International Symposium on Information Theory 2010 (ISIT10 E-Mo-4.2), June 13-18 in Austin, TX

Information theoretic model validation for clustering

Joachim M. Buhmann
Department of Computer Science
Swiss Federal Institute of Technology, ETH Zurich
Email: jbuhmann@inf.ethz.ch

***Abstract*—Model selection in clustering requires (i) to specify a suitable clustering principle and (ii) to control the model order complexity by choosing an appropriate number of clusters depending on the noise level in the data. We advocate an information theoretic perspective where the uncertainty in the measurements quantizes the set of data partitionings and, thereby, induces uncertainty in the solution space of clusterings. A clustering model, which can tolerate a higher level of fluctuations in the measurements than alternative models, is considered to be superior provided that the clustering solution is equally informative. This tradeoff between *informativeness* and *robustness* is used as a model selection criterion. The requirement that data partitionings should generalize from one data set to an equally probable second data set gives rise to a new notion of structure induced information.**

I. Introduction

Data clustering or data partitioning has emerged as the workhorse of *exploratory data analysis*. This unsupervised learning methodology comprises a set of data analysis techniques which group data into clusters by either optimizing a quality criterion or by directly employing a clustering algorithm. The zoo of models range from centroid based algorithms like k-means or k-medoids, spectral graph methods like Normalized Cut, Average Cut or Pairwise Clustering to linkage inspired grouping principles like Single Linkage, Average Linkage or Path-based Clustering.

The various clustering methods and algorithms ask for a unifying meta-principle how to choose the "right" clustering method dependent on the data source. This paper advocates a shift of viewpoint away from the problem *"What is the 'right' clustering model?"* to the question *"How can we algorithmically validate clustering models?"*. This conceptual shift roots in the assumption that ultimately, the data should vote for their prefered model type and model complexity[4]. Therefore, algorithms which are endowed with the ability to validate clustering concepts can maneuver through the space of clustering models and, dependent on the training and validation data sets, they can select a model with maximal information content and optimal robustness.

In this paper, we propose an information theoretic model validation strategy to select clustering models. A clustering model is used to generate a code for communication over a noisy channel. "Good" models are selected according to their robustness to noise. The approximation precision of clustering solutions is controlled by an algorithm called empirical risk approximation (ERA) [2] which quantizes the hypothesis class of clusterings. ERA employs an hypothetical communication framework where sets of approximate clustering solutions for the training and for the test data are used as a communication code. Approximations of the empirical minimizer with model averaging over approximate solutions favors stability of clusterings. Furthermore, it is well known that stability based model selection [8] yields highly satisfactory results in applications although the theoretical foundation of this model selection strategy is still controversially debated [1].

II. Statistical learning of clustering

Given are a **set of objects** $\mathbf{O} = \{o_1, \ldots, o_n\} \in \mathcal{O}$ and measurements $\mathbf{X} \in \mathcal{X}$ to characterize these objects. $\mathcal{O}, \mathcal{X}$ denotes the object or measurement space, respectively. Such measurements might be d-dimensional vectors $\mathbf{X} = \{X_i \in \mathbb{R}^d, 1 \leq i \leq n\}$ or relations $\mathbf{D} = (D_{ij}) \in \mathbb{R}^{n \cdot n}$ which describe the (dis)-similarity between object o_i and o_j. More complicated data structures than vectors or relations, e.g., three-way data or graphs, are used in various applications. In the following, we use the generic notation $\mathbf{X}$ for measurements. We have to distinguish between objects and measurements since repeated measurements might refer to the same object. Data denote object-measurement relations $\mathcal{O} \times \mathcal{X}$, e.g., vectorial data $\{X_i : 1 \leq i \leq n\}$ describe surjective relations between objects o_i and measurements $X_i := X(o_i)$.

The **hypotheses** for a clustering problem are the functions assigning data to groups, i.e.,

$$\begin{aligned} c : \mathcal{O} \times \mathcal{X} &\rightarrow \{1, \ldots, k\}^n \\ (\mathbf{O}, \mathbf{X}) &\mapsto c(\mathbf{O}, \mathbf{X}) \end{aligned} \tag{1}$$

The parameter $n = |\mathbf{O}|$ denotes the number of objects. In cases where $\mathbf{X}$ uniquely identifies the object set $\mathbf{O}$, i.e., there exists a bijective function between objects and measurements, then we omit the first argument of c to simplify notation. A clustering is then denoted by $c : \mathcal{X} \rightarrow \{1, \ldots, k\}^n$.

