Information Theoretic Model Selection for Pattern Analysis

Author(s):
Buhmann, Joachim M.; Haghir Chehreghani, Morteza; Frank, Mario; Streich, Andreas Peter

Publication date:
2011

Permanent link:
https://doi.org/10.3929/ethz-a-006611809

JMLR: Workshop and Conference Proceedings 7 (2011) 1–8 Workshop on Unsupervised and Transfer Learning

Information Theoretic Model Selection for Pattern Analysis

Joachim M. Buhmann JBUHMANN@INF.ETHZ.CH
Morteza Haghir Chehreghani MORTEZA.CHEHREGHANI@INF.ETHZ.CH
Mario Frank MARIO.FRANK@INF.ETHZ.CH
Andreas P. Streich ANDREAS.STREICH@ALUMNI.ETHZ.CH
Department of Computer Science, ETH Zurich, Switzerland

Editor: I. Guyon, G. Dror, V. Lemaire, G. Taylor, and D. Silver

Abstract

Exploratory data analysis requires (i) to define a set of patterns hypothesized to exist in the data, (ii) to specify a suitable quantification principle or cost function to rank these patterns and (iii) to validate the inferred patterns. For data clustering, the patterns are object partitionings into k groups; for PCA or truncated SVD, the patterns are orthogonal transformations with projections to a low-dimensional space. We propose an information theoretic principle for model selection and model-order selection. Our principle ranks competing pattern cost functions according to their ability to extract context sensitive information from noisy data with respect to the chosen hypothesis class. Sets of approximative solutions serve as a basis for a communication protocol. Analogous to Buhmann (2010), inferred models maximize the so-called approximation capacity that is the mutual information between coarsened training data patterns and coarsened test data patterns. We demonstrate how to apply our validation framework by the well-known Gaussian mixture model.

Keywords: Unsupervised learning, data clustering, model selection, information theory, maximum entropy, approximation capacity

1. Model Selection via Coding

Model selection and model order selection [Burnham and Anderson (2002)] are fundamental problems in pattern analysis. A variety of models and algorithms has been proposed to extract patterns from data, i.e., for clustering, but a comprehensive theory on how to choose the "right" pattern model given the data is still missing. Statistical learning theory as in Vapnik (1998) advocates to measure the generalization ability of models and to employ the prediction error as a measure of model quality. In particular, stability analysis of clustering solutions has shown very promising results for model order selection in clustering [Dudoit and Fridlyand (2002); Lange et al. (2004)], although discussed controversially by Ben-David et al. (2006). Stability is, however, only one aspect of statistical modeling, e.g., for unsupervised learning. The other aspect of the modeling tradeoff is characterized by the informativeness of the extracted patterns. A tolerable decrease in the stability of inferred patterns in the data (data model) might be compensated by a substantial increase of their information content (see also the discussion in Tishby et al. (1999)).

We formulate a principle that balances these two antagonistic objectives by transforming the model selection problem into a generic communication scenario. Thereby, the objective

function or cost function that maps patterns to quality scores is considered as a noisy channel. Different objectives are ranked according to their transmission properties and the cost function with the highest channel capacity is then selected as the most informative model for a given data set. Thereby, we generalize the set-based coding scheme proposed by Buhmann (2010) in order to simplify the embedding of pattern inference problems in a communication framework. Learning, in general, resembles communication from a conceptual viewpoint: For *communication*, one demands a high rate (a large amount of information transferred per channel use) together with a decoding rule that is stable under the perturbations of the messages by the noise in the channel. For *learning* patterns in data, the data analyst favors a rich model with high complexity (e.g., a large number of clusters in grouping), while the generalization error of test patterns is expected to remain stable and low. We require that solutions of pattern analysis problems are reliably inferred from noisy data.

This article first develops the information theoretic framework of weighted approximation set coding (wASC) for validating statistical models. We then demonstrate for the first time, how to practically transform a pattern recognition task to a communication setting and how to compute the capacity of clustering solutions. The feasibility of wASC is demonstrated for mixture models and real world data.

