Information Theoretic Model Validation for Spectral Clustering

Morteza Haghir Chehreghani
ETH Zurich
morteza.chehreghani@inf.ethz.ch

Alberto Giovanni Busetto
ETH Zurich and CC-SPMD, Zurich
alberto.busetto@inf.ethz.ch

Joachim M. Buhmann
ETH Zurich and CC-SPMD, Zurich
jbuhmann@inf.ethz.ch

Abstract

Model validation constitutes a fundamental step in data clustering. The central question is: *Which cluster model and how many clusters are most appropriate for a certain application?* In this study, we introduce a method for the validation of spectral clustering based upon approximation set coding. In particular, we compare correlation and pairwise clustering to analyze the correlations of temporal gene expression profiles. To evaluate and select clustering models, we calculate their reliable informativeness. Experimental results in the context of gene expression analysis show that pairwise clustering yields superior amounts of reliable information. The analysis results are consistent with the Bayesian Information Criterion (BIC), and exhibit higher generality than BIC.

1 Introduction

Clustering constitutes a fundamental task in exploratory data analysis to compress the data and to abstract concepts. Typically, the analyst has the possibility to select a clustering model from a plethora of alternatives. The central question is: *Which model is most appropriate for a certain application?* In this study, we introduce a method for the statistical validation of spectral clustering by searching for an information theorectically optimal tradeoff between stability and informativeness.

The structure of this manuscript is as follows. This section contains a summary of relevant work and of the preliminary background. Subsequently, we introduce the method and its properties. Finally, we report and discuss experimental results with gene expression profiles.

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX.

Motivation. Spectral clustering (that is, clustering with data characterized by relations rather than by vectors) lacks a reliable principle for model selection. This deficit exists despite the theoretical and practical importance of the task [6]. Typically, model selection is performed according to heuristic approaches or it requires expert knowledge. Heuristic strategies are also employed for model order selection, which concerns the number of clusters [11, 15]. At present, practical solutions appeal to intuition. General principles for validation are not available.

Related work. Due to simplicity and proved effectiveness, the Bayesian Information Criterion (BIC) is arguably the favored principle for model order selection [13]. Nevertheless, BIC exhibits two fundamental limitations. First, its applicability is limited to models with directly determined free parameters (e.g. mixture of Gaussians). Second, BIC can be applied only to finite dimensional data spaces. Its extension to spectral clustering remains unclear, because the effective dimensionality grows with the sample size.

More generally, model selection criteria indicate a tradeoff between two antagonistic goals: informativeness and solution stability. Abundant results demonstrate the power of stability criteria for model order selection [7, 11] (although not undisputedly [3]). In the same spirit, PAC-Bayesian generalization bounds have been derived for different clustering models [14]. Appeal to stability in data analysis is related to two fundamental scientific concepts: mathematical well-posedness [10] and experimental repeatability. In principle, controlled experiments are expected to obtain similar outcomes under similar (controlled) conditions. But what about informativeness? Overestimating the information content increases the overfitting risk. The explanatory power of models comes from their ability to generalize. Quantifying the "information content" of a solution constitutes a necessary condition to obtain the optimal tradeoff between stability and informativeness. Reasonable reductions of stability are acceptable, as long as they are compensated by better representations [16]. Such tradeoffs are conventionally calculated within the framework of statistical learning theory. In this context, a central concept is that of generalization capacity [17]. Absolute measures of capacity can be obtained through information theory. Ap-

proximation Set Coding (ASC) [4, 5] is centered around the identification of optimal tradeoffs between stability and informativeness. The process benefits from the analogy to communication: optimal models maximize the reliable information transfer between data sets.

Contribution. This study introduces a method for the validation of spectral clustering based upon the principle of ASC. The method converts sets of approximate clustering solutions into codes. The noisy channel of the equivalent communication scenario is defined (explicitly) by the cost function and (implicitly) by the fluctuations of the data. First, we demonstrate how to apply this principle to spectral clustering. Then, we select the number of clusters which maximize the reliable informativeness of the solution. Finally, we select alternative clustering models (correlation and pairwise clustering) for grouping genes, a scientific data analysis application in molecular biology.

The method is evaluated on temporal gene expression profiles of *M. galloprovincialis*. We show that pairwise clustering provides three times more informative solutions given the gene expression profiles than correlation clustering. We compare our results with BIC and the stability criterion, demonstrating consistency and wide applicability.

2 ASC based model validation

We start by introducing notation and basic definitions, i.e., summarizing the relevant background with particular emphasis to the ASC principle.

Notation and basic definitions. Let $\mathbf{O}$ be a set of n objects respectively associated with their measurements $\mathbf{X}$. Multiple representations of the measurements are admissible. In general, one might specify *vector-based* representations (where $\mathbf{x}_i$ denotes the vector of object i), as well as *relation-based* ones (where the entry X_{ij} refers to the pairwise similarity between objects i and j). Different clustering models are characterized by distinct cost functions, which are denoted by $R(c, \mathbf{X})$. Relevant parameters are incorporated in clustering solutions (denoted by c). $c(i)$ indicates the cluster index for object i. Cluster membership can be encoded by the binary co-clustering matrix $\mathbf{H}$. In this case, one has that $H_{ij} = 1$ if and only if objects i and j belong to the same cluster. Otherwise, $H_{ij} = 0$. In parametric models (such as K-means), the clustering solution c contains the inferred parameters as well (that is, the centroids $\boldsymbol{\mu}_k$). The solution $c^{\perp}(\mathbf{X})$ indicates the empirical minimizer of a given cost function.