The **hypothesis class** for a clustering problem is defined as the set of functions assigning data to groups, i.e., $\mathcal{C}(\mathbf{X}) = \{c(\mathbf{O}, \mathbf{X}) : \mathbf{O} \in \mathcal{O}\}$. For n objects we can distinguish $O(k^n)$ such functions. Specific clustering models might require additional parameters θ which characterize a cluster, e.g., the centroids in k-means clustering. The hypothesis class is then the product space of possible assignments and possible parameter values.

III. Clustering costs and empirical risk approximation

Exploratory pattern analysis and model selection for grouping requires to assess the quality of clustering hypotheses. Various criteria emphasize coherency of data or connectedness, e.g., k-means clustering measures the average distance of data vectors to the nearest cluster centroid or prototype. For the subsequent discussion on information theoretic model validation, a cost or risk function $R(c, \mathbf{X})$ is assumed to measure how well a particular clustering with assignments $c(\mathbf{X})$ and cluster parameters θ groups the objects. To simplify the notation, cluster parameters θ are not explicitly listed as arguments of clustering costs but are subsumed in the specification of the cost function R. A suitable metric for the space of hypotheses might be chosen based on such a cost function R.

The clustering solution $c^{\perp}(\mathbf{X})$ minimizes the empirical risk (ERM) of data clustering given the measurements $\mathbf{X}$, i.e.,

$$c^{\perp}(\mathbf{X}) = \arg\min_{c} R(c, \mathbf{X}). \tag{2}$$

Clustering solutions which are similar in costs to the ERM solution $c^{\perp}(\mathbf{X})$ define the set $\mathcal{C}_{\gamma}(\mathbf{X})$ of empirical risk approximations for clustering, i.e.,

$$\mathcal{C}_{\gamma}(\mathbf{X}) := \{c(\mathbf{X}) \ : \ R(c, \mathbf{X}) \leq R(c^{\perp}, \mathbf{X}) + \gamma\}. \tag{3}$$

The set $\mathcal{C}_{\gamma}(\mathbf{X})$ reduces to the ERM solution in the limit $\lim_{\gamma \to 0} \mathcal{C}_{\gamma}(\mathbf{X}) = \{c^{\perp}(\mathbf{X})\}$.

To validate clustering methods we have to define and estimate the generalization performance of partitionings. We adopt the two sample set scenario with training and test data which is widely used in statistics and statistical learning theory [11] i.e. to bound the deviation of empirical risk from expected risk, but also for two-terminal systems in information theory [6]. We assume for the subsequent discussion that training data and test data are described by respective object sets $\mathbf{O}^{(1)}, \mathbf{O}^{(2)}$ and measurements $\mathbf{X}^{(1)}, \mathbf{X}^{(2)} \sim \mathbb{P}(\mathbf{X})$ which are drawn i.i.d. from the same probability distribution $\mathbb{P}(\mathbf{X})$. Furthermore, $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}$ uniquely identify the training and test object sets $\mathbf{O}^{(1)}, \mathbf{O}^{(2)}$ so that it is sufficient to list $\mathbf{X}^{(j)}$ as references to object sets $\mathbf{O}^{(j)}, j = 1, 2$.

Statistical inference requires that clustering solutions have to generalize from training data to test data since noise in the data renders the ERM solution $c^{\perp}(\mathbf{X}^{(1)}) \neq c^{\perp}(\mathbf{X}^{(2)})$ unstable. How can we evaluate the generalization properties of clustering solutions? Before we can evaluate the clustering costs $R(., \mathbf{X}^{(2)})$ on test data of the ERM clustering on training data $c^{\perp}(\mathbf{X}^{(1)})$ we have to identify a clustering $c \in \mathcal{C}(\mathbf{X}^{(2)})$ which corresponds to $c^{\perp}(\mathbf{X}^{(1)})$. A priori, it is not clear how to compare clusterings $c(\mathbf{X}^{(1)})$ for measurements $\mathbf{X}^{(1)}$ with clusterings $c(\mathbf{X}^{(2)})$ for measurements $\mathbf{X}^{(2)}$. Therefore, we define the mapping

$$\begin{aligned} \psi : \mathcal{C}(\mathbf{X}^{(1)}) &\rightarrow \mathcal{C}(\mathbf{X}^{(2)}) \\ c(\mathbf{X}^{(1)}) &\mapsto \psi \circ c(\mathbf{X}^{(1)}) \end{aligned} \tag{4}$$