2. Brief Introduction to Approximation Set Coding

In this section, we briefly describe the theory of weighted Approximation Set Coding (wASC) for pattern analysis as proposed by Buhmann (2010).

Let $\mathbf{X} = \{X_1, \ldots, X_n\} \in \mathcal{X}$ be a set of n objects $\mathbf{O}$ and n measurements in a data space $\mathcal{X}$, where the measurements characterize the objects. Throughout the paper, we assume the special case of a bijective map between objects and measurements, i.e., the i^{th} object is synonymous with the vector $\mathbf{x}_i \in \mathbb{R}^D$. In general, the (object, measurement) relation might be more complex than an object-specific feature vector. A **hypothesis**, i.e. a solution of a pattern analysis problem, is a function c that assigns objects (e.g. data) to patterns of a pattern space $\mathcal{P}$:

$$c \; : \; \mathcal{X} \to \mathcal{P}, \qquad \mathbf{X} \mapsto c(\mathbf{X}). \tag{1}$$

Accordingly, the **hypothesis class** is the set of all such functions, i.e. $\mathcal{C}(\mathbf{X}) := \{c(\mathbf{X}) : \mathbf{X} \in \mathcal{X}\}$. For clustering, the patterns are object partitionings $\mathcal{P} = \{1, \ldots, k\}^n$. A model for pattern analysis is characterized by a **cost/objective function** $R(c, \mathbf{X})$ that assigns a real value to a pattern $c(\mathbf{X})$. To simplify the notation, model parameters $\boldsymbol{\theta}$ (e.g., centroids) are not explicitly listed as arguments of the objective function. Let $c^{\perp}(\mathbf{X})$ be the pattern that minimizes the cost function, i.e. $c^{\perp}(\mathbf{X}) := \arg\min_c R(c, \mathbf{X})$. For normalization purposes we assume that the minimal costs $R(c^{\perp}, \mathbf{X}) = 0$ vanish. As the measurements $\mathbf{X}$ are random variables, the global minimum $c^{\perp}(\mathbf{X})$ of the empirical costs is a random variable as well. Let $\mathbf{X}^{(q)}, q \in \{1, 2\}$, be two datasets with the same inherent structure but different noise instances. In most cases, their global minima differ, i.e. $c^{\perp}(\mathbf{X}^{(1)}) \neq c^{\perp}(\mathbf{X}^{(2)})$. In order to rank all solutions of the pattern analysis problem, we introduce *approximation weights*

$$w : \mathcal{C} \times \mathcal{X} \times \mathbb{R}_+ \to [0, 1]\,, \qquad (c, \mathbf{X}, \beta) \mapsto w_\beta(c, \mathbf{X})\,. \tag{2}$$

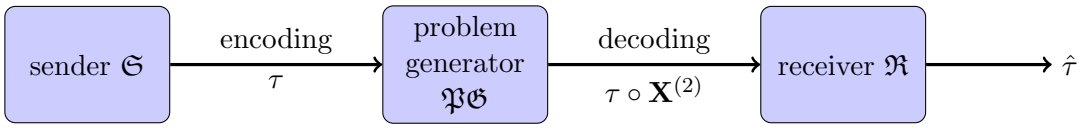


Figure 1: Communication process: (1) the sender selects transformation τ, (2) the problem generator draws $\mathbf{X}^{(2)} \sim \mathbb{P}(\mathbf{X})$ and applies τ to it, and (3) the receiver estimates $\hat{\tau}$ based on $\tilde{\mathbf{X}} = \tau \circ \mathbf{X}^{(2)}$.