The principle of ASC. Consider two data sets, denoted by $\mathbf{X}^{(m)}, m \in \{1, 2\}$. Let us assume that they share the same inherent structure. At the same time, the two data sets exhibit different noise instantiations. For a given cost function, the globally minimal solutions corresponding to the two data sets are generally different, meaning $c^{\perp}(\mathbf{X}^{(1)}) \neq c^{\perp}(\mathbf{X}^{(2)})$. This variability is caused by data fluctuations which induce stochastic variations in the solutions as well. Hence, partitioning a data set by calculating the empirical optimum of clustering costs lacks robustness. ASC addresses this problem by ranking all clustering solutions according to *approximation weights*, which are denoted by $w(c, \mathbf{X})$. For this purpose, one can use the family of Boltzmann weights $w(c, \mathbf{X}) := \exp(-\beta R(c, \mathbf{X}))$ that are parametrized by the inverse computational temperature β. They define the two *weight sums*

$$Z_m = \sum_{c \in \mathcal{C}(\mathbf{X}^{(m)})} \exp\left(-\beta R(c, \mathbf{X}^{(m)})\right),\ m = 1, 2 \quad (1)$$

and the *joint weight sum*

$$Z_{12} = \sum_{c \in \mathcal{C}(\mathbf{X}^{(2)})} \exp\left(-\beta (R(c, \mathbf{X}^{(1)}) + R(c, \mathbf{X}^{(2)}))\right). \quad (2)$$

The set of all clusterings of $\mathbf{X}^{(m)}$ is here denoted by $\mathcal{C}(\mathbf{X}^{(m)})$. Essentially, Z_m counts all statistically indistinguishable clustering solutions which approximate the point of minimum cost. At the same time, $\exp(-\beta(R(c, \mathbf{X}^{(1)}) + R(c, \mathbf{X}^{(2)})))$ measures how well a solution c minimizes costs on *both* data sets.

Learning problems can be defined in terms of identification of optimal "resolution", which is induced by the fluctuations of the noise. In ASC, this resolution scale is formalized by the choice of the parameter β. How large can β be chosen to still ensure stability of solution under variation of the data $\mathbf{X}$? Too low β yields excessively coarse resolutions. The maximal amount of predictive information is not captured in this limit.

The optimal β maximizes the mutual information [4]

$$\mathcal{I}_{\beta}(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}) = \frac{1}{n} \log \left(\frac{|\{\sigma\}| Z_{12}}{Z_1 \cdot Z_2} \right) . \quad (3)$$

For $\mathbf{X}^{(1)}$, the number of all distinct clustering solutions is given by the cardinality $|\{\sigma\}|$. The maximum of $\mathcal{I}_{\beta}$ as a function of β is called the *Approximation Capacity* (AC), $\mathrm{AC}(R) \equiv \mathcal{I}_{\beta^{\star}}$ with $\beta^{\star} = \arg\max_{\beta} \mathcal{I}_{\beta}$. AC determines the best resolution in the solution space achievable for cost function R, that is the best tradeoff between stability and informativeness.

ASC for clustering. ASC enables the comparative evaluation of cluster models. Consider a clustering problem with a number of clusters K. The potential h_{ik} indicates the costs of assigning object i to cluster k. For the moment, let us assume that the potentials h_{ik} are provided, for $1 \leq i \leq n, 1 \leq k \leq K$. In a factorial model such as K-means, the cost function can be expressed as

$R(c, \mathbf{X}) = \sum_i \|\mathbf{x}_i - \boldsymbol{\mu}_{c(i)}\|^2$. In this case, the potential $h_{i,c(i)} = \|\mathbf{x}_i - \boldsymbol{\mu}_{c(i)}\|^2$ corresponds to squared distance between data vector and centroid. For non-factorial models, however, computing the potentials is not as straightforward. In the method section, we explain how to compute the potentials for non-factorial models.

Given the potentials, the AC is computed in two main steps:

1. Calculate the cardinality $|\{\sigma\}|$, the weight sums Z_1,Z_2 and the joint weight sum Z_{12}.
2. Maximize $\mathcal{I}_\beta$ (3), which yields $\beta^\star = \arg\max_\beta \mathcal{I}_\beta$.

The cardinality $|\{\sigma\}|$ is determined by the entropy of the empirical minimizer $c^\perp(\mathbf{X}^{(1)})$ (depending on its type). In other words, one has

$$\frac{1}{n} \log |\{\sigma\}| = -\sum_{k=1}^{K} P_k \log P_k, \tag{4}$$

where P_k is the probability of the k^{th} cluster in $c^\perp(\mathbf{X}^{(1)})$.

The weight sums ($m = 1, 2$) are calculated by

$$\begin{aligned} Z_m &= \sum_{c \in \mathcal{C}(\mathbf{X}^{(m)})} \exp\left(-\beta \sum_{i=1}^{n} h_{i,c(i)}^{(m)}\right) \\ &= \prod_{i=1}^{n} \sum_{k=1}^{K} \exp\left(-\beta h_{ik}^{(m)}\right). \end{aligned} \tag{5}$$

Similarly, the joint weight sum amounts to

$$\begin{aligned} Z_{12} &= \sum_{c \in \mathcal{C}(\mathbf{X}^{(2)})} \exp\left(-\beta \sum_{i=1}^{n} (h_{i,c(i)}^{(1)} + h_{i,c(i)}^{(2)})\right) \\ &= \prod_{i=1}^{n} \sum_{k=1}^{K} \exp\left(-\beta \left(h_{ik}^{(1)} + h_{ik}^{(2)}\right)\right). \end{aligned} \tag{6}$$

In general, there might not exist any analytic solution which satisfies $\frac{d\mathcal{I}_\beta}{d\beta} = 0$. This limitation can be overcome by numerical optimization.