which identifies a clustering hypothesis for training data $c \in \mathcal{C}(\mathbf{X}^{(1)})$ with a clustering hypothesis for test data $\psi \circ c \in \mathcal{C}(\mathbf{X}^{(2)})$. The reader should note that such a mapping ψ might change the object indices. In cases when the measurements are elements of an underlying metric space, then a natural choice for ψ is the nearest neighbor mapping $\nu(i) = \arg\min_{\ell} \|X_{\ell}^{(2)} - X_{i}^{(1)}\|^2$ where we identify clustering $c(\mathbf{X}^{(1)})$ with $\psi \circ c(\mathbf{X}^{(1)}) = \Big(c(X_{\nu(1)}^{(2)}), c(X_{\nu(2)}^{(2)}), \ldots, c(X_{\nu(n)}^{(2)})\Big)$.

The mapping ψ enables us to evaluate clustering costs on test data $\mathbf{X}^{(2)}$ for clusterings $c(\mathbf{X}^{(1)})$ selected on the basis of training data $\mathbf{X}^{(1)}$. Consequently, we can determine how many γ-optimal training solutions are also γ-optimal on test data, i.e., $\Delta\mathcal{C}(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}) := \big|\big(\psi \circ \mathcal{C}_{\gamma}(\mathbf{X}^{(1)})\big) \cap \mathcal{C}_{\gamma}(\mathbf{X}^{(2)})\big|$. A large overlap means that the training approximation set generalizes to the test data, whereas a small or empty intersection indicates the lack of generalization. Essentially, γ parametrizes a coarsening of the hypothesis class such that sets of data partitionings become stable w.r.t measurement fluctuations. The tradeoff between stability and informativeness is controlled by minimizing γ under the constraint of large $\Delta\mathcal{C}(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})/|\mathcal{C}_{\gamma}(\mathbf{X}^{(2)})|$ for given risk function $R(., \mathbf{X})$.

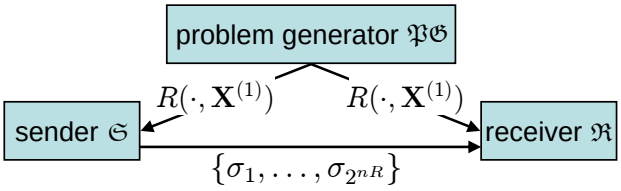

Fig. 1. Generation of a set of 2^{nR} code problems for communication by e.g. permuting the object indices.

IV. Coding by Approximation

In the following, we describe a communication scenario with a sender $\mathfrak{S}$, a receiver $\mathfrak{R}$ and a problem generator $\mathfrak{PG}$ where the problem generator serves as a noisy channel between sender and receiver. Communication takes place by approximately optimizing clustering cost functions, i.e., by calculating approximation sets $\mathcal{C}_{\gamma}(\mathbf{X}^{(1)}), \mathcal{C}_{\gamma}(\mathbf{X}^{(2)})$. This coding concept will be refered to as approximation set coding (ASC). The noisy channel is characterized by a clustering cost function $R(c, \mathbf{X})$ which determines the channel capacity of the ASC scenario. Validation and selection of clustering models is then achieved by maximizing the channel capacity over a set of cost functions $R_{\theta}(., \mathbf{X}), \theta \in \Theta$ where θ indexes the various clustering models.

Sender $\mathfrak{S}$ and receiver $\mathfrak{R}$ agree on a clustering principle $R(c, \mathbf{X}^{(1)})$ and on a mapping function ψ. The following procedure is then employed to generate the code for the communication process:

1) Sender $\mathfrak{S}$ and receiver $\mathfrak{R}$ obtain a data set $\mathbf{X}^{(1)}$ from the problem generator $\mathfrak{PG}$.
2) $\mathfrak{S}$ and $\mathfrak{R}$ calculate the γ-approximation set $\mathcal{C}_{\gamma}(\mathbf{X}^{(1)})$.
3) $\mathfrak{S}$ generates a set of (random) permutations $\Sigma := \{\sigma_1, \ldots, \sigma_{2^{nR}}\}$ to rename the objects. The permutations define a set of optimization problems $R(c, \sigma_j \circ \mathbf{X}^{(1)})$

