psi(\mathcal{Y}) = \overline{\mathcal{Y}}$, and $U=1$ if and only if $\overline{Z}_1\in\overline{\mathcal{X}}$.

We now investigate $I(\tilde{Z}_1;\tilde{Z}_2)$, where $\tilde{Z}_i$ is either Z_i or $\overline{Z}_i$. We get

$$\begin{aligned} &I(\tilde{Z}_1;\tilde{Z}_2) \\ &\overset{(a)}{=} I(\tilde{Z}_1,U;\tilde{Z}_2) \\ &\overset{(b)}{=} I(\tilde{Z}_1;\tilde{Z}_2|U) + I(U;\tilde{Z}_2) \\ &\overset{(c)}{=} \frac{1}{2}I(\tilde{Z}_1;\tilde{Z}_2|U=1) + \frac{1}{2}I(\tilde{Z}_1;\tilde{Z}_2|U=0) + H(U), \end{aligned} \tag{12}$$

where (a) is because U is a function of Z_1 and $\overline{Z}_1$, (b) is the chain rule of mutual information, and (c) follows because U is also a function of Z_2 and $\overline{Z}_2$ and from the definition of conditional mutual information.

Now suppose $\tilde{Z}_1 = \overline{Z}_1$ and $\tilde{Z}_2 = Z_2$. If $U=1$, then $\overline{Z}_1\in\overline{\mathcal{X}}$ and $Z_2\in\mathcal{Y}$, and the joint distribution $P_{\overline{Z}_1,Z_2}$ equals the joint distribution $P_{\overline{X},Y}$. With similar considerations for $U=0$ we hence get

$$\begin{aligned} I(\overline{Z}_1;Z_2) &= \frac{1}{2}I(\overline{Z}_1;Z_2|U=1) + \frac{1}{2}I(\overline{Z}_1;Z_2|U=0) + H(U) \\ &= \frac{1}{2}I(\overline{X};Y) + \frac{1}{2}I(X;\overline{Y}) + H(U). \end{aligned} \tag{13a}$$

Along the same lines we obtain

$$I(Z_1;Z_2) = I(X;Y) + H(U), \tag{13b}$$

$$I(\overline{Z}_1;\overline{Z}_2) = I(\overline{X};\overline{Y}) + H(U), \tag{13c}$$

$$I(Z_1;\overline{Z}_2) = \frac{1}{2}I(X;\overline{Y}) + \frac{1}{2}I(\overline{X};Y) + H(U). \tag{13d}$$

Inserting these in the cost function in Lemma 1 and applying the definition of relevant information loss in Definition 2 completes the proof. □

We now present our cost function for information-theoretic co-clustering:

Definition 3 (Generalized Information-Theoretic Co-Clustering). *The generalized information-theoretic co-clustering problem concerns finding a minimizer $(\hat{\Phi},\hat{\Psi})$ of*

$$\min_{\Phi\colon \mathcal{X}\to\overline{\mathcal{X}},\ \Psi\colon \mathcal{Y}\to\overline{\mathcal{Y}}} \mathcal{L}_\beta(\Phi,\Psi), \tag{14}$$

where the minimization is over all functions $\Phi\colon \mathcal{X}\to\overline{\mathcal{X}}$ *and* $\Psi\colon \mathcal{Y}\to\overline{\mathcal{Y}}$ *and where* $\mathcal{L}_\beta(\Phi,\Psi)$ *is as in the setting of Proposition 1.*

The presented cost function admits an intuitive explanation for the effect of the parameter β: In the context of the words/documents co-clustering example above, minimizing $L_X(Y\to\overline{Y})$ means that we are looking for word clusters that tell us much about documents. In contrast, minimizing $L_{\overline{X}}(Y\to\overline{Y})$ means that we are looking for word and document clusters such that the word clusters tell us much about the document clusters. The parameter β thus determines how strongly the two clusterings should be coupled. We show in Sections 6 and 7 that the choice of β can have a prominent effect on the clustering performance.

4.2 Adapting a Sequential Optimization Heuristic

In general, finding a minimizer of our cost function (14) is a combinatorial problem with exponential computational complexity in $|\mathcal{X}|$ and $|\mathcal{Y}|$. Hence heuristics for combinatorial or non-convex optimization are used to find good sub-optimal solutions with reasonable complexity. In particular, it can be optimized by adapting heuristics proposed for information-theoretic co-clustering by other authors (see Sections 2 and 5). Since our cost function is derived from the generalized information-theoretic Markov aggregation problem, co-clustering solutions can be obtained by employing the aggregation algorithm proposed in [23] taking into account the additional mutual exclusivity constraint. The algorithm is a simple sequential heuristic for minimizing $\mathcal{L}_\beta$, similar to the sequential IB algorithm proposed in [11] and the algorithm proposed by Dhillon et al. for information-theoretic co-clustering [2]. This algorithm is random in the sense that it is started with two random functions Φ and Ψ with desired output cardinalities. In each iteration, these two functions are altered successively in order to reduce the cost function, either until we reach a maximum number of iterations or until the cost function has converged to within a chosen threshold of a local minimum. The authors of [23] introduced an annealing procedure for the β-parameter to escape local optima, which is particularly important for small values of β. The pseudocodes for the sequential heuristic, sGITCC, and the annealing heuristic, AnnITCC, are given in Algorithms 1 and 2, respectively; for details, the reader is referred to [23]. It can be shown along the lines of the corresponding result in [23] that, by storing intermediate results, the computational complexity of computing $\mathcal{L}_\beta(\Phi,\Psi_\ell)$ and $\mathcal{L}_\beta(\Phi_j,\Psi)$ can be brought down to $\mathcal{O}(|\mathcal{X}|)$ and $\mathcal{O}(|\mathcal{Y}|)$, respectively. Thus, one iteration of Algorithm 1 has computational complexity of $\mathcal{O}(|\mathcal{X}|\cdot|\mathcal{Y}|\cdot\max\{|\overline{\mathcal{Y}}|,|\overline{\mathcal{X}}|\})$.

The following example shows how the sequential heuristic in Algorithm 1 can get stuck in a poor local optimum for $\beta=\frac{1}{2}$. The same example is unproblematic for $\beta=1$. Since one can certainly find heuristics that perform optimally in this example even for $\beta=\frac{1}{2}$, matching the heuristic to the cost function seems to be an important issue. We will see further evidence for the impact of heuristics on performance in our experiments with the Newsgroup20 dataset in Section 7.1.

Example 1. Consider the following 3×4 matrix describing the joint probability distribution between X and Y:We are

$$P_{X,Y} = \left[\begin{array}{c|c|cc} 0.25 & 0 & 0 & 0 \\ \hline 0 & 0.25 & 0 & 0 \\ \hline 0 & 0 & 0.25 & 0.25 \end{array}\right].$$

interested in two row clusters and two column clusters, i.e., $|\overline{\mathcal{X}}|=|\overline{\mathcal{Y}}|=2$. Suppose that during some iteration, the clustering functions Φ and Ψ induce the partition indicated by the thin black lines in the matrix $P_{X,Y}$. At this stage, for $\beta=\frac{1}{2}$ the sequential algorithm will terminate since this Φ is the optimal choice for Ψ fixed, and this Ψ is the optimal choice for Φ fixed. In other words, changing either clustering function alone increases the cost $\mathcal{L}_{\frac{1}{2}}=I(X;Y)-I(\overline{X};\overline{Y})$. Nevertheless, it is clear from looking at $P_{X,Y}$, that the cost is minimized ($I(\overline{X};\overline{Y})$ is maximized) for the partition indicated by the thick black lines. The algorithm thus gets stuck for $\beta=\frac{1}{2}$ because the cost function in this case only depends on the clustered variables, and because it updates the clustering functions subsequently rather than jointly. For larger values of β, the coupling between the clustering functions is weaker. In particular, for $\beta=1$, the clustering functions can be optimized independently of each other, and the algorithm hence terminates at a partition consistent with the vertical thick line, even if it was started at the partition indicated by the thin lines.