Thereby, we require that i) weights should be non-negative and ii) solutions with lower costs should have a larger weight, i.e.,

$$R(c, \mathbf{X}) \leq R(\tilde{c}, \mathbf{X}) \quad \Longleftrightarrow \quad w_\beta(c, \mathbf{X}) \geq w_\beta(\tilde{c}, \mathbf{X}) \; . \tag{3}$$

The family of (Boltzmann) weights $w_\beta(c, \mathbf{X}) := \exp(-\beta R(c, \mathbf{X}))$, parameterized by the inverse computational temperature β, fulfils these requirements. These weights define the two *weight sums* $\mathcal{Z}_q$ and the *joint weight sum* $\mathcal{Z}_{12}$

$$\mathcal{Z}_q := \mathcal{Z}(\mathbf{X}^{(q)}) \;\; = \sum_{c \in \mathcal{C}(\mathbf{X}^{(q)})} \exp(-\beta R(c, \mathbf{X}^{(q)})), \; q = 1, 2 \tag{4}$$

$$\mathcal{Z}_{12} := \mathcal{Z}(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}) \;\; = \sum_{c \in \mathcal{C}(\mathbf{X}^{(2)})} \exp(-\beta(R(c, \mathbf{X}^{(1)}) + R(c, \mathbf{X}^{(2)}))), \tag{5}$$

where $\exp(-\beta(R(c, \mathbf{X}^{(1)}) + R(c, \mathbf{X}^{(2)})))$ measures how well a solution c minimizes costs on *both* datasets. The sums (4,5) play a central role in our framework. If $\beta = 0$, all weights $w_\beta(c, \mathbf{X}) = 1$ are independent of the costs. In this case, $\mathcal{Z}_q = |\mathcal{C}(\mathbf{X}^{(q)})|$ indicates the size of the hypothesis space, and $\mathcal{Z}_{12} = \mathcal{Z}_1 = \mathcal{Z}_2$. For high β, all weights are small compared to the weight $w_\beta(c^\perp, \mathbf{X}^{(q)})$ of the global optimum and the weight sum essentially counts the number of globally optimal solutions. For intermediate β, $\mathcal{Z}(\cdot)$ takes a value between 0 and $|\mathcal{C}(\mathbf{X}^{(q)})|$, giving rise to the interpretation of $\mathcal{Z}(\cdot)$ as the effective number of patterns that approximately fit the dataset $\mathbf{X}^{(q)}$, where β defines the precision of this approximation. Essentially, $\mathcal{Z}_q$ counts all statistically indistinguishable data patterns that approximate the minimum of the objective function. The global optimum $c^\perp(\mathbf{X})$ can change whenever we optimize on another random subset of the data, whereas, for a well-tuned β, $\mathcal{Z}(\mathbf{X})$ remains approximately invariant. Therefore, $\mathcal{Z}(\mathbf{X})$ defines the resolution of the hypothesis class that is relevant for inference. Noise in the measurements $\mathbf{X}$ reduces this resolution and thus coarsens the hypothesis class. As a consequence, the key problem of learning is to control the resolution optimally: How high can β be chosen to still ensure identifiability of $\mathcal{Z}(\mathbf{X})$ in the presence of data fluctuations? Conversely, choosing β too low yields a too coarse resolution of solutions and does not capture the maximal amount of information in the data.

We answer this key question by means of a communication scenario. The communication architecture includes a *sender* $\mathfrak{S}$ and a *receiver* $\mathfrak{R}$ with a *problem generator* $\mathfrak{PG}$ between the two terminals $\mathfrak{S}$, $\mathfrak{R}$ (see Fig. 1). The communication protocol is organized in two stages: (i) design of a communication code and (ii) the communication process.

For the communication code, we adapt Shannon's random coding scenario, where a codebook of random bit strings covers the space of all bit strings. The sender sends a bit string and the receiver observes a perturbed version of this bit string. For decoding, the receiver has to find the most similar codebook vector in the codebook which is the decoded message. In the same spirit, for our scenario, the sender must communicate patterns to the receiver via noisy datasets. Since we are interested in patterns with low costs, the optimal pattern $c^{\perp}(\mathbf{X}^{(1)})$ can serve as a message. The other patterns in the codebook are generated by transforming the training data $\tau \circ \mathbf{X}^{(1)}$ with the transformation $\tau \in \mathbb{T} := \{\tau_1, ..., \tau_{2^{n\rho}}\}$. The number of codewords is $2^{n\rho}$ and ρ is the rate of the protocol. The choice of such transformations depends on the hypothesis class and they have to be equivariant, i.e., the transformed optimal pattern equals the optimal pattern of the transformed data $\tau \circ c(\mathbf{X}^{(1)}) = c(\tau \circ \mathbf{X}^{(1)})$. In data clustering, *permuting* the indices of the objects defines the group of transformations to cover the pattern space. Each clustering solution $c \in \mathcal{C}(\mathbf{X}^{(1)})$ can be transformed into another solution by a permutation τ on the indices of c.