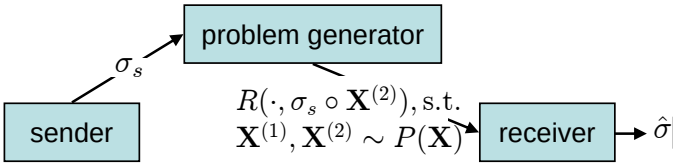

Fig. 2. Communication process: (1) the sender selects transformation σ_s, (2) the problem generator draws $\mathbf{X}^{(2)} \sim \mathbb{P}(\mathbf{X})$ and applies σ_s to it, and the receiver estimates $\sigma^\star$ based on $\tilde{\mathbf{X}} = \sigma_s \circ \mathbf{X}^{(2)}$.

with associated approximation sets $\mathcal{C}_\gamma(\sigma_j \circ \mathbf{X}^{(1)}), 1 \leq j \leq 2^{nR}$.

4) $\mathfrak{S}$ sends the set of permutations Σ to $\mathfrak{R}$ who determines the approximation sets $\{\mathcal{C}_\gamma(\sigma_i \circ \mathbf{X}^{(1)})\}_{i=1}^{2^{nR}}$.

The rationale behind this procedure is the following: Given the measurements $\mathbf{X}^{(1)}$ the sender has randomly covered the set of clusterings $\mathcal{C}(\mathbf{X}^{(1)})$ by respective approximation sets $\{\mathcal{C}(\sigma_i \circ \mathbf{X}^{(1)}) : 1 \leq i \leq 2^{nR}\}$. Communication succeeds if the approximation sets are stable under the stochastic fluctuations of the measurements. The criterion for reliable communication is defined by the ability of the receiver to identify a specific permutation that has been selected by the sender. The approximation sets $\mathcal{C}(\sigma_i \circ \mathbf{X}^{(1)})$ play the role of codebook vectors in Shannon's theory of communication.

After this setup procedure, both sender and receiver have a list of approximation sets available or can algorithmically determine membership of clusterings in one of the 2^{nR} approximation sets.

How is the communication between sender and receiver organized? During communication, the following steps take place as depicted in fig. 2:

1) The sender $\mathfrak{S}$ selects a permutation σ_s as message and send it to the problem generator $\mathfrak{PG}$.
2) $\mathfrak{PG}$ generates a new data set $\mathbf{X}^{(2)}$ and it applies the selected permutation to $\mathbf{X}^{(2)}$, yielding $\tilde{\mathbf{X}} = \sigma_s \circ \mathbf{X}^{(2)}$.
3) $\mathfrak{PG}$ send $\tilde{\mathbf{X}}$ to the receiver $\mathfrak{R}$ without revealing σ_s.
4) $\mathfrak{R}$ calculates the approximation set $\mathcal{C}_\gamma(\tilde{\mathbf{X}})$
5) $\mathfrak{R}$ estimates the applied permutation σ_s by using the decoding rule

$$\hat{\sigma} = \arg\max_{\sigma \in \Sigma} \left| \left(\psi \circ \mathcal{C}_\gamma(\sigma \circ \mathbf{X}^{(1)}) \right) \cap \mathcal{C}_\gamma(\tilde{\mathbf{X}}) \right| \qquad (5)$$

This communication channel supports to communicate at most $n \log k$ nats if two conditions hold: (i) the channel is noise free $\mathbf{X}^{(1)} \equiv \mathbf{X}^{(2)}$; (ii) all clusters have the same number of objects assigned to.

It is worth mentioning that ASC is conceptually not restricted to clustering problems although we focus the discussion here to this problem domain.

V. Error analysis of Approximation Set Coding

To determine the optimal approximation precision for an optimization problem $R(., \mathbf{X})$ we have to determine necessary and sufficient conditions which have to hold in order to reliably identify approximation sets. Reliable identification of approximation sets enable us to define a communication protocol using the above described coding scheme. Therefore, we analyse the error probability of *approximation set coding* and the channel capacity which is associated with a particular cost function $R(., \mathbf{X})$. This channel capacity will be refered to as *approximation capacity* since it determines the approximation precision of the coding scheme.

A communication error occurs if the sender selects σ_s and the receiver decodes $\hat{\sigma} = \sigma_j, j \neq s$. To estimate the probability of this event, we introduce the sets

$$\Delta\mathcal{C}_j := \left(\psi \circ \mathcal{C}_\gamma(\sigma_j \circ \mathbf{X}^{(1)}) \right) \cap \mathcal{C}_\gamma(\tilde{\mathbf{X}}^{(2)}), \ \sigma_j \in \Sigma. \qquad (6)$$

The set $\Delta\mathcal{C}_j$ measures the intersection between the approximation set $\mathcal{C}_\gamma(\sigma_j \circ \mathbf{X}^{(1)})$ for σ_j-permuted measurements and the approximation set which has been calculated by the receiver based on the test data $\tilde{\mathbf{X}}$.