Algorithm 1. Sequential Generalized Information-Theoretic Co-Clustering (sGITCC)

function: $(\Phi,\Psi)=$ sGITCC($P_{X,Y}$, β, $|\overline{\mathcal{X}}|$, $|\overline{\mathcal{Y}}|$, #iter$_{\max}$, tol, optional: intial clustering $(\Phi_{\text{init}},\Psi_{\text{init}})$)
  if $(\Phi_{\text{init}},\Psi_{\text{init}})$ is empty **then** ▷ *Inizialization*
    $(\Phi,\Psi)\leftarrow$ Random Clustering
  else
    $(\Phi,\Psi)\leftarrow(\Phi_{\text{init}},\Psi_{\text{init}})$
  end if
  #iter $\leftarrow 0$
  while #iter $<$ #iter$_{\max}\wedge\delta>$ tol **do** ▷ *Main Loop*
    $C_{old}\leftarrow\mathcal{L}_\beta(\Phi,\Psi)$
    for all elements $i\in\mathcal{X}$ **do** ▷ *Optimizing* Φ
      for all clusters $j\in\overline{\mathcal{X}}$ **do**
        $\Phi_j(x)=\begin{cases}\Phi(x) & \forall x\neq i\\ j & x=i\end{cases}$
      end for
      $\Phi(i)=\arg\min_j\mathcal{L}_\beta(\Phi_j,\Psi)$
    end for
    for all elements $k\in\mathcal{Y}$ **do** ▷ *Optimizing* Ψ
      for all clusters $\ell\in\overline{\mathcal{Y}}$ **do**
        $\Psi_\ell(y)=\begin{cases}\Psi(y) & \forall y\neq k\\ \ell & y=k\end{cases}$
      end for
      $\Psi(k)=arg\,min_\ell\mathcal{L}_\beta(\Phi,\Psi_\ell)$ ▷ *Break ties*
    end for
    $\delta\leftarrow C_{old}-\mathcal{L}_\beta(\Phi,\Psi)$
    #iter $\leftarrow$ #iter $+1$
  end while
end function:

Algorithm 2. β-Annealing Information-Theoretic Co-Clustering (AnnITCC)

```
function: (Φ, Ψ) = AnnITCC(P_{X,Y}, β, |𝒳|, |𝒴|, #iter_max, tol, Δ)
    α ← 1
    (Φ, Ψ) = sGITCC(P_{XY}, β, |𝒳|, |𝒴|, #iter_max, tol)
    while α > β do
        α ← max{α − Δ, β}
        (Φ, Ψ) = sGITCC(P_{XY}, α, |𝒳|, |𝒴|, #iter_max, tol, (Φ, Ψ))
    end while
end function:
```

5 Special Cases of Generalized Information-Theoretic Co-Clustering

We next show that our generalized information-theoretic co-clustering cost function from Definition 3 contains, for appropriate settings of the parameter β, previously proposed cost functions as special cases. For example, for $\beta = 1$, we obtain

$$\mathcal{L}_1(\Phi,\Psi) = L_X(Y \to \overline{Y}) + L_Y(X \to \overline{X}). \tag{15}$$

This cost function consists of two IB functionals: The first term considers clustering Y with X the relevant variable, while the second term considers clustering X with Y the relevant variable. This approach rewards clustering solutions for X and Y that are completely decoupled. To minimize this cost function, one can use the fixed-point equations derived in [15] or the agglomerative IB method (aIB) that merges clusters until the desired cardinality is reached [16]. Finally, a sequential IB method (sIB) has been proposed that iteratively moves an element from its current cluster to the cluster that minimizes the cost until a local minimum is reached [11].

More interestingly, we can rewrite the cost function that Dhillon et al. proposed in [2] for information-theoretic co-clustering (ITCC) and obtain

$$\mathcal{L}_{\mathrm{ITCC}}(\Phi,\Psi) := I(X;Y) - I(\overline{X};\overline{Y}) = \mathcal{L}_{\frac{1}{2}}(\Phi,\Psi). \tag{16}$$

Thus, ITCC is a special case of our cost function for $\beta = \frac{1}{2}$. The authors of [2] proposed a sequential algorithm, similar to sIB, alternating between optimizing Φ and Ψ. Furthermore, $\mathcal{L}_{\mathrm{ITCC}}(\Phi,\Psi)$ can be optimized via non-negative matrix tri-factorization [2, Lemma 2.1] and thus yields a generative model as a result. We are not aware if a similar connection to generative models holds for other values of β.

In [4], the cost function $\mathcal{L}_{\frac{1}{2}}$ is generalized to pairwise interactions of multiple variables (the two-dimensional case is equivalent to co-clustering). The authors introduce a multilevel heuristic that schedules the splitting of clusters, merges clusters following the ideas of aIB [1], and optimizes intermediate results sequentially with sIB.

The authors of [1] proposed applying aIB twice to obtain the co-clustering. In the first step, in which the set $\mathcal{X}$ is clustered, they treat Y as the relevant variable; in the second step, in which the set $\mathcal{Y}$ is clustered, they treat the clustered variable $\overline{X}$ as relevant. In essence, the authors of [1] thus minimize the functional

$$\mathcal{L}_{\mathrm{IB-double}}(\Phi,\Psi) = L_Y(X \to \overline{X}) + L_{\overline{X}}(Y \to \overline{Y}) = \mathcal{L}_{\frac{1}{2}}(\Phi,\Psi), \tag{17}$$

in a greedy manner: They first optimize over Φ to minimize only the first term and then optimize over Ψ to minimize the second term. Comparing (16) and (17) reveals that [1] and [2] optimize the same cost function; the fact that they report different performance results can only be explained by differences in the optimization heuristic and (possibly) preprocessing steps. We will elaborate on this topic in our experiments with the Newsgroup20 dataset in Section 7.1.

Another approach related to IB, called information bottleneck co-clustering (IBCC), was proposed in [3]. The functional being maximized by IBCC is

$$\begin{aligned}\mathcal{L}_{\mathrm{IBCC}}(\Phi,\Psi) :&= I(X;\overline{Y}) + I(\overline{X};Y) + I(\overline{X};\overline{Y})\\ &= 3I(X;Y) - 2\mathcal{L}_{\frac{3}{4}}(\Phi,\Psi).\end{aligned} \tag{18}$$

Hence, also IBCC is a special case of the generalized Markov aggregation framework for $\beta = \frac{3}{4}$. The authors of [3] propose two algorithms: One is an agglomerative, i.e., a greedy merging algorithm, the other is an iterative update of fixed-point equations in the spirit of [15].

Finally, for $\beta = 0$ we obtain the functional

$$\mathcal{L}_0(\Phi,\Psi) = L_{\overline{X}}(Y \to \overline{Y}) + L_{\overline{Y}}(X \to \overline{X}). \tag{19}$$

As previously mentioned, for Markov aggregation and $\beta = 0$ the cost function is linked to the phenomenon of lumpability. In the co-clustering framework, lumpability means that the two clustering solutions that are coupled. Precisely, we have $\mathcal{L}_0(\Phi,\Psi) = 0$ if the rows X and columns Y do not share more information with the column clusters $\overline{Y}$ and row clusters $\overline{X}$, respectively, than the row clusters and column clusters share with each other. Unfortunately, we also have $\mathcal{L}_0(\Phi,\Psi) = 0$ if $\overline{X}$ and $\overline{Y}$ are independent, which suggests an inherent drawback of $\mathcal{L}_0$ for co-clustering (despite its justification in Markov aggregation [21]). This leads to $\mathcal{L}_0$ (and, in general, $\mathcal{L}_\beta$ for small β) having multiple bad local optima in which any heuristic tends to get stuck.

6 Strengths and Limitations of Generalized Information-Theoretic Co-Clustering

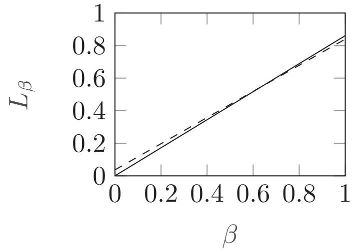
In this section we use examples and experiments on synthetic datasets to highlight different aspects of using $\mathcal{L}_\beta$ and our proposed optimization heuristic for co-clustering. Specifically, we will point at limitations and strengths of co-clustering in comparison with one-sided clustering ($\beta = 1$), which leads to guiding principles for the choice of β depending on characteristics of the considered dataset.