To communicate, $\mathfrak{S}$ selects a transformation $\tau_s \in \mathbb{T}$ and sends it to a *problem generator* $\mathfrak{PG}$ as depicted in Fig. 1. $\mathfrak{PG}$ then generates a new dataset $\mathbf{X}^{(2)}$, applies the transformation τ_s, and sends the resulting data $\tilde{\mathbf{X}} := \tau_s \circ \mathbf{X}^{(2)}$ to $\mathfrak{R}$. On the receiver side, the lack of knowledge on the transformation τ_s is mixed with the stochastic variability of the source generating the data $\mathbf{X}$. $\mathfrak{R}$ has to estimate the transformation $\hat{\tau}$ based on $\tilde{\mathbf{X}}$. The decoding rule of $\mathfrak{R}$ selects the pattern transformation $\hat{\tau}$ that yields the highest joint weight sum of $\hat{\tau} \circ \mathbf{X}^{(1)}$ and $\tilde{\mathbf{X}}$

$$\hat{\tau} = \arg\max_{\tau \in \mathbb{T}} \sum_{c \in \mathcal{C}(\mathbf{X}^{(1)})} \exp(-\beta(R(c, \tau \circ \mathbf{X}^{(1)}) + R(c, \tilde{\mathbf{X}}))) \ . \tag{6}$$

In the absence of noise in the data, we have $\mathbf{X}^{(1)} = \mathbf{X}^{(2)}$, and error-free communication works even for $\beta \to \infty$. The higher the noise level, the lower we have to choose β in order to obtain weight sums that are approximately invariant under the stochastic fluctuations in the measurements thus preventing decoding errors. The error analysis of this protocol investigates the probability of decoding error $\mathbb{P}(\hat{\tau} \neq \tau_s | \tau_s)$. As derived for an equivalent channel in Buhmann (2011), an asymptotically vanishing error rate is achievable for rates

$$\rho \ \leq \ \mathcal{I}_{\beta}(\tau_j, \hat{\tau}) \ = \ \frac{1}{n} \log \left(\frac{|\{\tau_j\}| \mathcal{Z}_{12}}{\mathcal{Z}_1 \cdot \mathcal{Z}_2} \right) \ = \ \frac{1}{n} \left(\log \frac{|\{\tau_j\}|}{\mathcal{Z}_1} + \log \frac{|\mathcal{C}^{(2)}|}{\mathcal{Z}_2} - \log \frac{|\mathcal{C}^{(2)}|}{\mathcal{Z}_{12}} \right) \tag{7}$$

The three logarithmic terms in eq.(7) denote the mutual information between the coarsening of the pattern space on the sender side and the coarsening of the pattern space on the receiver side.

The cardinality $|\{\tau_j\}|$ is determined by the number of realizations of the random transformation τ, i.e. by the entropy of the type (in an information theoretic sense) of the empirical minimizer $c^{\perp}(\mathbf{X})$. As the entropy increases for a large number of patterns, $|\{\tau_j\}|$ accounts for the model complexity or informativeness of the solutions. For noisy data, the communication rate is reduced as otherwise the solutions can not be resolved by the receiver. The relative weights are determined by the term $\mathcal{Z}_{12}/(\mathcal{Z}_1 \cdot \mathcal{Z}_2) \in [0, 1]$ which accounts for the stability of the model under noise fluctuations.

