Multivariate Information Bottleneck

Nir Friedman **Ori Mosenzon** **Noam Slonim** **Naftali Tishby**
School of Computer Science & Engineering, Hebrew University, Jerusalem 91904, Israel
{nir, mosenzon, noamm, tishby }@cs.huji.ac.il

Abstract

The *Information bottleneck method* [11] is an unsupervised non-parametric data organization technique. Given a joint distribution $P(A, B)$, this method constructs a new variable T that extracts partitions, or clusters, over the values of A that are informative about B. The information bottleneck has already been applied to document classification, gene expression , neural code, and spectral analysis. In this paper, we introduce a general principled framework for multivariate extensions of the information bottleneck method. This allows us to consider multiple systems of data partitions that are inter-related. Our approach utilizes Bayesian networks for specifying the systems of clusters and what information each captures. We show that this construction provides insight about bottleneck variations and enables to characterize solutions of these variations. We also present a general framework for iterative algorithms for constructing solutions, and apply it to several examples.

1 Introduction

Clustering, or data partitioning, is a common data analysis paradigm. A central question is understanding general underlying principles for clustering. One information theoretic approach to clustering is to require that clusters should capture only the "relevant" information in the data, where the relevance is explicitly determined by various components of the data itself. A common data type which calls for such a principle is co-occurrence data, such as verbs and direct objects in sentences [6], words and documents [1, 3, 9], tissues and gene expression patterns [12], galaxies and spectral components [8], etc. In most such cases the objects are discrete or categoric and no obvious "correct" measure of similarity exists between them. Thus, we would like to rely purely on the joint statistics of the co-occurrences and organize the data such that the "relevant information" among the variables is captured in the best possible way.

Formally, we can quantify the relevance of one variable, A, with respect to another one, B, in term of the *mutual information*, $I(A; B)$. This well known quantity, defined as,

$$I(A; B) = \sum_{a,b} P(a, b) \log \frac{P(a, b)}{P(a)P(b)}$$

is symmetric, non-negative, and equals to zero if and only if the variables are independent. It measures how many bits are needed to convey the information A has about B (or vice versa). The aim of information theoretic clustering is to find (soft) partitions of A's values that are informative about B. This requires balancing two goals: we want to loose irrelevant distinctions made by A, and at the same time to maintain relevant ones. A possible principle for extracting such partitions is the information bottleneck (IB) method [11]. Clustering is posed as a construction of a new variable T that represents partitions of A. The principle is described by a variational tradeoff between the information we try to minimize, $I(A; T)$, and the one we try to maximize $I(T; B)$. We briefly review this principle and its consequences in the next section.

The main contribution of this paper is a general formulation of a *multivariate* extension of the information bottleneck principle. This extension allows us to consider cases where the clustering is relevant with respect to several variables, or where we construct two or more systems of clusters at the same time.

To give concrete motivation, we briefly mention two examples that we treat in detail in later sections. In *symmetric clustering* (also called *two-sided or double clustering*) we want to find two systems of clusters: one of A and one of B that are informative about each other. A possible application is relating documents to words, where we seek clustering of documents according to word usage, and a corresponding clustering of words. This procedure aims to find document clusters that correspond to different topics and at the same time identify cluster of words that charcterize these topics [9]. Clearly, the two systems of clusters are in interaction, and we want a unifying principle that shows how to construct them simultaneously. In *parallel clustering* we attempt to build two systems of clusters of the values of A. Our aim here is to capture independent aspects of the information A conveys about B. A biological example appear in analysis of gene expression data, where multiple independent distinctions about tissues (healthy vs. tumor, epithelial vs. muscle, etc.) are relevant for the expression of genes.

We present such tasks, and others, in our framework by specifying a pair of Bayesian networks. One network, G_{in}, represents which variables are compressed versions of the observed variables (each new variable compresses it parents in the network). The second network, G_{out}, represents which relations should be maintained or predicted (each variable is predicted by its parents in the network). We formulate the general principle as a tradeoff between the information each network carries. We want to minimize the information maintained by G_{in} and to maximize the information maintained by G_{out}.

We further give another interpretation to this principle, as a tradeoff between compression of the source (given by G_{in}) and fitness to a *target model*, where the model is described by G_{out}. Using this interpretation we can think of our new principle as a generalized compression distortion tradeoff (as in rate-distortion theory). This interpretation may allow us to investigate the principle in a general parametric setup. In addition, we show that, as with the original IB, the new principle provides us with self-consistent equations in the unknown probabilistic partition(s) which can be iteratively solved and shown to converge. We show how to combine this in a deterministic annealing procedure which enables us to explore the information tradeoff in an hierarchical manner. There are many possible applications for our new principle and algorithm. To mention just few, we consider semantic clustering of words based on multiple parts of speech, complex gene-expression data analysis, and neural code analysis.