6.1 Examples

In the previous section we have discovered an inherent shortcoming of $\mathcal{L}_0$ in that it leads to co-clusterings with (near-)independent cluster RVs. In this section, we point at further limitations of information-theoretic cost functions for co-clustering. These shortcomings are independent of the employed optimization heuristic, but rather reflect that in some scenarios not even the global optimum of the cost function coincides with the ground truth (or an otherwise desired co-clustering solution). Sometimes this is simply caused by the fact that the cost function does not fit the underlying model—e.g., if $\mathbf{W}$ is generated according to a

$$\left[\begin{array}{c|c:c} 0.12 & 0 & 0 \\ \hline 0 & 0.39 & 0.05 \\ \hdashline 0 & 0.05 & 0.39 \end{array}\right]$$

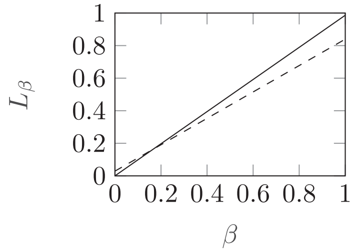
(a)



(b)

$$\left[\begin{array}{c|c:c} 0.12 & 0 & 0 \\ \hline 0 & 0.4 & 0.04 \\ \hdashline 0 & 0.04 & 0.4 \end{array}\right]$$

(c)



(d)

Fig. 1. Trading entropy for conditional entropy. (a) and (c) show joint distributions $P_{X,Y}$ together with two possible co-clusterings, while (b) and (d) show the corresponding values of the cost function for different values of β. Solid and dashed curves in (b) and (d) correspond to co-clusterings indicated by dashed and solid lines in (a) and (c).

Poisson latent block model, then maximizing the likelihood of the co-clustering is equivalent to minimizing $\mathcal{L}_{\frac{1}{2}}$ only if the clusters have all the same cardinality [9, Sec. 2.2]. In contrast, the following two scenarios make no assumptions on an underlying model but illustrate shortcomings inherent to the considered information-theoretic cost functions.

6.1.1 Largely Different $|\overline{\mathcal{X}}|$ and $|\overline{\mathcal{Y}}|$

An advantage of information-theoretic co-clustering approaches over, e.g., spectral [6], [8] or certain block model-based approaches [9] is that the former admit different cardinalities for the clustered sets $|\overline{\mathcal{X}}|$ and $|\overline{\mathcal{Y}}|$. If, however, these cardinalities differ greatly, then minimizing $\mathcal{L}_\beta$ becomes problematic especially for small values of β. Let us assume w.l.o.g. that $|\overline{\mathcal{Y}}| < |\overline{\mathcal{X}}|$. Then, the optimization term $L_{\overline{Y}}(X \to \overline{X})$ is limited by the information contained in $\overline{Y}$ rather than by the information loss induced by clustering X to $\overline{X}$; many functions Φ may bring $L_{\overline{Y}}(X \to \overline{X})$ close to zero simply because $\overline{Y}$ itself already contains little information. Similarly, the term $L_{\overline{X}}(Y \to \overline{Y})$ may be large for many choices of Φ, because, again, the limiting factor is the coarse clustering from Y to $\overline{Y}$. These terms get more importance in (14) if β is small. In other words, coupled co-clustering fails because the clustered variables contain little information. We illustrate this with a particular example, in which the joint probability distribution between X and Y is

$$P_{X,Y} = \left[\begin{array}{cc|cc} 0.125 & 0 & 0 & 0 \\ \hline 0.125 & 0 & 0 & 0 \\ \hline 0 & 0.125 & 0 & 0 \\ \hline 0 & 0.125 & 0 & 0 \\ \hline\hline 0 & 0 & 0.125 & 0 \\ 0 & 0 & 0.125 & 0 \\ \hline 0 & 0 & 0 & 0.125 \\ \hline 0 & 0 & 0 & 0.125 \end{array}\right].$$

Our aim is to obtain a co-clustering with $|\overline{\mathcal{Y}}| = 2$ and $|\overline{\mathcal{X}}| = 4$. In $P_{X,Y}$, the thick vertical line indicates one possibility for Ψ (a plausible ground truth). The horizontal lines indicate two possible options, Φ_1 (thick lines) and Φ_2 (thin lines) for the row clustering, where Φ_1 corresponds to a plausible ground truth.

For $\beta = 1$, (Φ_1, Ψ) has a lower cost than (Φ_2, Ψ), as desired. Furthermore, one can show that (Φ_1, Ψ) minimizes the cost function; $\mathcal{L}_1$ has its global minimum at the ground truth. For $\beta = \frac{1}{2}$, by evaluating $I(\overline{X}; \overline{Y})$ we see that both (Φ_1, Ψ) and (Φ_2, Ψ) have the same cost. In fact, any row clustering function Φ that shares the cluster boundary with the thick horizontal line in the middle has the same $I(\overline{X}; \overline{Y})$ for the given column clustering function Ψ: In this case, $\overline{X}$ determines $\overline{Y}$, hence we achieve the maximum $I(\overline{X}; \overline{Y}) = H(\overline{Y}) = 1$; the cost function has multiple global minima, only one of which lies at the ground truth. Finally, for $\beta = 0$, (Φ_1, Ψ) has a higher cost than (Φ_2, Ψ). This implies that even if we initialize our algorithm at the ground truth (this could be the case if we do β-annealing) we move away from this clustering solution when we optimize the cost function for smaller values of β.

6.1.2 Trading Entropy for Conditional Entropy

Consider the joint distribution in Fig. 1a that describes a dataset with a well-separated co-cluster structure for $|\overline{\mathcal{X}}| = |\overline{\mathcal{Y}}| = 2$ (based on zeros and indicated by solid lines, denoted by $(\Phi^\bullet, \Psi^\bullet)$). We evaluate our cost function for different values of β, both for $(\Phi^\bullet, \Psi^\bullet)$ and for an alternative co-clustering indicated by dashed lines, denoted by (Φ, Ψ). It can be seen in Fig. 1b that, for $\beta \in [0.65, 1]$, we have $\mathcal{L}_\beta(\Phi^\bullet, \Psi^\bullet) > \mathcal{L}_\beta(\Phi, \Psi)$, i.e., the "incorrect" solution has a lower cost than the ground truth. While in this case, e.g., ITCC [2] would probably terminate with $(\Phi^\bullet, \Psi^\bullet)$, it is easy to construct an example where ITCC fails. Changing our example only slightly leads to generalized information-theoretic co-clustering preferring (Φ, Ψ) over $(\Phi^\bullet, \Psi^\bullet)$ for all β in $[0.15, 1]$ (see Figs. 1c and 1d).

These examples show that even for datasets with a well-separated co-cluster structure, for a range of β there can be (local and global) minima having a lower cost $\mathcal{L}_\beta$ than the ground truth. This can be explained by the fact that optimizing the cost function for a given value of β boils down to maximizing/minimizing a combination of several mutual information terms. For example, for $\beta = \frac{1}{2}$ we aim to maximize, cf. (16)

$$I(\overline{X}; \overline{Y}) = H(\overline{X}) - H(\overline{X}|\overline{Y}). \quad (20)$$

This leads to two competing goals: entropy maximization (preferring clusters with roughly equal probabilities) and conditional entropy minimization (preferring row clusters that determine column clusters, and vice-versa). For the range of β where $\mathcal{L}_\beta(\Phi^\bullet, \Psi^\bullet)$ is not the global minimum, the first goal outweighs the second.

Note that for joint distributions with a well-separated co-cluster structure we have $\mathcal{L}_0(\Phi^\bullet, \Psi^\bullet) = 0$ since $I(X; \overline{Y}) = I(\overline{X}; Y) = I(\overline{X}; \overline{Y})$. Nevertheless, due to the shortcoming discussed in Section 5, this global optimum may not found because many other co-clusterings lead to $\mathcal{L}_0(\Phi, \Psi) \approx 0$.