In analogy to information theory, we define the *approximation capacity* as

$$\mathcal{CAP}(\tau_s, \hat{\tau}) = \max_{\beta} \mathcal{I}_\beta(\tau_s, \hat{\tau})\,. \tag{8}$$

Using these entities, we can describe how to apply the wASC principle for model selection from a set of cost functions $\mathcal{R}$: Randomly split the given dataset $\mathbf{X}$ into two subsets $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$. For each candidate cost function $R(c, \mathbf{X}) \in \mathcal{R}$, compute the mutual information (eq. 7) and maximize it with respect to β. Then select the cost function that achieves highest capacity at the best resolution $\beta^\star$.

3. Approximation Capacity for Mixture of Gaussians

In this section, we demonstrate the principle of maximum approximation capacity on the well known Gaussian mixture model (GMM). We first derive how to calculate the approximation capacity for GMMs and then we experimentally compare it against other model selection principles.

3.1. Calculation of Approximation Capacity for Gaussian Mixture Models

A GMM with K components is defined as $p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma})$, with non-negative π_k and $\sum_k \pi_k = 1$. For didactical reasons, we do not optimize the covariance matrix $\boldsymbol{\Sigma}$ and simply fix it to $\boldsymbol{\Sigma} = 0.5 \cdot \mathbf{I}$. Then, maximizing the GMM likelihood essentially boils down to centroid-based clustering. As follows, we calculate the approximation capacity in two steps: i) calculate the weight sums $\mathcal{Z}_q$, $q = 1, 2$, and the joint weight sum $\mathcal{Z}_{12}$; ii) maximize $\mathcal{I}_\beta$ w.r.t. β (eq. 7).

For calculating the weight sums, we introduce two auxiliary variables. Let $\boldsymbol{M}$ be the cluster assignment matrix encoding the clustering solution c. It assigns objects to clusters with $M_{i,k} \in \{0, 1\}$ and $\sum_k M_{i,k} = 1, \forall i$. Furthermore, let $\epsilon_{i,k}^{(q)} := \|\mathbf{x}_i^{(q)} - \boldsymbol{\mu}_k^{(q)}\|^2$ be the short-hand notation for the negative log-likelihood of object i from dataset q given cluster k. As the full data likelihood factorizes, the cost function of GMM (with fixed covariance) is $R = \sum_{i=1}^{n} \sum_{k=1}^{K} M_{i,k} \epsilon_{i,k}^{(q)}$. The hypothesis space for this centroid-based clustering model is parameterized by the cluster assignments. The transformations used in the coding scenario modify these clusterings. Therefore, we obtain the individual weight sums and the joint weight sum by summing over all possible clustering solutions

$$\mathcal{Z}_q = \sum_{c \in C(\mathbf{X}^{(q)})} \exp\left(-\beta \sum_{i=1}^{n} \sum_{k=1}^{K} M_{i,k} \epsilon_{i,k}^{(q)}\right) = \prod_{i=1}^{n} \sum_{k=1}^{K} \exp\left(-\beta \epsilon_{i,k}^{(q)}\right)\,, \tag{9}$$

$$\mathcal{Z}_{12} = \sum_{c \in C(\mathbf{X}^{(2)})} \exp\left(-\beta \sum_{i=1}^{n} \sum_{k=1}^{K} M_{i,k} (\epsilon_{i,k}^{(1)} + \epsilon_{i,k}^{(2)})\right) = \prod_{i=1}^{n} \sum_{k=1}^{K} \exp\left(-\beta \left(\epsilon_{i,k}^{(1)} + \epsilon_{i,k}^{(2)}\right)\right)\,. \tag{10}$$

By substituting these weight sums to eq. 7, the mutual information amounts to

$$\mathcal{I}_\beta = \log K + \frac{1}{n} \sum_{i=1}^{n} \left(\log \sum_{k=1}^{K} e^{-\beta\left(\epsilon_{i,k}^{(1)} + \epsilon_{i,k}^{(2)}\right)} - \log \sum_{k=1}^{K} e^{-\beta \epsilon_{i,k}^{(1)}} \sum_{k'=1}^{K} e^{-\beta \epsilon_{i,k'}^{(2)}} \right). \tag{11}$$

The approximation capacity is numerically determined as the maximum of $\mathcal{I}_\beta$ over β.