Model (order) selection. AC evaluates models and model orders. To identify the optimal number of clusters, one has to select an initial number of clusters K_{max} and then select the correct number according to its score. **Algorithm 1** describes this procedure.

At each (computational) temperature β^{-1}, potentials h_{ik} and other parameters are updated. In the case of $K \geq K^{\text{opt}}$, the optimal parameters are expected to be found at the optimal temperature. The result corresponds to the identification of the correct number of clusters. Greater detail about this task is provided in the method section, where concrete case studies are considered.

Algorithm 1 Calculate_Model_Order

1: **for** $1 \leq K \leq K_{\text{max}}$ **do**
2: Perform either annealed Gibbs sampling or annealed *mean-field approximation* to compute the potentials h_{ik} at different temperatures.
3: At each temperature $\beta^{-1} \in$, calculate $\mathcal{I}_\beta^K$.
4: Compute $\text{AC}^{\text{K}} := \max_\beta \mathcal{I}_\beta^{\text{K}}$ (that is, the AC).
5: **end for**
6: **return** $K^{\text{opt}} := \arg\max_{1 \leq K \leq K_{\text{max}}} \text{AC}^{\text{K}}$.

Similarly, a set of L candidate cost functions $\mathcal{R} = \{R^l(., \mathbf{X}) : 1 \leq l \leq L\}$ can be ranked by their ACs: after the evaluation, the model exhibiting the highest AC is selected.

Algorithm 2 Model_Comparison

1: Fix initial number of clusters K_{init} sufficiently large.
2: **for** $R^l(., \mathbf{X}) \in \mathcal{R}$ **do**
3: Compute the approximation capacity AC^{R^l}.
4: **end for**
5: **return** $R^{\text{opt}} := \arg\max_{1 \leq l \leq L} \text{AC}^{\text{R}^l}$.

Computational complexity. ASC requires to estimate weight sums. Thus, its computational costs are dominated by the computational complexity of approximating partition functions. Algorithmic techniques to estimate partition functions such as MCMC methods or variational Bayes methods are applicable. Furthermore, subsampling can be beneficial to compute AC for large data sets.

3 Method

In this section, we introduce the new validation method for spectral clustering. The study investigates two spectral models: pairwise clustering and correlation clustering. Both respective subsections follow parallel structure, consisting of problem formulation and calculation of AC.

A spectral clustering problem is often mathematically characterized by an attributed graph $(\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$. The vertices have to be clustered into groups $\mathcal{G}_u := \{i : c(i) = u\}, 1 \leq u \leq K$. The set of edges between elements of group $\mathcal{G}_u$ and $\mathcal{G}_v$ is denoted by $\mathcal{E}_{uv} := \{(i, j) : c(i) = u \wedge c(j) = v\}$.

Algorithm sketch: Given are data and a class of models,

1: Split the data set in two subsets.
2: **for** all models **do**
3: evaluate the AC of all model orders (that is, find the correct number of clusters).
4: select the correct model order.
5: **end for**
6: evaluate the AC of all models.

7: select the correct cluster model.

The output consists of the model which exhibits the optimal tradeoff between informativeness and stability.

3.1 AC of pairwise clustering

Arguably, K-means is one of the preferred choices for clustering in many application domains. This preference assumes vector data in a Euclidean space. What if the measurements are characterized by relations, such as pairwise similarities? Then the potentials $\{h_{i,c(i)}\}$ cannot be directly extracted from the cost function. As a general strategy, we propose to use *mean-field approximations* to estimate the potentials. However, for the specific case of pairwise clustering [8], there exists a straightforward way to compute the potentials by embedding relational data into a Euclidean vector space (without distortions of the clustering solutions).

Problem formulation. Given a set of pairwise similarities $\mathbf{X}$, the pairwise clustering cost function is defined as

$$R^{pc}(c, \mathbf{X}) = -\frac{1}{2} \sum_{k=1}^{K} |\mathcal{G}_k| \sum_{(i,j) \in \mathcal{E}_{kk}} \frac{X_{ij}}{|\mathcal{E}_{kk}|}. \qquad (7)$$

This cost function sums the average similarities per cluster (weighted by the respective cluster sizes). Therefore, adding a constant to all pairwise similarities shifts the cost value by a constant multiplied by the number of objects. At the same time, it does not modify the order of the clusterings induced by the costs [12], nor it changes the approximation capacity (3). This invariance renders the embedding of the objects into a $n-1$ dimensional kernel space possible. The similarity X_{ij} is then interpreted as a scalar product between two vectors representing object i and j. If we appropriately convert similarities $\mathbf{X}$ into dissimilarities ($\mathbf{D} = \text{const} - \mathbf{X}$), then pairwise clustering equivalently performs K-means clustering in kernel space. Using mean-field approximations, the calculation of the weight sums (5), which can be performed analytically for K-means clustering, is hence exact for pairwise clustering (see [12]).