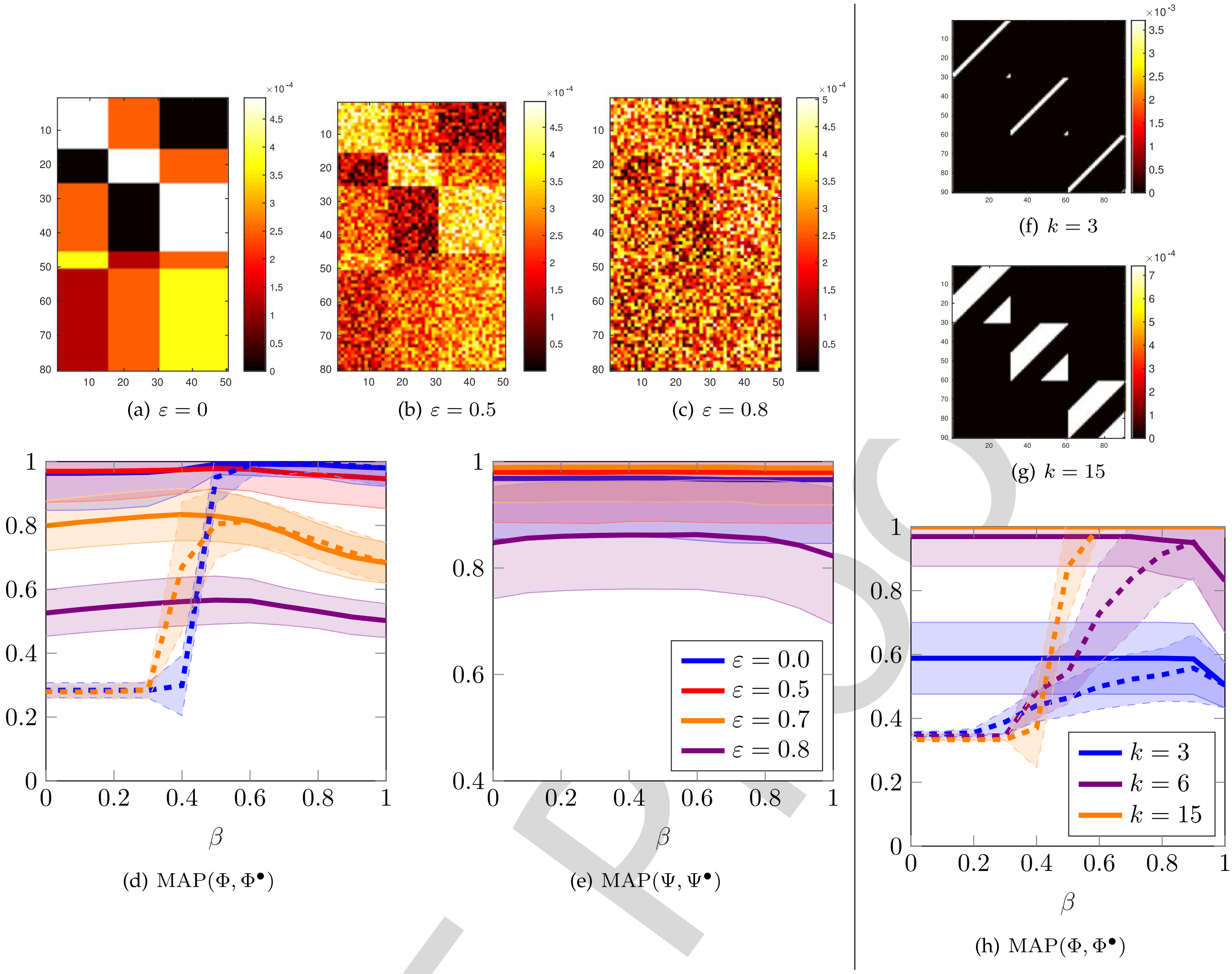


Fig. 2. (a)-(c) and (f)-(g) Colorplots of $P_{X,Y}$ for different noise levels ε and different parameters k. It can be seen that the true cluster structure becomes less obvious with increasing noise levels. (d), (e), and (h) Micro-averaged precision curves show the average over 500 random experiments (center line) and the standard deviation (shaded area). Solid curves correspond to AnnITCC, dashed curves to sGITCC. See text for details.

6.2 Synthetic Datasets

Next, we perform experiments with two different synthetic datasets to explore further the relation between suitable choices of β and the characteristics of the dataset. Since our focus is on providing a better understanding of information-theoretic co-clustering, we assume that the true numbers of clusters and the true clustering functions $\Phi^\bullet$ and $\Psi^\bullet$ are known. As an accuracy measure, we employ the micro-averaged precision, which we define as follows:

$$\mathrm{MAP}(\Phi, \Phi^\bullet) := \max_{\pi} \frac{\sum_{j \in \overline{\mathcal{X}}} |\Phi^{-1}(j) \cap \Phi^{\bullet -1}(\pi(j))|}{|\mathcal{X}|}, \tag{21}$$

where the maximization is over all permutations π of the set $\overline{\mathcal{X}}$. The micro-averaged precision $\mathrm{MAP}(\Psi, \Psi^\bullet)$ is computed along the same lines. Note that $\mathrm{MAP}(\cdot,\cdot)$ requires that the clustering solution found by the algorithm has the same number of clusters as are present in the ground truth. Since we assume the true number of clusters to be known, this is unproblematic. If the number of clusters is unknown, one can resort to more sophisticated measures such as the adjusted Rand index or normalized mutual information. In the present case, all of these measures will lead to similar qualitative results.

Unless noted otherwise, we set $\mathrm{tol} = 0$, $\#\mathrm{iter}_{\max} = 20$, and $\Delta = 0.1$ and ran AnnITCC for values of β between 0 and 1 in steps of 0.1. The simulation code for these and the real-world experiments in Section 7 is publicly accessible.[1]

The first experiment looks at the clustering performance in the presence of noise. We generated a joint probability distribution $T_{X,Y}$ with 80 rows and 50 columns, i.e., $|\mathcal{X}| = 80$ and $|\mathcal{Y}| = 50$, and planted co-clusters such that $T_{X,Y}$ is constant within each co-cluster. A colorplot of $T_{X,Y}$ is shown in Fig. 2a. The figure also shows the ground truth $\Phi^\bullet$ ($|\overline{\mathcal{X}}| = 5$) and $\Psi^\bullet$ ($|\overline{\mathcal{Y}}| = 3$). We moreover constructed a random probability distribution N and constructed $P_{X,Y}$ from a weighted average of $T_{X,Y}$ and N, i.e.,

$$P_{X,Y} = (1-\varepsilon) T_{X,Y} + \varepsilon N, \tag{22}$$

where $\varepsilon \in \{0, 0.5, 0.7, 0.8\}$. Colorplots of $P_{X,Y}$ are shown in Figs. 2b and 2c for $\varepsilon = 0.5$ and $\varepsilon = 0.8$, respectively.

1. bitbucket.org/bernhard_geiger/coclustering_markovaggregation

We repeated the whole procedure for 500 different probability matrices N. The MAP values, averaged over these 500 runs, are reported in Fig. 2d and 2e (solid lines). First of all, it can be seen that even in the noiseless case, the clusters are not always identified correctly. Since we identified the correct clusters in over 90 percent of the simulation runs, we believe that this effect can be explained by the algorithm getting stuck in a local optimum. Second, one can observe the natural effect that large noise levels lead to lower MAP values—interestingly, though, co-clustering seems to be quite robust to noise, as the MAP values in this experiment seem to decrease significantly only for $\varepsilon > 0.5$, i.e., when noise starts to dominate the data matrix. Finally, for large noise levels, it turns out that the intermediate values of β perform better. The performance drop for larger values of β is not due to the optimization heuristic getting stuck in bad local optima: We found that the cost of the co-clustering solution found by AnnITCC for large β is lower than the cost of the ground truth. Rather, the reason is that for $\beta = 1$ the clustering solutions are uncoupled, i.e., the relevant RV for clustering rows is the noisy column RV. For a certain amount of coupling, i.e., for intermediate values of β, the relevant RV for clustering rows is more strongly related to the column *clusters*, in which noise is reduced due to the averaging effect of clustering. Performance drops again when decreasing β further; the reason is the inherent shortcoming of $\mathcal{L}_0(\Phi, \Psi)$ which is discussed at the end of Section 5 and in [23].