3.2. Experimental Evaluation

We define $K = 5$ Gaussians with parameters $\pi_k = 1/K$, $\boldsymbol{\mu} = [(1,0);(0,1.5);(-2,0);(0,-3);(4.25,-4)]$, and with covariance $\boldsymbol{\Sigma} = 0.5 \cdot \mathbf{I}$. Let $\mathbf{X}^{(q)}, q \in \{1,2\}$ be two datasets of identical size $n = 10,000$ drawn from these Gaussians. We optimize the assignment variables and the centroid parameters of our GMM model via annealed Gibbs sampling. Thereby, we provide double as many clusters to the model in order to enable overfitting. Starting from a high temperature, we successively cool down while optimizing the model parameters. In Figure 2(a), we illustrate the positions of the centroids with respect to the center of mass. At high temperature, all centroids coincide, indicating that the optimizer favors one cluster. As we cool down, the centroids separate into increasingly many clusters until, finally, all 10 clusters are used to fit the data.

Figure 2(b) shows the numerical analysis of the mutual information (eq. 11). When the stopping temperature of the Gibbs sampler coincides with the temperature β^{-1} that maximizes mutual information, we expect the best tradeoff between robustness and informativeness. And indeed, as illustrated in Figure 2(a), the correct model-order $\hat{K} = 5$ is found at this temperature. At lower stopping temperatures, the clusters split into many instable clusters which increases the decoding error, while at higher temperatures informativeness of the clustering solutions decreases.

Relation to generalization ability: A properly regularized clustering model explains not only the dataset at hand, but also new datasets from the same source. The inferred model parameters and assignment probabilities from the first dataset $\mathbf{X}^{(1)}$ can be used used to compute the costs for the second dataset $\mathbf{X}^{(2)}$. The appropriate clustering model yields low costs on $\mathbf{X}^{(2)}$, while very informative but unstable structures and also very stable but little informative structures have high costs due to overfitting and underfitting, respectively.

We measure this generalization ability by computing the "transfer costs" $R(c^{(1)}, \mathbf{X}^{(2)})$: At each stopping temperature of the Gibbs sampler, we take the current parameters $\boldsymbol{\mu}^{(1)}$ and assignment probabilities $\mathbf{P}^{(1)}$ inferred from $\mathbf{X}^{(1)}$ and transfer them to $\mathbf{X}^{(2)}$. The assignment probabilities $\mathbf{P}^{(1)}$ follow a Gibbs distribution

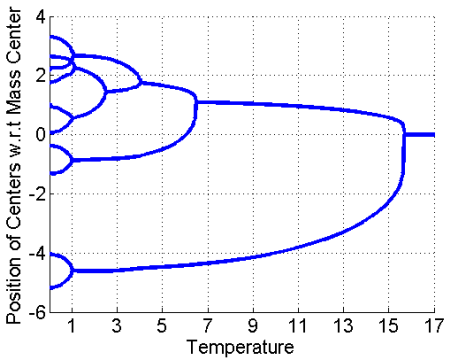
$$p(\boldsymbol{\mu}_k^{(1)}|\mathbf{x}_i^{(1)}) = Z_x^{-1} \exp\left(-\beta \|\mathbf{x}_i^{(1)} - \boldsymbol{\mu}_k^{(1)}\|^2\right), \tag{12}$$

with Z_x as the normalization constant. The expected transfer costs with respect to these probabilities are then