The probability of a communication error is given by a substantial overlap $\Delta\mathcal{C}_j$ with $\sigma_j \in \Sigma \setminus \{\sigma_s\}$, i.e.,

$$\begin{aligned} \mathbb{P}(\hat{\sigma} \neq \sigma_s | \sigma_s) &= \mathbb{P}\left(\max_{\sigma_j \in \Sigma \setminus \{\sigma_s\}} |\Delta\mathcal{C}_j| \geq |\Delta\mathcal{C}_s| \middle| \sigma_s \right) \\ &\leq \sum_{\sigma_j \in \Sigma \setminus \{\sigma_s\}} \mathbb{P}\left(|\Delta\mathcal{C}_j| \geq |\Delta\mathcal{C}_s| \middle| \sigma_s \right) \qquad (7) \\ &= \sum_{\sigma_j \in \Sigma \setminus \{\sigma_s\}} \mathbb{E}_{\mathbf{X}^{(1,2)}} \mathbb{E}_{\sigma_j} \left[\mathbb{I}_{\{|\Delta\mathcal{C}_j| \geq |\Delta\mathcal{C}_s|\}} \middle| \sigma_s \right] \end{aligned}$$

The notation $\mathbf{X}^{(1,2)} = (\mathbf{X}^{(1)}, \mathbf{X}^{(2)})$ and $\mathbb{I}_{\{f\}} = \begin{cases} 1 & f \text{ is true} \\ 0 & \text{otherwise} \end{cases}$ is used. The inequality in (7) is caused by the union bound. The confusion probability with message $\sigma_j, j \neq s$ for given training data $\mathbf{X}^{(1)}$ and test data $\mathbf{X}^{(2)}$ conditioned on σ_s is defined by

$$\begin{aligned} \mathbb{E}_{\sigma_j}\left[\mathbb{I}_{\{|\Delta\mathcal{C}_j| \geq |\Delta\mathcal{C}_s|\}} \right] &= \frac{1}{|\{\sigma_j\}|} \sum_{\{\sigma_j\}} \mathbb{I}_{\{\log|\Delta\mathcal{C}_j| \geq \log|\Delta\mathcal{C}_s|\}} \\ &\overset{(a)}{\leq} \sum_{\{\sigma_j\}} \frac{\exp\left(\log|\Delta\mathcal{C}_j| - \log|\Delta\mathcal{C}_s|\right)}{|\{\sigma_j\}|} \\ &= \frac{1}{|\{\sigma_j\}|} \sum_{\{\sigma_j\}} \frac{|\Delta\mathcal{C}_j|}{|\Delta\mathcal{C}_s|} \\ &\overset{(b)}{=} \frac{|\mathcal{C}_\gamma(\mathbf{X}^{(1)})||\mathcal{C}_\gamma(\mathbf{X}^{(2)})|}{|\{\sigma_j\}||\Delta\mathcal{C}_s|} \\ &\overset{(c)}{=} \exp\left(-n\mathcal{I}_\gamma(\sigma_j, \hat{\sigma})\right) \qquad (8) \end{aligned}$$

The expectation $\mathbb{E}_{\sigma_j}\left[\mathbb{I}_{\{|\Delta\mathcal{C}_j| \geq |\Delta\mathcal{C}_s|\}} \right]$ in derivation (8) is conditioned on σ_s which has been omitted to increase the readability of the formulas. The summation $\{\sigma_j\}$ is indexed by all possible realizations of the transformation σ_j that are uniformly selected. (a) we have used the inequality $\mathbb{I}_{\{x \geq 0\}} \leq \exp(x)$; (b) averaging over a random permutation σ_j of object indices breaks any statistical dependence between sender and receiver approximation sets which corresponds to the error case in jointly typical coding [5]; (c) we have introduced the mutual information between the uniform distribution of the sender

message σ_j and the receiver message $\hat{\sigma}$

$$\begin{aligned} \mathcal{I}_\gamma(\sigma_j, \hat{\sigma}) &= \frac{1}{n} \log \left(\frac{|\{\sigma_j\}||\Delta\mathcal{C}_s|}{|\mathcal{C}_\gamma^{(1)}||\mathcal{C}_\gamma^{(2)}|} \right) \\ &= \frac{1}{n} \left(\log \frac{|\{\sigma_j\}|}{|\mathcal{C}_\gamma^{(1)}|} + \log \frac{|\mathcal{C}^{(2)}|}{|\mathcal{C}_\gamma^{(2)}|} - \log \frac{|\mathcal{C}^{(2)}|}{|\Delta\mathcal{C}_s|} \right) \end{aligned} \quad (9)$$