The second experiment investigates the effect of intra-cluster coupling between X and Y. We choose $|\mathcal{X}| = |\mathcal{Y}| = 90$ and $|\overline{\mathcal{X}}| = |\overline{\mathcal{Y}}| = 3$ to avoid the effects discussed in Section 6.1.1 and generate a joint probability distribution

$$P_{X,Y} = \begin{bmatrix} \mathbf{C} & 0 & 0 \\ 0 & \mathbf{C} & 0 \\ 0 & 0 & \mathbf{C} \end{bmatrix}, \tag{23}$$

where $\mathbf{C}$ is a 30×30 circulant matrix the first row of which consists of $30 - k$ zeros followed by k entries equal to $\frac{1}{k|\mathcal{X}|}$. Each subsequent row of $\mathbf{C}$ is obtained by a circular shift of the previous row. Figs. 2f and 2g show $P_{X,Y}$ for $k = 3$ and $k = 15$, respectively. The ground truth co-clustering is given by the block structure of $P_{X,Y}$.

It is clear that, as k decreases, the intra-cluster coupling between X and Y increases. To see this note that, for $k = 30$, X does not contain more information about Y than the ground truth cluster $\overline{X}$ does, whereas for $k = 1$, X specifies Y uniquely. Fig. 2h shows the average MAP values obtained by running AnnITCC 500 times with random initializations. Since the experimental setup is symmetric we only show the results for Φ. First, we observe that with decreasing k the performance deteriorates. This is intuitive considering that with decreasing k the clustering structure becomes less obvious. For $k = 30$, $P_{X,Y}$ is uniform in the the blocks whereas for $k = 1$, the colums of $P_{X,Y}$ can be reordered such that $P_{X,Y}$ is a diagonal matrix with no clear co-clustering structure. Second, $\beta = 1$ does not lead to the best results for increased coupling, despite the fact that the global optimum of $\mathcal{L}_1$ coincides with the ground truth. Apparently, the optimization heuristic tends to terminate in poor local optima more often for $\beta = 1$ than for smaller values of β. This is because for $\beta = 1$ the two clustering solutions are decoupled, i.e., Φ and Ψ are determined independently of each other, while smaller β explicitly assumes coupled clusterings. We thus conclude that smaller values of β detect intra-cluster coupled co-clusters more robustly.

Finally we noticed that for both synthetic datasets, the MAP curves are relatively flat in many scenarios. One may think that this is due to AnnITCC getting stuck in a local optimum for a certain β, which it is not able to escape from for the subsequent lower β values. This is not the case: Figs. 2h and 2d show that the results obtained by running sGITCC (dotted lines) are almost identical to those obtained from AnnITCC for larger values of β until where both of them reach the peak performance. Subsequently, for smaller values of β, the performance of sGITCC dropped significantly due to the reasons outlined at the end of Section 5, justifying using AnnITCC for these values of β.

6.3 Guiding Principles for Choosing β

Although in this paper we do not propose a heuristic to find the suitable value (or range) of β for a given dataset, the examples and experiments in this section admit providing the following guiding principles:

- For large differences between target cardinalities $|\overline{\mathcal{X}}|$ and $|\overline{\mathcal{Y}}|$, larger values of β may lead to better results due to the increasingly decoupled nature of the cost function for increasing β.
- For datasets with highly imbalanced (co-)clusters, smaller values of β are more suitable (but only when one can manage to avoid optimization issues linked to smaller values of β).
- In general, co-clustering using $\mathcal{L}_\beta$ and β-annealing seems to be robust to noise. For large noise levels, however, intermediate values of β tend to perform better due to noise averaging.
- In presence of intra-cluster coupling, local optima of $\mathcal{L}_\beta$ are more prominent for β close to 1. The correct co-clusterings are found more robustly for intermediate values of β.

7 Real-World Experiments

7.1 Document Classification by Co-Clustering of Words and Documents - Newsgroup20 Data Set

7.1.1 *Dataset, Preprocessing, and Simulation Settings*

The Newsgroup20 (NG20) dataset[2] consists of approximately 18800 documents containing 50000 different words. In this section, we evaluate co-clustering performance only via document clusters since there is no ground truth for word clusters. Nevertheless, word clustering was claimed to improve the document clustering performance, cf. [1], [2].

We refer to the RV over words as W, the set of words as $\mathcal{W}$, the RV over the documents as D, and the set of documents as $\mathcal{D}$. The respective clustered RVs and sets are denoted by an overline. The joint distribution of W and D is obtained by normalizing the contingency table (counting the number of times a word appears in a document) to a probability distribution. During preprocessing, we removed

2. qwone.com/~jason/20Newsgroups

TABLE 1
Overview of the Different Subsets Drawn from NG20

Dataset	Discussion Groups	$\frac{\text{docs}}{\text{class}}$	$\lvert\mathcal{D}\rvert$
Binary	talk.politics.mideast, talk.politics.misc	250	500
Multi5	rec.motorcycles, comp.graphics, sci.space, rec.sport.basketball, talk.politics.mideast	100	500
Multi10	comp.sys.mac.hardware, misc.forsale, rec.autos, talks.politics.gun, sci.med, alt.atheism, sci.crypt, sci.space, sci.electronics, rec.sport.hockey	50	500

newsgroup-identifying headers and lowered upper-case letters. We moreover reduced $\mathcal{W}$ to the 2000 words with the highest contribution to $I(D;W)$, which is consistent with the preprocessing in [1], [2], [11]. Finally, we constructed various subsets of the NG20 dataset by randomly selecting 500 documents evenly distributed among the document classes. An overview of the used datasets is given in Table 1.

Note that there are significant differences in the preprocessing steps performed in previous studies. For example, [11] included the newsgroup-identifying header, which may improve clustering performance.

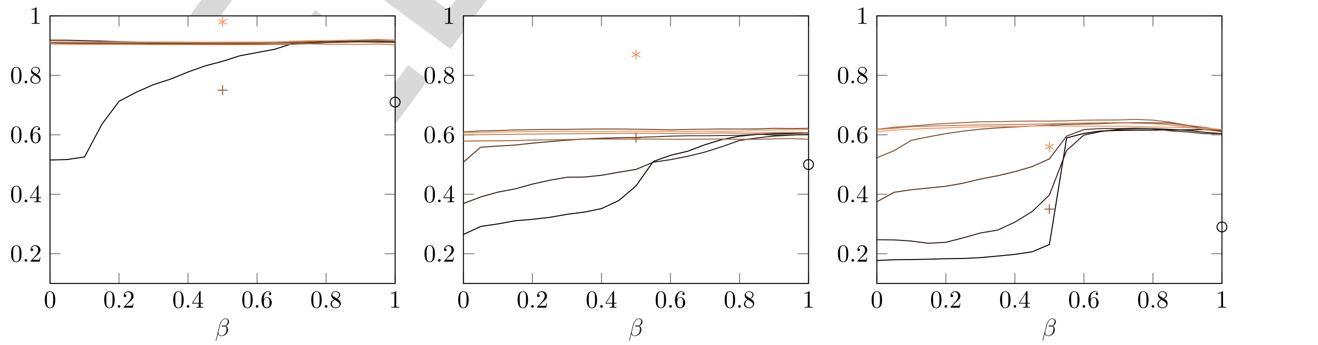
We ran AnnITCC with $\text{tol} = 10^{-3}$, $\Delta = 0.05$ and $\#\text{iter}_{\max} = 20$. For initialization, we slightly changed line 3 in Algorithm 2: Instead of running sGITCC with $\beta = 1$, which is equivalent to the completely decoupled case, we run sIB for both the word and document clusterings separately, where 25 restarts are performed and the best result w.r.t. the cost function is taken. Since there is no ground truth available for the word clusters, we executed AnnITCC for $|\overline{\mathcal{W}}| \in \{2, 4, 8, 16, 32, 64, 128\}$. This is consistent with the simulation settings described in [2], for example.