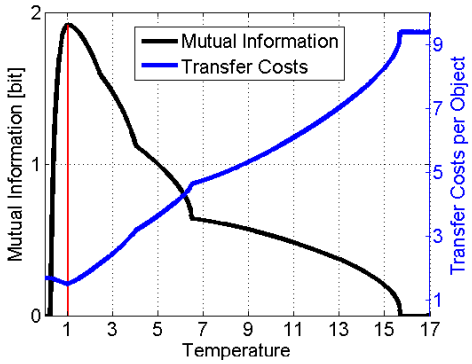
$$\left\langle R(c^{(1)}, \mathbf{X}^{(2)}) \right\rangle = \sum_{i=1}^{n} \sum_{k=1}^{K} p(\mathbf{x}_i^{(1)}, \boldsymbol{\mu}_k^{(1)}) \|\mathbf{x}_i^{(2)} - \boldsymbol{\mu}_k^{(1)}\|^2 \approx \frac{1}{n} \sum_{n=1}^{n} \sum_{k=1}^{K} p(\boldsymbol{\mu}_k^{(1)}|\mathbf{x}_i^{(1)}) \|\mathbf{x}_i^{(2)} - \boldsymbol{\mu}_k^{(1)}\|^2 , \tag{13}$$

Figure 2(b) illustrates the transfer costs as a function of β and compares it with the approximation capacity. The optimal transfer costs are obtained at the stopping temperature that corresponds to the approximation capacity.

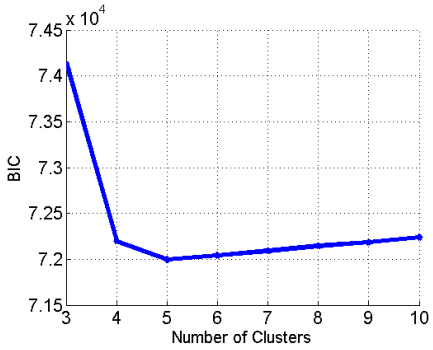
Relation to BIC Arguably the most popular criterion for model-order selection is BIC as proposed by Schwarz (1978). It is, like wASC, an asymptotic principle, i.e. for sufficiently



(*a*) Clustering hierarchy



(*b*) Mutual info. and transfer costs



(*c*) BIC measure

Figure 2: Annealed Gibbs sampling for GMM: Influence of the stopping temperature for annealed optimization on the mutual information, on the transfer costs and on the positions of the cluster centroids. The lowest transfer cost is achieved at the temperature with highest mutual information. This is the lowest temperature at which the correct number of clusters $\hat{K} = 5$ is found. The hierarchy in Fig. 2(a) is obtained by projecting the two-dimensional centroids at each stopping temperature to the optimal one-dimensional subspace using multidimensional scaling. BIC verifies correctness of $\hat{K} = 5$.

many observations, the fitted model preferred by BIC ideally corresponds to the candidate which is a posteriori most probable. However, the application of BIC is limited to models where one can determine the number of free parameters as here with GMM. Figure 2(c) confirms the consistency of wASC with BIC in finding the correct model order.

4. Conclusion

Model selection and model order selection pose critical design issues in all unsupervised learning tasks. The principle of maximum approximation capacity (wASC) offers a theoretically well-founded approach to answer these questions. We have motivated this principle and derived the general form of the capacity. As an example, we have studied the ap-

proximation capacity of Gaussian mixture models (GMM). Thereby, we have demonstrated that the choice of the optimal number of Gaussians based on the approximation capacity coincides with the configurations yielding optimal generalization ability. Weighted approximation set coding finds the true number of Gaussians used to generate the data.

Weighted approximation set coding is a very general model selection principle which is applicable to a broad class of pattern recognition problems (for SVD see Frank and Buhmann (2011)). We have shown how to use wASC for model selection and model order selection in clustering. Future work will address the generalization of wASC to discrete continuous optimization problems, such as sparse regression, and to algorithms without cost functions.

References

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.

Joachim M. Buhmann. Information theoretic model validation for clustering. In *International Symposium on Information Theory, Austin Texas*, pages 1398 – 1402. IEEE, 2010. doi: 10.1109/ISIT.2010.5513616.

Joachim M. Buhmann. Context sensitive information: Model validation by information theory. In J.-F. Martnez-Trinidad et al., editor, *MCPR 2011*, volume 6718 of *LNCS*, pages 21–21. Springer, 2011.

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

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

Mario Frank and Joachim M. Buhmann. Selecting the rank of SVD by maximum approximation capacity. In *International Symposium on Information Theory (ISIT 2011), St. Petersburg*. IEEE, 2011.

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

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

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.

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