To compactify the formula, the following notation is introduced: $\mathcal{C}^{(i)} := \mathcal{C}(\mathbf{X}^{(i)}), \mathcal{C}_\gamma^{(i)} := \mathcal{C}_\gamma(\mathbf{X}^{(i)}), i = 1, 2$. The interpretation of eq. (9) is straightforward: The first logarithm measures the entropy of the number of transformations which can be resolved with an uncertainty of $\mathcal{C}_\gamma^{(1)}$ in the space of clusterings on the sender side. The logarithm $\log(|\mathcal{C}^{(2)}|/|\mathcal{C}_\gamma^{(2)}|)$ calculates the entropy of the receiver clusterings which are quantized by $\mathcal{C}_\gamma^{(2)}$. The third logarithm measures the joint entropy of $(\sigma_j, \hat{\sigma})$ which depends on the size of the intersection $|\Delta\mathcal{C}_s| = |\left(\psi \circ \mathcal{C}_\gamma(\sigma_s \circ \mathbf{X}^{(1)})\right) \cap \mathcal{C}_\gamma(\sigma_s \circ \mathbf{X}^{(2)})|$.

Inserting (8) into (7) yields the upper bound for the error probability

$$\begin{aligned} \mathbb{P}(\hat{\sigma} \neq \sigma_s | \sigma_s) &\leq \exp(nR \log 2) \exp\left(-n\mathcal{I}_\gamma(\sigma_j, \hat{\sigma})\right) \\ &= \exp(-n(\mathcal{I}_\gamma(\sigma_j, \hat{\sigma}) - R \log 2)) \end{aligned} \quad (10)$$

The communication rate $nR \log 2$ is limited by the mutual information $\mathcal{I}_\gamma(\sigma_j, \hat{\sigma})$ for asymptotically error-free communication.

VI. Information theoretical model selection

The analysis of the error probability suggests the following inference principle for model selection: the approximation precision is controlled by γ which has to be minimized to derive more expressive clusterings. For large γ the rate R will be low since we resolve the space of clusterings in only a coarse grained fashion. For too small γ the error probability does not vanish which indicates confusions between σ_j and σ_s. The optimal γ-value is given by the smallest γ or, equivalently the highest approximation precision

$$\gamma^\star = \arg \max_{\gamma \in [0,\infty)} \mathcal{I}_\gamma(\sigma, \hat{\sigma}). \quad (11)$$

Another choice to be made in modeling is to select a suitable cost function for clustering $R(., \mathbf{X})$. Let us assume that a number of cost functions $\{R_1(., \mathbf{X}), R_2(., \mathbf{X}), \ldots, R_m(., \mathbf{X})\}$ are considered as candidates. The cost function to be selected is

$$R^\star(c, \mathbf{X}) = \arg \max_{1 \leq j \leq m} \mathcal{I}_\gamma(\sigma(R_j), \hat{\sigma}(R_j)) \quad (12)$$

where both the random variables σ and $\hat{\sigma}$ depend on $R(c, \mathbf{X})$. The selection rule (12) prefers the model which is "expressive" enough to exhibit high information content (e.g., many clusters) and, at the same time robustly resists to noise in the data set. The bits or nats which are measured in the ASC communication setting are context sensitive since they refer to a hypothesis class $\mathcal{C}(\mathbf{X})$, i.e., how finely or coarsely functions can be resolved in $\mathcal{C}$.

VII. Computation of the approximation capacity

To estimate the mutual information $\mathcal{I}_\gamma(\sigma, \hat{\sigma})$ computationally, we have to calculate the size of the sets $|\mathcal{C}_\gamma(\mathbf{X}^{(1)})|$, $|\mathcal{C}_\gamma(\mathbf{X}^{(2)})|$, $|\{\sigma_j\}|$, $|\Delta\mathcal{C}_s|$.