Calculation of AC. For the purpose of illustration, let us consider two sets of objects $\mathbf{O}^{(1)}$ and $\mathbf{O}^{(2)}$ of identical size, consisting of $n = 800$ objects each. The data sets are drawn from four isotropic Gaussian sources. For each source, the component parameters are $\pi_k = 1/4$, the means are $\boldsymbol{\mu} = [(4,4); (-4,4); (-4,-4); (4,-4)]$ and the covariances are isotropic $\boldsymbol{\Sigma} = 5 \cdot \mathbf{I}$. Now, we convert the pairwise squared Euclidean distances into similarity matrices. On the basis of this information, the potentials $\{h_{i,c(i)}, 1 \leq i \leq n\}$ are calculated by performing annealed Gibbs sampling for different numbers of initial clusters (varying from 1 to 10). For different β, the mutual information is calculated using the potentials $\{h_{i,c(i)}\}$. The approximation capacity is then obtained by maximization.

Figure 1 shows the data set (a) and the approximation capacity for different numbers of initial clusters. The AC saturates at approximately 1.5 bits per object since objects have probabilistic assignments to clusters at the optimal $\beta^\star$ value. Note that in the case of very well separated clusters, AC leads to $\log_2(4) = 2$ bits. As shown in Figure 1(c), BIC and AC selections are consistent for these data sets.

For the case $K = 8$, the trajectories of the clusters are illustrated in detail. As a function of the (inverse) temperature, the positions of the centroids diverge as the system cools down (increase of β) and the model parameters are optimized. Figure 2(a) shows the positions of the inferred centroids. The colors of the trajectories indicate the value of β. $\beta \approx 0$ is dark blue and $\beta \gg 1$ is red. The transition from green to yellow denotes the simultaneous split of four to eight clusters at $\beta^\star$. At high temperature all centroids coincide, indicating that the optimizer favors a single cluster. At very low to zero temperature, the algorithm estimates 8 clusters with locations strongly determined by fluctuations. Figure 2(b) depicts the mutual information as a function of β. The optimal temperature corresponding to the approximation capacity is the lowest temperature at which the correct number of clusters is found.

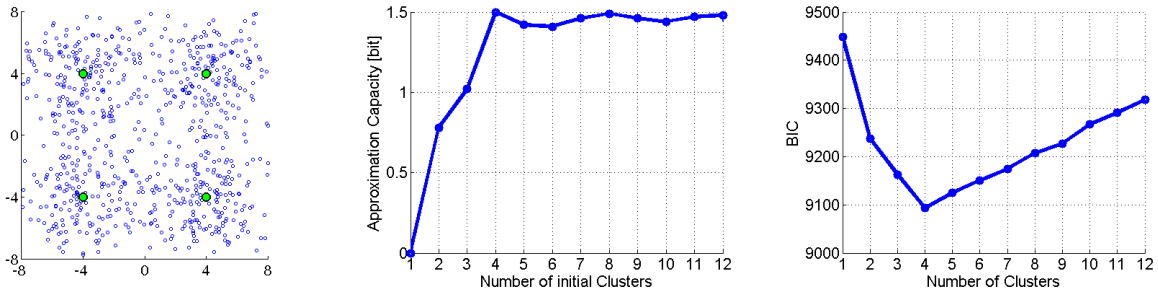
3.2 AC of correlation clustering

An alternative clustering principle for relational data is defined by correlation clustering [2]. Objects are partitioned based on a multi-cut of a similarity graph that is annotated with positive and negative edge weights. For correlation clustering, the potentials have to be approximated through e.g. *mean-field approximation* since no product form of the weight sums (1, 2) is known. Approximate estimates of $Z_m,\ Z_{12}$ can be inserted into Eq. (3) to calculate the AC.

Problem formulation. We consider the complete graph with vertex set $\mathbf{O}$ and edge weights given by a similarity matrix $\mathbf{X} := \{X_{ij}\}$ (between objects i and j). The cost function for correlation clustering consists of the sum of the disagreements. The disagreements correspond to negative intra-cluster edges and positive inter-cluster edges. For the tuple $(c, \mathbf{X})$, the costs are

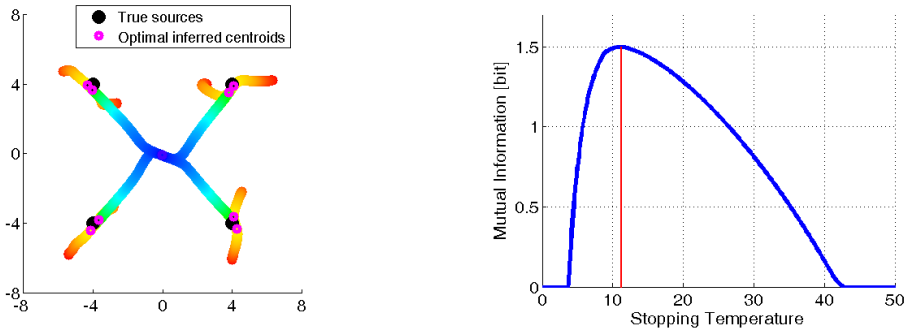
$$R^{cc}(c, \mathbf{X}) = \frac{1}{2} \sum_{1 \leq u \leq K} \sum_{(i,j) \in \mathcal{E}_{uu}} (|X_{ij}| - X_{ij}) + \frac{1}{2} \sum_{1 \leq u \leq K} \sum_{1 \leq v < u} \sum_{(i,j) \in \mathcal{E}_{uv}} (|X_{ij}| + X_{ij}). \qquad (8)$$

Non-factorial mean-field approximation. In statistical models where Boltzmann weights do not assume a product form, such as correlation clustering, the weight sums



(a) the data generated with $\mathbf{\Sigma} = 5 \cdot \mathbf{I}$ (b) approximation capacity (c) BIC score

Figure 1: Approximation capacity and BIC score for pairwise clustering as a function of K. Both principles yield consistent results on these two dimensional data (Figure 1(a)).