For a fair comparison of different values of β, we do not apply further heuristics to improve the performance of AnnITCC. In contrast, the authors of [2] initialize their co-clustering algorithm for $|\overline{\mathcal{W}}|$ word clusters with the result obtained for $|\overline{\mathcal{W}}|/2$ word clusters, where each word cluster is split randomly. In [4], the authors introduce an additional correction parameter which leads to clusters of approximately the same size (which matches the evenly distributed classes in the NG20 dataset). Therefore, even for those values of β for which we obtain the same cost functions, our results need not be equal to those reported in the literature.

7.1.2 Results and Comparison

The results obtained by Algorithm 2 - averaged over 20 runs—for the different subsets of NG20 are visualized in Fig. 3. As it can be seen, AnnITCC can discover the true document labels with high accuracy. For the Binary dataset, AnnITCC was able to achieve a micro-averaged precision of approximately 90 percent, for the Multi5 dataset 60 percent and for the Multi10 dataset approximately 60–65 percent. In comparison, experiments with sGITCC confirm the observations from [23] that small $\beta \in [0, 0.4]$ lead to meaningless results in the range of random clustering, while high $\beta \in [0.6, 1]$ produce results in the range of Fig. 3. Fig. 3 further shows that the stronger the word and document clustering solutions are coupled, the worse are the results for small numbers of word clusters. This is most obvious for the Multi10 dataset for $\overline{\mathcal{W}} \in \{2, 4, 8\}$ word clusters, where the MAP values increase sharply if β increases from 0.4 to 0.6 (see Fig. 3c). For small β, the document clusters are obtained from the word clusters and, e.g., two word clusters do not contain sufficient information to distinguish between ten document clusters. This agrees with our discussion in Section 6.1.1. However, for very large $|\overline{\mathcal{W}}|$, there were no further improvements. This suggests that there exists a number of word clusters that are sufficient to achieve the same (or better, see below) performance as document clustering based on words.

One major issue to observe from Fig. 3 is that for the Binary and Multi5 data, the results are almost independent of β (for sufficiently many word clusters). Only for Multi10 there was a mild increase in performance for intermediate values of β. This confirms the observations from Section 6.2: Clustering words removes noise, hence document clustering based on word clusters may be slightly more robust than document clustering based on words. Nevertheless, since the effect is only small for Multi10 (and not present for Binary and Multi5), we doubt that co-clustering of words and documents



(a) Binary (b) Multi5 (c) Multi10

Fig. 3. Micro-averaged precision for different NG20 subsets and AnnITCC. Results are shown for different numbers of word clusters, $|\overline{\mathcal{W}}| = \{2, 4, 8, 16, 32, 64, 128\}$ (darker colors for fewer clusters). For comparison, we added results reported in the literature. $(*)$: Taken from [2, Table 5]; $|\overline{\mathcal{W}}|$ is unclear. $(+, \circ)$: Taken from [1, Table 3]; the best results for each dataset are displayed. These results were obtained by applying aIB for different numbers of word clusters, $|\overline{\mathcal{W}}| = \{10, 20, 30, 40, 50\}$; the displayed MAP values are averages of the individual MAP values. We were not able to compare our results to those of [3] because they used different subsets of the NG20 dataset. Since the cost functions from the literature are the same as ours for the respective values of β, the difference in the performance can only be attributed to preprocessing steps, the optimization heuristics, and/or the choice of favorable data subsets.

is indeed significantly superior to one-sided document clustering w.r.t. the classification results. The classification results from [4] point towards similar conclusions, since also there sIB performed very well compared to the respective co-clustering methods. Still, the authors of [1], [2], [3] claim that their proposed algorithms and/or cost functions for co-clustering outperform one-sided clustering. In the light of our results, we suggest that the choice of the cost function has less effect on the performance than algorithmic details, preprocessing steps, and additional heuristics for, e.g., initialization.

7.2 MovieLens100k

7.2.1 *Dataset, Preprocessing, and Simulation Settings*

The MovieLens100k dataset[3] consists of 100000 ratings of 1682 movies by 943 users [32]. The user ratings take integer values 1 (worst) to 5 (best). We construct a user-movie matrix $\mathbf{R} := [R_{ij}]$ where R_{ij} is the rating user i gave to the movie j ($R_{ij} = 0$ if user i did not rate movie j). Note that $\mathbf{R}$ is a sparse matrix with only 100000 out of approximately 1.59 million entries being nonzero.

We refer to the RV over the users as U, the set of users as $\mathcal{U}$, the RV over movies as M, and the set of movies as $\mathcal{M}$. The respective clustered RVs and sets are denoted by an overline. The joint distribution between U and M is obtained by normalizing $\mathbf{R}$ to a probability distribution.

For initializing AnnITCC we ran sGITCC 25 times with random initializations for $\beta = 1$ with $\mathrm{tol} = 10^{-3}$ and $\#\mathrm{iter}_{\max} = 20$. We chose the best co-clustering (Φ, Ψ) among these 25 restarts w.r.t. the cost and used this as the initialization for AnnITCC. We ran AnnITCC with $\mathrm{tol} = 10^{-3}$, $\Delta = 0.1$ and $\#\mathrm{iter}_{\max} = 20$. We defined 10 user clusters, i.e., $|\overline{\mathcal{U}}| = 10$, as was done in [18], [19]. Furthermore, we defined $|\overline{\mathcal{M}}| = 19$ since the MovieLens100k dataset categorizes the movies into 19 different genres.

7.2.2 *Evaluation Metrics*

Evaluating co-clustering performance for the MovieLens100k dataset is difficult. The authors of [19] proposed to assess co-clustering performance based on recommendations, i.e., a portion of the dataset is used for co-clustering, based on which the "taste" of the users is predicted. The remaining portion of the dataset (i.e., the validation set) is used to assess this prediction. We believe that such an approach is not effective. Indeed, the available ratings in $\mathbf{R}$ are skewed in the sense that approximately 82.5 percent of the ratings are above 3. Hence, a naive recommendation system suggesting a positive rating for every user-movie pair in the validation set matches the user's taste with approximately 82.5 percent. In comparison, the authors of [19] claim a match of 89 percent for their approach.

A second option is to compare the co-clustering results to a plausible ground truth. For the users, demographic information is available which theoretically admits constructing such a ground truth; we nevertheless refrain from doing so, since no choice can be justified without evoking critique. For the movies, genre information is available which lends itself to evaluating movie clusters. However, not every movie is assigned to a unique genre, but may belong to

3. grouplens.org/datasets/movielens/100k

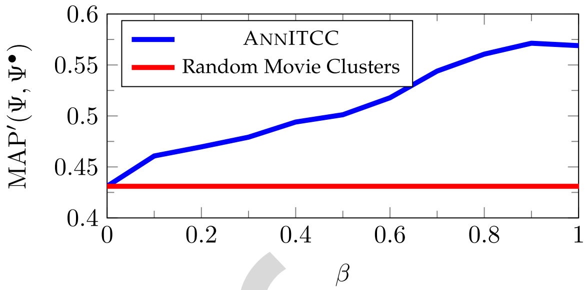


Fig. 4. AnnITCC performance for movie genre matching.

multiple genres. The ground truth $\Psi^\bullet$ is therefore not a function, but a distribution over the set of genres $\overline{\mathcal{M}}$. This is problematic for (21), which is why we replace it here by

$$\mathrm{MAP}'(\Psi, \Psi^\bullet) := \frac{1}{|\mathcal{M}|} \sum_{j \in \overline{\mathcal{M}}} \max_{i \in \overline{\mathcal{M}}} |\Psi^{-1}(j) \cap \Psi^{\bullet -1}(i)|. \qquad (24)$$