The cardinality $|\{\sigma_j\}|$ is determined by the type of the empirical minimizer $c^\perp(\mathbf{X})$, i.e., the probabilities $\overline{p}_\nu := \mathbb{P}(c^\perp(\mathbf{X}^{(1)}) = \nu), 1 \leq \nu \leq k$ with

$$|\{\sigma_j\}| \doteq \exp(n\mathcal{H}(\overline{p}_1, \ldots, \overline{p}_k)) \quad (13)$$

where $\mathcal{H}(\overline{p}_1, \ldots, \overline{p}_k) = -\sum_{\nu=1}^k \overline{p}_\nu \log \overline{p}_\nu$ denotes the entropy of the type of $c^\perp(\mathbf{X}^{(1)})$, ($a_n \doteq b_n \Leftrightarrow \lim_{n\to\infty} \frac{1}{n} \log \frac{a_n}{b_n} = 0$).

The cardinality of the approximation sets can be estimated estimated using concepts from statistical physics. The approximation sets $\mathcal{C}_\gamma(\mathbf{X}^{(1,2)})$ are known as microcanonical ensembles in statistical mechanics. Estimating their size is achieved up to logarithmic corrections by calculating the partition function

$$|\mathcal{C}_\gamma(\mathbf{X}^{(1,2)})| \doteq \sum_{c \in \mathcal{C}(\mathbf{X}^{(1,2)})} \exp(-\beta R(c, \mathbf{X}^{(1,2)})). \quad (14)$$

The scaling factor β, also know as inverse computational temperature, is determined such that the average costs of the ensemble $\mathcal{C}_\gamma(\mathbf{X}^{(1)})$ yields $R(c^\perp, \mathbf{X}^{(1)}) + \gamma$. The weights $\exp(-\beta R(c, \mathbf{X}^{(1,2)}))$ are known as Boltzmann factors.

The joint entropy in the mutual information, which is related to the intersection

$$\begin{aligned} |\Delta\mathcal{C}| &= \left| \left(\psi \circ \mathcal{C}_\gamma(\mathbf{X}^{(1)})\right) \cap \mathcal{C}_\gamma(\mathbf{X}^{(2)}) \right| \\ &= \sum_{c \in \mathcal{C}(\mathbf{X}^{(2)})} \mathbb{I}_{\{c \in \psi \circ \mathcal{C}_\gamma(\mathbf{X}^{(1)})\}} \mathbb{I}_{\{c \in \mathcal{C}_\gamma(\mathbf{X}^{(2)})\}} \\ &\doteq \sum_{c \in \mathcal{C}(\mathbf{X}^{(2)})} \exp(-\beta R(\psi^{-1} \circ c, \mathbf{X}^{(1)})) \cdot \\ &\qquad \exp(-\beta R(c, \mathbf{X}^{(2)})), \end{aligned} \quad (15)$$

involves a product of Boltzmann factors.

The identification of approximation sets with microcanonical ensembles provides access to a rich source of computational and analytical methods from statistical physics to calculate the mutual information $\mathcal{I}_\gamma(\sigma, \hat{\sigma})$. This analogy is by no means accidental since information theory and statistical mechanics are both specializations of empirical process theory with large deviation analysis of many particle systems. The central role of entropy and free energy is reflected in ASC coding where the logarithm of the partition function arises in the mutual information (9) twice.

The cardinalities of the approximation sets can also be numerically estimated by sampling using Markov Chain Monte Carlo methods or by employing analytical techniques like deterministic annealing [9], [3].

VIII. Why information theory for clustering validation?

There exists a long history of information theoretic approaches to model selection, which traces back at least to

Akaike's extension of the Maximum Likelihood principle. AIC penalizes fitted models by twice the number of free parameters. The Bayesian Information Criterion (BIC) suggests a stronger penalty than AIC, i.e., number of model parameters times logarithm of the number of samples. Rissanen's minimum description length principles is closely related to BIC (see e.g. [7] for model selection penalties). Tishby et al [10] proposed to select the number of clusters according to a difference of mutual informations which is closely related to rate distortion theory with side information.

All these information criteria regularize model estimation of the data source. Approximation set coding pursues a different strategy for the following reason: Quite often the measurement space $\mathcal{X}$ has a much higher "dimension" than the solution space. Consider for example the problem of spectral clustering with k groups based on dissimilarities $\mathbf{D}$: The measurements are elements of $\mathbb{R}^{n(n-1)/2}$ for real valued, symmetric weights with vanishing self-dissimilarities, but we can at most distinguish $O(k^n)$ different clusterings. Any approach which relies on estimating the probability distribution $\mathbb{P}(\mathbf{X})$ of the data ultimately will fail since we require far too many observations than needed to identify one hypothesis or a set of hypotheses, i.e., one clustering or a set of clusterings.