(a) Cluster trajectories (b) Mutual Information

Figure 2: Annealed Gibbs sampling for pairwise clustering. Influence of the stopping temperature for annealed optimization on the mutual information and on the positions of the cluster centroids.

have to be calculated numerically by sampling or analytically by a mean-field approximation. The potentials h_{ik} are determined by approximating the Gibbs distribution $p(c) = w(c, \mathbf{X})/Z$ by a factorial distribution with the mean-fields as adjustable parameters. Given these potentials, the assignments $c(.)$ of objects to clusters are independent, i.e., $c(i)$ is not influencing $c(j), j \neq i$. The family of factorial distributions is defined as

$$\mathcal{Q} = \left\{ \mathbf{Q} : \mathbf{Q}(c) = \prod_{i=1}^{n} q_{i,c(i)}, \quad q_{i,c(i)} \in [0,1] \right\}. \quad (9)$$

The closest factorial distribution (in an information theoretic sense) can be determined by minimizing the Kullback-Leibler divergence (see[8])

$$D_{KL}\left(\mathbf{Q}||\mathbf{P}^{cc}\right) = \sum_{c \in C} \mathbf{Q} \log \frac{\mathbf{Q}}{\exp(-\beta(R^{cc} - F^{cc}))}$$
$$= \sum_{i=1}^{n} \sum_{k=1}^{K} q_{ik} \log q_{ik} + \beta \mathbb{E}_{\mathbf{Q}}\{R^{cc}\} - \beta F^{cc}. \quad (10)$$

The free energy $F^{cc} := -\beta \log Z(\mathbf{X})$ does not depend on q_{ik}. To find the optimal factorial distribution, we minimize $D_{KL}\left(\mathbf{Q}||\mathbf{P}^{cc}\right)$ w.r.t q_{ik} observing the normalization constraint $\sum_{k=1}^{K} q_{ik} = 1, \forall i$:

$$0 = \frac{\partial}{\partial q_{ik}} \left[D_{KL}\left(\mathbf{Q}||\mathbf{P}^{cc}\right) + \sum_{j=1}^{n} \lambda_j \left(\sum_{k=1}^{K} q_{jk} - 1 \right) \right] \quad (11)$$
$$= \sum_{c \in \mathcal{C}} \prod_{j \leq n : j \neq i} q_{j,c(j)}^{\mathbb{I}_{\{c(i)=k\}}} R^{cc} + \frac{1}{\beta}(\log q_{ik} + 1) + \lambda_i.$$

For the extremum of the bound, the necessary condition determines the mean-field assignments

$$q_{ik} = \frac{\exp(-\beta h_{ik})}{\sum_{k'} \exp(\beta h_{ik'})}, \text{ with } h_{ik} = \mathbb{E}_{\mathbf{Q}_{i \to k}}\{R^{cc}\}. \quad (12)$$

$\mathbb{E}_{\mathbf{Q}_{i \to k}}\{R^{cc}\}$ is the expectation over all configurations subject to the constraint of assigning object i to cluster k. Thereby, to calculate the mean-fields, R^{cc} is decomposed into contributions which depend on object i and on the costs of all other objects. Each q_{ik} is influenced uniquely by the terms which depend on object i.

For correlation clustering, mean-field approximation yields

the equations

$$h_{ik} = \frac{1}{2} \sum_{j \le n: j \neq i} (|X_{ij}| + X_{ij})(1 - \mathbb{E}_{\mathbf{Q}_{i \to k}}\{\mathbb{I}_{\{c(j)=k\}}\}) + \frac{1}{2} \sum_{j \le n: j \neq i} (|X_{ij}| - X_{ij}) \mathbb{E}_{\mathbf{Q}_{i \to k}}\{\mathbb{I}_{\{c(j)=k\}}\} + \texttt{const.}$$
$$= \frac{1}{2} \sum_{j \le n: j \neq i} (|X_{ij}| + X_{ij})(1 - q_{jk}) + \frac{1}{2} \sum_{j \le n: j \neq i} (|X_{ij}| - X_{ij}) q_{jk} + \texttt{const.} \quad (13)$$

Through the annealing scheme, at each temperature an iterative EM-type algorithm approximates the mean-fields and the probabilities by mutual conditioning. The t-th iteration of the algorithm consists of two main steps. First, $q_{ik}^{(t)}$ is estimated as a function of ${h_{ik}}^{(t-1)}$. Second, ${h_{ik}}^{(t)}$ is calculated for given $q_{ik}^{(t)}$. Finally, the weight sums (1, 2) and the AC are calculated from the potentials.