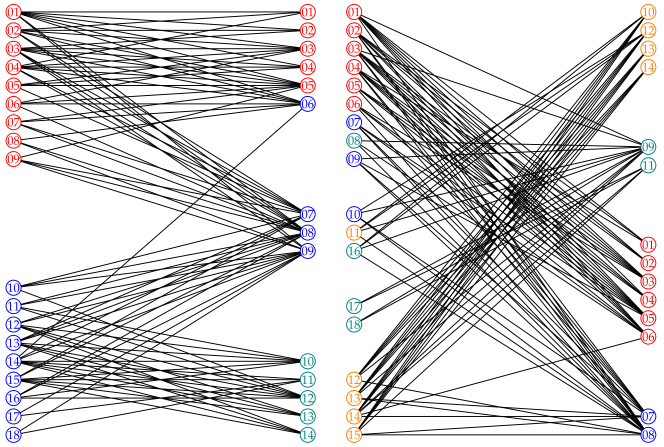
For each movie cluster, we look for the genre with which this cluster has the greatest overlap. Unlike for MAP, two different clusters can now be mapped to same movie genre in MAP'. Hence, MAP', sometimes referred to as *purity*, is essentially the average of the fraction of movies in each cluster that belong to the same genre. As a side result, MAP' gets rid of the maximum over all permutations π, which is intractable for large numbers of genres.

7.2.3 *Results*

The results are shown in Fig. 4. First, note that the MAP' value for randomly generated clusters is remarkably high. This is because the number of movies in different genres varies greatly; for example, 725 movies are assigned to genre "Drama" and 505 to genre "Comedy", whereas only 24 movies belong to the genre "Film-Noir". Noting this, quantitative results based on movie genres are useful to observe trends and general behavior, but the numbers should be taken with a grain of salt. On the other extreme, the maximum value for MAP' in Fig. 4 is significantly smaller than 1. This is reasonable since co-clustering is based on a sparse matrix of user-movie rating pairs: While some users are *genre-addicts* rating movies mainly based on their genre, other users may rate movies based on completely different aspects unrelated to genre. Hence, one cannot expect a value $\mathrm{MAP}' = 1$ for co-clustering based on user-movie rating pairs.

We observe that MAP' generally decreases with decreasing β and the maximum value is at $\beta = 0.9$, albeit only slightly larger than for $\beta = 1$. This shows that our algorithm is capable of outperforming ITCC ($\beta = \frac{1}{2}$), IBCC ($\beta = \frac{3}{4}$), and (albeit only slightly) IB-based ($\beta = 1$) movie clustering. For β close to 0, we obtain results which are very close to what we obtain for randomly generated movie clusters. A closer analysis revealed that the solution found for $\beta = 0$ has a lower cost than the solution found for $\beta = 1$, which means that β-annealing was successful in escaping bad local optima, but that the ground truth does not coincide with the global optimum of the cost function for $\beta = 0$. We believe that, in this particular example, this phenomenon is linked to the user-movie rating matrix $\mathbf{R}$ being sparse.

We finally complement this quantitative evaluation by a qualitative evaluation of the movie clusters. Again, we



(a) Two women communities (left), three event clusters (right). Colors taken from [13].

(b) Four women communities (left), four event clusters (right). Colors taken from [12].

Fig. 5. Community structure of the southern women event participation dataset. The separation between nodes indicates the clustering obtained from AnnITCC with $\beta = 0.7$, the color of the nodes is taken from reference clusterings from the literature.

observe meaningful results for higher values of β when compared to smaller values of β. For example, looking at movie clusters for $\beta = 0.9$, we notice that many classics are clustered into one group, including *Gone With The Wind*, *Breakfast at Tiffany's (1961)*, *12 Angry Men*, *The Graduate*, *The Bridge on River Kwai*, *Citizen Kane*, *Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb*, *Vertigo*, *Casablanca*, *His Girl Friday (1940)*, *A Street Car Named Desire*, *It Happened One Night*, *The Great Dictator*, *The Great Escape*, *Philadelphia Story*. Similarly, many animated/kids movies have been assigned to a cluster, including *The Lion King*, *Alladin*, *Snow White and the Seven Dwarfs*, *Homeward Bound*, *Pinocchio*, *Turbo: A Power Rangers Movie*, *Mighty Morphin Power Rangers: The Movie*, *Cinderella*, *Alice in Wonderland (1951)*, *Dumbo (1941)*, *Beauty and the Beast*, *Winnie the Pooh and the Blustery Day*, *The Jungle Book*, *The Fox and the Hound*, *Parent Trap*, *Jumanji*, *Casper*, etc. Furthermore, our approach clustered various sequences of movies, e.g., 6 out of 8 Star Trek movies and all 7 Amityville movies have been assigned to one cluster each. In contrast, the results for $\beta = 0$ did not yield clusters one would consider meaningful.

7.3 Community Detection in Bipartite Graphs

Community detection is a common problem in social network analysis and is usually concerned with (random) unipartite graphs, see [33]. In this section, we look at the related problem for bipartite graphs. There, the two sets of vertices could be the characters and the scenes of a play, and the goal could be to group characters in a meaningful way.

We apply our algorithm to the Southern Women Event Participation Dataset [12], [33]. The dataset consists of 18 women ($|\mathcal{X}| = 18$) and 14 events ($|\mathcal{Y}| = 14$), and the weight matrix $\mathbf{W}$ contains a one if the corresponding woman participated in the corresponding event and a zero otherwise. We restarted AnnITCC 50 times for $\beta = 1$ to obtain a good initial co-clustering for the annealing process. To get results comparable to those in the literature, we chose $|\overline{\mathcal{X}}| = 2$, $|\overline{\mathcal{Y}}| = 3$ and $|\overline{\mathcal{X}}| = |\overline{\mathcal{Y}}| = 4$. The results are displayed in Fig. 5 for $\beta = 0.7$.

The two women communities we obtained match with those communities reported in the literature [13], [33]. The authors of [13] also clustered the events into three clusters: The events are clustered into a group in which only women of the first women community participated, a group in which only women of the second women community participated, and a group in which women from both communities participated. Our result in Fig. 5a is remarkably similar to theirs, with the exception that the event with label 6 is put in a different group. Note, however, that in this event only one woman of the opposite community participated. Remarkably, we obtained the same co-clustering for all values of β.

For four women communities and four event clusters, we compared our results with those of Barber [12], who employed a modularity-based approach. Our event clusters in Fig. 5b are identical to those of [12], and our women communities are largely consistent. We found in a separate set of experiments that the women communities show a greater agreement for $\beta = 1$, and less agreement for $\beta = \frac{1}{2}$; the MAP values for the chosen value of $\beta = 0.7$ lie in between. Thus, community detection via ITCC can be outperformed by our algorithm for larger values of β.

8 Conclusion

We introduced a generalized framework for information-theoretic co-clustering that arises from recent results on the theory of Markov aggregation. The generalized cost function we proposed allows for trading between completely coupled and decoupled clusterings of two variables connected via a probability table. We obtain well-known previous approaches, e.g., Information-Theoretic Co-Clustering from Dhillon et al., as special cases of our cost function. Using this framework, we provided better understanding of information-theoretic co-clustering in general and discussed some shortcomings inherent to such co-clustering methods.

We performed experiments on both synthetic and real-world data, such as document classification, movie clustering, and community detection. We also demonstrated that our framework can be used to fairly compare various previously proposed cost functions. For example, for the Newsgroup20 dataset, we observed that performance depended little on the cost function, but rather on the optimization heuristic, preprocessing steps, and/or choice of data subsets. We furthermore provide guiding principles for choosing the parameter β of our cost function depending upon the characteristics of the dataset.

Acknowledgments

The work of Rana Ali Amjad has been funded by the German Ministry of Education and Research in the framework of an Alexander von Humboldt Professorship. The work of Bernhard C. Geiger has been funded by the Erwin Schrödinger Fellowship J 3765 of the Austrian Science Fund. The authors contributed equally to this work.

References

[1] N. Slonim and N. Tishby, "Document clustering using word clusters via the information bottleneck method," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2000, pp. 208–215.