Using an information theoretic perspective, we might ask the question how the uncertainty in the measurements reduces the resolution in the hypothesis class. How similar can two hypotheses be so that they are still statistically distinguishable given a cost function $R(c, \mathbf{X})$? This research program is based on the idea that approximation sets of clustering cost functions can be used as a reliable code. The capacity of such a coding scheme then answers the question how sensitive a particular cost function is to data noise.

IX. Conclusion

Model selection and validation requires to estimate the generalization ability of models from training to test data. "Good" models show a high expressiveness and they are robust w.r.t. noise in the data. This tradeoff between *informativeness* and *robustness* ranks different models when they are tested on new data and it quantitatively describes the underfitting/overfitting dilemma. In this paper we have explored the idea to use approximation sets of clustering solutions as a communication code. Since clustering solutions with k clusters can be represented as strings of n symbols with a k-ary alphabet, the significant problem of model order selection in clustering can be naturally phrased as a communication problem. The *approximation capacity* of a cost function provides a selection criterion which renders various models comparable in terms of their respective bit rates. The number of reliably extractable bits of a clustering cost function $R(., \mathbf{X})$ define a "task sensitive information measure" since it only accounts for the fluctuations in the data $\mathbf{X}$ which actually have an influence on identifying an individual clustering solution or a set of clustering solutions.

The maximum entropy inference principle suggests that we should average over the statistically indistinguishible solutions in the optimal approximation set $\mathcal{C}_{\gamma\star}(\mathbf{X})$. Such a model averaging strategy replaces the original cost function with the free energy and, thereby, it defines a continuation methods with maximal robustness. The urgent question in many data analysis applications, which regularization term should be used without introducing an unwanted bias, is naturally answered by the entropy. The second question, how the regularization parameter should be selected, in answered by ASC: Choose the parameter value which maximizes the approximation capacity!

ASC for model selection can be applied to all combinatorial or continuous optimization problems which depend on noisy data. The noise level is characterized by two samples $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}$. Two samples provide by far too little information to estimate the probability density of the measurements but two large samples contain sufficient information to determine the uncertainty in the solution space. The equivalence of ensemble averages and time averages of ergodic systems is heavily exploited in statistical mechanics and it also enables us in this paper to derive a model selection strategy based on two samples.

Future work also includes the study of algorithmic complexity issues. The question how hard are properly regularized optimization problems hints at a relationship between computational complexity and statistical complexity.

Acknowledgment

The author appreciated valuable and insightful discussions with S. Ben-David, T. Lange, F. Pla, V. Roth and N. Tishby. This work has been partially supported by the DFG-SNF research cluster FOR916 and by the FP7 EU project SIMBAD.

References

[1] Shai Ben-David, Ulrike von Luxburg, and Dávid Pál. A sober look at clustering stability. In G. Lugosi and H.U. Simon, editors, *Learning Theory, Proceedings of 19th Annual Conference on Learning Theory, COLT 2006, Pittsburgh, PA, USA*, volume 4005 of *LNAI*, pages 5–19. Springer-Verlag Berlin Heidelberg, 2006.

[2] Joachim M. Buhmann. Empirical risk approximation: An induction principle for unsupervised learning. Technical Report IAI-TR-98-3, Department of Computer Science III / University of Bonn, 1998.

[3] Joachim M. Buhmann and Hans Kühnel. Vector quantization with complexity costs. *IEEE Transactions on Information Theory*, 39(4):1133–1145, July 1993.

[4] Gerda Claeskens and Nils Lid Hjort. *Model Selection and Model Averaging*. Cambridge University Press, 2008.

[5] Thomas M. Cover and Jay A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 2nd edition, 2006.

[6] Imre Csiczár and Janos Körner. *Information Theory: Coding theorems for discrete memoryless systems*. Academic Press, New York, San Francisco, London, 1981.

[7] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Verlag, New York, 2008.

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

[9] K. Rose, E. Gurewitz, and G. Fox. Vector quantization by deterministic annealing. *IEEE Transactions on Information Theory*, 38(4):1249–1257, 1992.

[10] 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.

[11] Vladimir N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, Berlin, Heidelberg, 1982.