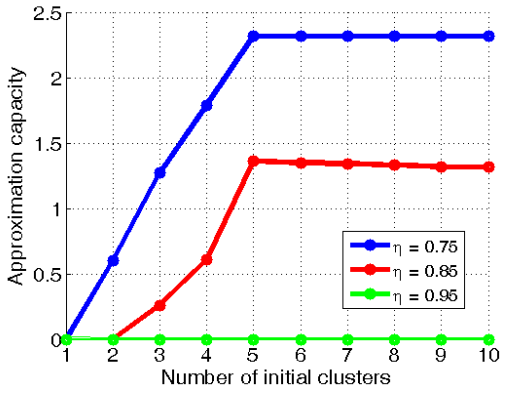


Figure 3: Approximation capacity in three different settings of correlation clustering for $\xi = 0.35$.

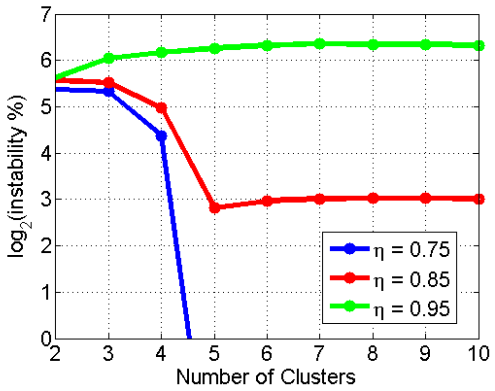


Figure 4: Instability measure in three different settings of correlation clustering for $\xi = 0.35$. For $\eta = 0.75$ instability is always zero for 5 and more clusters.

Calculation of AC. For the experiments on synthetic data, we generate two correlation graphs $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$. Given the noise parameter η and the complexity parameter ξ, the correlation graph is constructed as follows:

1. First, a perfect graph is constructed. In other words, $+1$ is assigned to intra-cluster edges, while -1 to inter-cluster edges.

2. Then, each edge in $\mathcal{E}_{uv}, v \neq u$ is flipped to $+1$ with probability ξ. This step tends to increase the complexity of the structure.

3. Finally, each edge ($\mathcal{E}_{uv}, v \neq u$ and $\mathcal{E}_{uu}$) is replaced by a random edge with probability η.

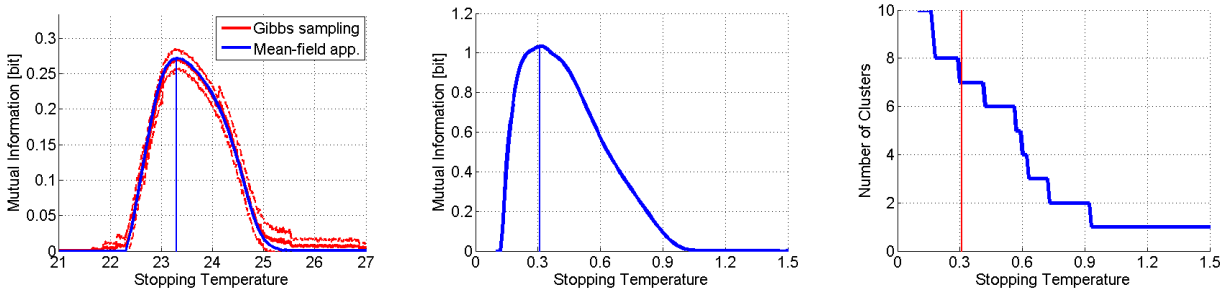
By construction, each graph consists of 1500 nodes and 5 clusters. Identity mapping between objects in $\mathbf{O}^{(1)}$ and $\mathbf{O}^{(2)}$ is guaranteed by the same order of construction of the graphs. Structure complexity is anchored at $\xi = 0.35$ and noise level η varies from 0.75 to 0.95, thus generating data sets with a broad range of difficulty.

Varying the number of initial clusters from 1 to 10, the mean-field algorithm is executed with 10 random initializations per model order. The best result in terms of cost value is taken at each round, and on the basis of that the mutual information is optimized over different values of β. Figure 3 illustrates the results of this procedure. The results of mean-field approximation are verified by checking the consistency with Gibbs sampling.

The data analysis problem is easy for $\eta = 0.75$. In this regime, Gibbs samplers or mean-field annealing procedures select the correct number of clusters even when initialized with a large number of clusters. In fact, superfluous clusters are simply left empty as the cost function prefers large clusters for low noise levels. The effective number of clusters remains 5 regardless of the initialization and, hence, the approximation capacity is invariant.

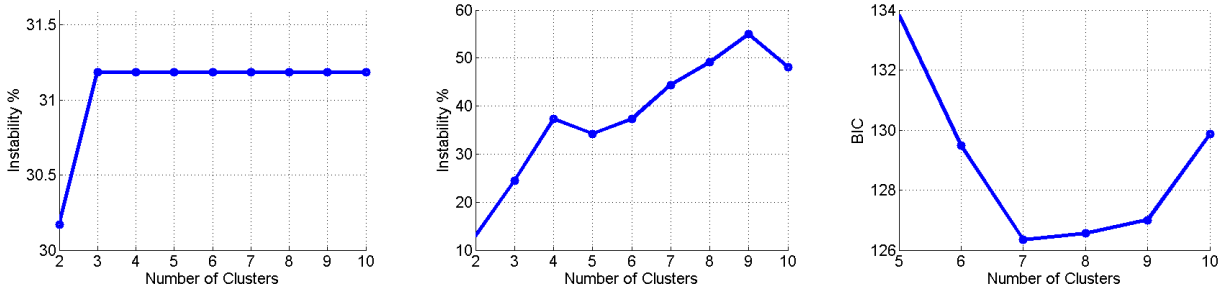
At $\eta = 0.85$ the problem is rather complicated due to noise but still learnable. In this regime, substantial variations are exhibited in the inferred clustering for different choices of the number of clusters. Approximation capacity systematically selects the correct number of clusters. For larger numbers of clusters, the effective number is still 5 and the capacity reduces slightly due to degeneracy. Both for $\eta = 0.75$ and $\eta = 0.85$, the instability measure (computed as proposed in [11]) is consistent with the ASC principle (see Figure 4).