[2] I. Dhillon, S. Mallela, and D. Modha, "Information-theoretic co-clusterings," in *Proc. ACM Int. Conf. Knowl. Discovery Data Mining*, Aug. 2003, pp. 89–98.
[3] P. Wang, C. Domeniconi, and K. B. Laskey, "Information bottleneck co-clustering," in *Proc. Workshop Text Mining Held Conjunction SIAM Int. Conf. Data Mining*, May 2010.
[4] R. Bekkerman, R. El-Yaniv, and A. McCallum, "Multi-way distributional clustering via pairwise interactions," in *Proc. Int. Conf. Mach. Learn.*, 2005, pp. 41–48.
[5] R. El-Yaniv and O. Souroujon, "Iterative double clustering for unsupervised and semi-supervised learning," in *Proc. Advances Neural Inf. Process. Syst.*, Dec. 2001, pp. 1025–1032.
[6] I. S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2001, pp. 269–274.
[7] L. Labiod and M. Nadif, "Co-clustering for binary and categorical data with maximum modularity," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2011, pp. 1140–1145.
[8] M. Ailem, F. Role, and M. Nadif, "Co-clustering document-term matrices by direct maximization of graph modularity," in *Proc. ACM Int. Conf. Inf. Knowl. Manag.*, 2015, pp. 1807–1810.
[9] M. Ailem, F. Role, and M. Nadif, "Sparse Poisson latent block model for document clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 7, pp. 1563–1576, Jul. 2017.
[10] S. Sra and I. S. Dhillon, "Nonnegative matrix approximation: Algorithms and applications," Dept. Comput. Sci., Univ. Texas at Austin, Austin, TX, USA, Tech. Rep. TR-06–27, 2006.
[11] N. Slonim, N. Friedman, and N. Tishby, "Unsupervised document classification using sequential information maximization," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2002, pp. 129–136.
[12] M. J. Barber, "Modularity and community detection in bipartite networks," *Phys. Rev. E*, vol. 76, Dec. 2007, Art. no. 066102.
[13] P. Doreian, V. Batagelj, and A. Ferligoj, "Generalized blockmodeling of two-mode network data," *Social Netw.*, vol. 26, no. 1, pp. 29–53, 2004.
[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.
[15] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. Allerton Conf. Commun., Control Comput.*, Sep. 1999, pp. 368–377.
[16] N. Slonim and N. Tishby, "Agglomerative information bottleneck," in *Proc. Advances Neural Inf. Process. Syst.*, Nov. 1999, pp. 617–623.
[17] I. S. Dhillon, S. Mallela, and R. Kumar, "A divisive information-theoretic feature clustering algorithm for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1265–1287, 2003.
[18] A. Banerjee, I. Dhillon, J. Ghosh, S. Merugu, and D. S. Modha, "A generalized maximum entropy approach to Bregman co-clustering and matrix approximation," *J. Mach. Learn. Res.*, vol. 8, pp. 1919–1986, 2007.
[19] C. Laclau, I. Redko, B. Matei, Y. Bennani, and V. Brault, "Co-clustering through optimal transport," in *Proc. Int. Conf. Mach. Learn.*, vol. 70, pp. 1955–1964, Aug. 2017.
[20] K. Deng, P. G. Mehta, and S. P. Meyn, "Optimal Kullback-Leibler aggregation via spectral theory of Markov chains," *IEEE Trans. Autom. Control*, vol. 56, no. 12, pp. 2793–2808, Dec. 2011.
[21] B. C. Geiger, T. Petrov, G. Kubin, and H. Koeppl, "Optimal Kullback-Leibler aggregation via information bottleneck," *IEEE Trans. Autom. Control*, vol. 60, no. 4, pp. 1010–1022, Apr. 2015.
[22] Y. Xu, S. M. Salapaka, and C. L. Beck, "Aggregation of graph models and Markov chains by deterministic annealing," *IEEE Trans. Autom. Control*, vol. 59, no. 10, pp. 2807–2812, Oct. 2014.
[23] R. A. Amjad, C. Blöchl, and B. C. Geiger, "A generalized framework for Kullback-Leibler Markov aggregation," *IEEE Trans. Autom. Control*, 2017.
[24] D. H. Wolpert, J. A. Grochow, E. Libby, and S. DeDeo, "Optimal high-level descriptions of dynamical systems," Jun. 2015.
[25] J. G. Kemeny and J. L. Snell, *Finite Markov Chains*, 2nd ed. Berlin, Germany: Springer, 1976.
[26] M. Rosenblatt, "Functions of a Markov process that are Markovian," *J. Math. Mech.*, vol. 8, pp. 585–596, 1959.
[27] P. Buchholz, "Exact and ordinary lumpability in finite Markov chains," *J. Appl. Probability*, vol. 31, no. 1, pp. 59–75, 1994.
[28] L. Gurvits and J. Ledoux, "Markov property for a function of a Markov chain: A linear algebra approach," *Linear Algebra Appl.*, vol. 404, pp. 85–117, 2005.
[29] B. C. Geiger and C. Temmel, "Lumpings of Markov chains, entropy rate preservation, and higher-order lumpability," *J. Appl. Probab.*, vol. 51, no. 4, pp. 1114–1132, Dec. 2014.
[30] O. Pfante, N. Bertschinger, E. Olbrich, N. Ay, and J. Jost, "Comparison between different methods of level identification," *Advances Complex Syst.*, vol. 17, no. 02, 2014, Art. no. 1450007.
[31] M. Plumbley, "Information theory and unsupervised neural networks," Cambridge Univ. Eng. Dept., Cambridge U.K., Tech. Rep. CUED/F-INFENG/TR. 78, 1991.
[32] F. M. Harper and J. A. Konstan, "The movielens datasets: History and context," *ACM Trans. Interactive Intell. Syst.*, vol. 5, 2015, Art. no. 19.
[33] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3, pp. 75–174, 2010.

Clemens Blöchl received the BSc and MSc degrees in electrical engineering and information technology from the Technical University of Munich, Germany, in 2014 and 2017, respectively. In 2017, he joined Rohde & Schwarz GmbH & Co. KG in Munich, where he is currently involved in the field of radio communication testing. His research interests include channel coding theory and information-theoretic system design with particular interest in the field of information-theoretic clustering.

Rana Ali Amjad (S'13) received the bachelor's degree in electrical engineering (with highest distinction) from the University of Engineering and Technology, Lahore, Pakistan, in 2011, and the master's degree in communication engineering (with highest distinction) from the Technical University of Munich, Germany, in 2013. Since 2014, he has been working toward the PhD degree at the Institute for Communication Engineering, Technical University of Munich. He has received various awards in his academic career including the Faculty Award for Best Master Thesis, Award for Outstanding Performance in Master's Degree, and Gold Medal for Best Performance in Communications major during his bachelor's degree. His research interests cover information and communication theory, machine learning, channel coding, and information-theoretic security. He is a student member of the IEEE.

Bernhard C. Geiger (S'07-M'14) received the Dipl.-Ing. degree in electrical engineering (with distinction) and the Dr. techn. degree in electrical and information engineering (with distinction) from the Graz University of Technology, Austria, in 2009 and 2014, respectively. In 2009, he joined the Signal Processing and Speech Communication Laboratory, Graz University of Technology, as a project assistant and took a position as a research and teaching associate at the same lab, in 2010. He was a senior scientist and Erwin Schrödinger Fellow at the Institute for Communications Engineering, Technical University of Munich, Germany, from 2014 to 2017 and a postdoctoral researcher with the Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria, from 2017 to 2018. He is currently a senior researcher with the Know-Center GmbH, Graz, Austria. His research interests cover information theory for machine learning, and information-theoretic model reduction for Markov chains and hidden Markov models. He is a member of the IEEE.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/publications/dlib.

Queries to the Author

Q1. Please either annotate the IEEE Proof PDF or send a list of corrections. Do not send new source files as we do not reconvert them at this production stage.

Q2. There was discrepancy between source and pdf file. We followed source file.

Q3. Please provide page range for Ref. [3].

Q4. Please provide complete bibliography details for Ref. [23] and [24].