At $\eta = 0.95$ the edge labels are almost entirely random, obfuscating all structure in the data. Therefore, as shown in Figure 3, the number of learnable clusters is just 1. In this regime, instability cannot be used to determine the number of clusters as it remains undefined for $K = 1$.



(a) correlation clustering (CC) (b) pairwise clustering (PC) (c) number of clusters in PC

Figure 5: Approximation capacity for correlation clustering and for pairwise clustering applied to Gene expression data. For correlation clustering the mutual information is computed by mean-field approximation and by Gibbs sampling to compare both methods. Figure 5(c) shows the number of clusters at different temperatures.



(a) instability of CC (b) instability of PC (c) BIC score for PC

Figure 6: Instability measure and BIC score computed for Gene expression data. BIC relies upon the calculation of the effective dimensionality. In some cases, such as pairwise clustering, the number of free parameters is unclear but heuristics exist. In other cases, such as correlation clustering, it is rather problematic.

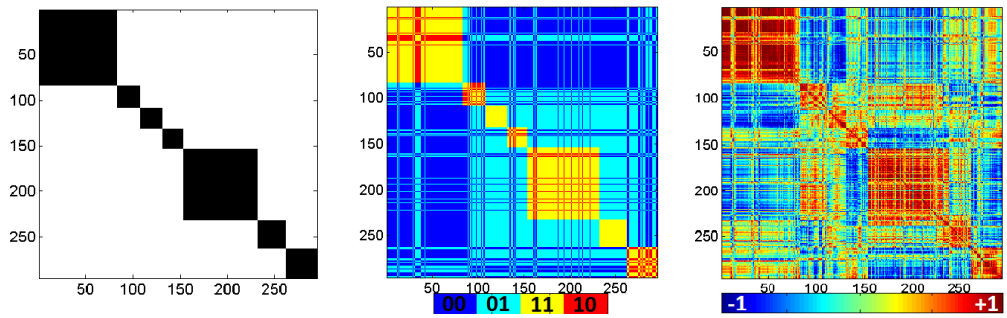
4 Clustering of gene expression profiles

In this section, we analyze experimental gene expression profiles with the introduced method. This analysis provides a worked-through procedure for model comparison and validation. We show how to select the number of clusters and how to evaluate the informativeness of clustering methods. In particular, we compare pairwise and correlation clustering and demonstrate that the former yields significant amount of additional (reliable) information. The results show consistency with BIC and wider applicability.

Experiment description. Conventional K-means approaches rely upon the definition of explicit metrics in the feature space. However, there exist numerous applications (such as gene expression analysis) in which the choice of a natural metric is far from obvious. Spectral clustering techniques have the potential to overcome this limitation, since they rely upon implicit metrics whose specification may be straightforward. In this experiment, the analyzed data set consists of gene expression profiles from *Mytilus galloprovincialis* female digestive gland [1]. Time points corresponds to 12 consecutive months, chosen to study how seasonal environmental changes affect physiology across the annual cycle. The first sample, which corresponds to January, is defined as reference. Logarithmic values are obtained for the 295 differentially expressed genes [1].

Taking advantage of the temporal structure of the data, the two object sets $\mathbf{O}^{(1)}$ and $\mathbf{O}^{(2)}$ are constructed by splitting the feature vector, i.e., the measurements for the months *(Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec)* are separated into the values for *(Mar, May, Jul, Sep, Nov)* and for *(Apr, Jun, Aug, Oct, Dec)*. This interleaved separation captures the statistical dependence of the samples due to time proximity. Thus it avoids the risk of undersampling small clusters of high biological relevance (as in this study) by having too few genes per cluster. Pearson correlation coefficients are calculated for each pair of genes in each set to construct the similarity matrices $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$.

Gene expression clustering. To obtain the approximation capacity, we fix the number of initial clusters at 10 and compute the mutual information at different temperatures. Figure 5 demonstrates the approximation capacity for the two models. In Figure 5(a) we have investigated the accuracy of the mean-field approximation by Gibbs sam-



(a) pairwise clustering (b) pairwise vs. correlation clustering (c) correlation matrix

Figure 7: Comparison of optimal pairwise and correlation clustering solutions. The objects are permuted according to pairwise clustering. In the color coding, the first bit refers to H_{ij}^{pc} and the second bit corresponds to H_{ij}^{cc}.

pling. Comparing the capacities shows the advantage of pairwise clustering over correlation clustering. This result means that under identical noise effects, pairwise clustering is able to discover more structure from data than correlation clustering. In this data set, ASC validates pairwise clustering ($\max_\beta \mathcal{I}_\beta = 1.03$) as 3.5 times more informative than correlation clustering ($\max_\beta \mathcal{I}_\beta = 0.272$).

Figure 5(c) shows the number of clusters identified by pairwise clustering at different temperatures. At the optimal temperature, 7 clusters are discovered by pairwise clustering. This diversity is in contrast to the 2 clusters identified by correlation clustering. Correlation clustering exhibits by construction a bias that favors clusters with equal sizes. In contrast, pairwise clustering is unbiased to size due to its shift-invariant property. Subcluster consistency is substantial: $6398/8231$ of the co-clustered pairs in pairwise clustering result co-clustered in correlation clustering as well.

To provide a more detailed analysis of the problem, two well known model order selection criteria are considered for comparison: the BIC score and the instability measure. BIC is a principle which is hard to apply in cases where the effective dimensionality is unclear. Such a situation arises for pairwise clustering and it is even less well defined in the case of correlation clustering. In the former model, the BIC score has been computed according to the effective number of dimensions, calculated as the ratio between the trace and the largest eigenvalue [9]. On the other hand, instability is a heuristic in the spirit of two-instance cross validation. It is applicable to both models but its generality remains confined to alternatives with comparable informativeness. BIC and instability constitute potentially inconsistent criteria, as apparent in Figures 6(a) and 6(c).

Figure 7 provides an in depth comparison study. Figure 7(a) shows the permutation of the objects based on the co-clustering induced by pairwise clustering at optimal temperature. Thereby the inferred clusters appear as diagonal blocks of the co-clustering matrix $\mathbf{H}^{pc}$. In Figure 7(b) the correlation clustering of the permuted objects, i.e. $\mathbf{H}^{cc}$, is compared with pairwise clustering. For the pair of objects i and j, the following encoding is used: i) *'yellow'* if $(H_{ij}^{pc}, H_{ij}^{cc}) = (1,1)$, ii) *'red'* if $(H_{ij}^{pc}, H_{ij}^{cc}) = (1,0)$, iii) *'light blue'* if $(H_{ij}^{pc}, H_{ij}^{cc}) = (0,1)$, and iv) *'dark blue'* if $(H_{ij}^{pc}, H_{ij}^{cc}) = (0,0)$. A substantial consistency between pairwise and correlation clusterings is observed. However, the pairwise clustering finds finer representations (more detailed and still validated structures) than correlation clustering that only identifies coarser structures. Figure 7(c) shows the correlation matrix for gene expressions.

5 Discussion

The introduced method addresses the task of model validation for spectral clustering. The optimal tradeoff between stability and informativeness is achieved by maximizing approximation capacity. Assignment potentials and resulting partition sums are computed either by embedding into a vector space (for pairwise clustering) or by mean-field approximation (for correlation clustering). In particular, this study showed the following properties:

1. self-consistency of ASC and the consistency with BIC (in contrast to the instability criterion),
2. greater generality of ASC in comparison to BIC,
3. applicability in the biological context of gene expression analysis,
4. three-fold higher informativeness of pairwise clustering in comparison to correlation clustering in a biological application.

Future work addresses generalizations of the introduced method to algorithms without an explicit cost function.

Acknowledgements

This work was partially supported by the DFG-SNF research cluster FOR916.

References

[1] Mohamed Banni, Alessandro Negri, Flavio Mignone, Hamadi Boussetta, Aldo Viarengo, and Francesco Dondero. Gene expression rhythms in the mussel Mytilus galloprovincialis (lam.) across an annual cycle. *PLoS ONE*, 6(5):e18904, 05 2011.

[2] Nikhil Bansal, Avrim Blum, and Shuchi Chawla. Correlation clustering. *Machine Learning*, 56:89–113, 2004.

[3] Shai Ben-David, Ulrike von Luxburg, and Dávid Pál. A sober look at clustering stability. In G. Lugosi and H.U. Simon, editors, *COLT'06, Pittsburgh, PA, USA*, pages 5–19, 2006.

[4] Joachim M. Buhmann. Information theoretic model validation for clustering. In *International Symposium on Information Theory, Austin Texas*, pages 1398 – 1402. IEEE, 2010.

[5] Joachim M. Buhmann. Context sensitive information: Model validation by information theory. In *Mexican Conference on Pattern Recognition*, pages 12–21, 2011.

[6] Kenneth P. Burnham and David R. Anderson. *Model selection and inference: a practical information-theoretic approach, 2nd ed.* Springer, New York, 2002.

[7] Sandrine Dudoit and Jane Fridlyand. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, 3(7), 2002.

[8] Thomas Hofmann and Joachim M. Buhmann. Pairwise data clustering by deterministic annealing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(1):1–14, 1997.

[9] Mark Kirkpatrick. Patterns of quantitative genetic variation in multiple dimensions. *Genetica*, 136:271–284, 2009.

[10] Herbert Kocha and Daniel Tataru. Well-posedness for the Navier-Stokes equations. *Advances in Mathematics*, 157(1):22 –35, 2001.

[11] Tilman Lange, Mikio Braun, Volker Roth, and Joachim M. Buhmann. Stability-based validation of clustering solutions. *Neural Computation*, 16(6):1299–1323, June 2004.

[12] Volker Roth, Julian Laub, Motoaki Kawanabe, and Joachim M. Buhmann. Optimal cluster preserving embedding of non-metric proximity data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(12), 2003.

[13] Gideon Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[14] Yevgeny Seldin and Naftali Tishby. Pac-bayesian analysis of co-clustering and beyond. *J. Mach. Learn. Res.*, 11:3595–3646, 2010.

[15] Noam Slonim, Gurinder Singh Atwal, Gasper Tracik, and William Bialek. Information-based clustering. *Proceedings of the National Academy of Science (PNAS)*, 102:18297–1830, 2005.

[16] Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.

[17] Vladimir N. Vapnik. *Statistical learning theory*. Wiley, New York, 1998.