2 The Information Bottleneck

Tishby et al [11] considered two variables, A and B, with their (assumed given) joint distribution $P(A, B)$. Here A is the variable we try to compress, with respect to the "relevant" variable B. Namely, we seek a (soft) partition of A through an auxiliary variable T and the probabilistic mapping $P(T \mid A)$, such that the the mutual information $I(A; T)$ is minimized (maximum compression) while the relevant information $I(T; B)$ is maximized. The dependency relations between the 3 variables can be described by the relations: T independent of B given A; and on the other hand we want to predict B from T.

By introducing a positive Lagrange multiplier β, Tishby *et al* formulate this tradeoff by minimizing the following Lagrangian,

$$\mathcal{L}[P(T \mid A)] = I(A; T) - \beta I(T; B)$$

where we take $P(A, B, T) = P(A, B)P(T \mid A)$.

By taking the variation (i.e derivative in the finite case) of $\mathcal{L}$ w.r.t. $P(T \mid A)$, under the proper normalization constraints, it was shown in [11] that the optimal partition satisfy,

$$P(T \mid A) = \frac{P(T)}{Z(A, \beta)} \exp\left(-\beta D(P(B \mid A) \| P(B \mid T))\right)$$

where $D(P\|Q) = E_P[\log \frac{P}{Q}]$ is the familiar KL divergence [2]. This equation must be satisfied self consistently. The practical solution of these equations can be done by repeated iterations of the self-consistent equations, for every given value of β, similar to clustering by deterministic annealing [7]. The convergence of these iterations to a (generally local) optimum was proven in [11] as well.

3 Bayesian Networks and Multi-Information

A Bayesian network *structure* over a set of random variables $X_1, \ldots, X_n$ is a DAG G in which vertices are annotated by names of random variables. For each variable X_i, we denote by $\mathbf{Pa}^G_{X_i}$ the (potentially empty) set of parents of X_i in G. We say that a distribution P is *consistent* with G, if P can be factored in the form:

$$P(X_1, \ldots, X_n) = \prod_i P(X_i \mid \mathbf{Pa}^G_{X_i})$$

and use the notation $P \models G$ to denote that.

One of the main issues that we will deal with is the amount of information that variables $X_1, \ldots, X_n$ contain about each other. A quantity that captures this is the *multi-information* given by

$$\begin{aligned} \mathcal{I}(X_1, \ldots, X_n) &= D(P(X_1, \ldots, X_n) \| P(X_1) \cdots P(X_n)) \\ &= E\left[\log \frac{P(X_1, \ldots, X_n)}{P(X_1) \cdots P(X_n)}\right] . \end{aligned}$$

The multi-information captures how close is the distribution $P(X_1, \ldots, X_n)$ to the factored distribution of the marginals. This is the natural generalization of the pairwise concept of mutual information. If this quantity is small, we do not "lose" much by approximating P by the product distribution. Like mutual information, it measures the number of bits that can be gained by a joint compression of the variables vs. independent compression.

When P has additional known independence relations, we can rewrite the multi-information in terms of the dependencies among the variables:

Proposition 3.1: *Let G be a Bayesian network structure over $\mathbf{X} = \{X_1, \ldots, X_n\}$, and let P be a distribution over $\mathbf{X}$ such that $P \models G$. Then,*

$$\mathcal{I}(X_1, \ldots, X_n) = \sum_i I(X_i; \mathbf{Pa}^G_{X_i}) .$$

That is, the multi-information is the sum of *local* mutual information terms between each variable and its parents. We denote the sum of these informations with respect to a network structure as:

$$\mathcal{I}^G = \sum_i I(X_i; \mathbf{Pa}^G_{X_i}) .$$

When P is not consistent with the DAG G, we often want to know how close is P to a distribution that is consistent with G. That is, what is the "distance" (or *distortion*) of P from its *projection* onto the sub-space of distributions consistent with G. We naturally define this distortion as

$$D(P\|G) = \min_{Q \models G} D(P\|Q) \, .$$

This measure has two natural interpretation in terms of the graph G.

Proposition 3.2: *Let G be a Bayesian network structure over $\mathbf{X} = \{X_1, \ldots, X_n\}$, and let P be a distribution over $\mathbf{X}$. Assume that the order $X_1, \ldots, X_n$ is consistent with the DAG, then*

$$\begin{aligned} D(P\|G) &= \sum_i I(X_i; \{X_1, \ldots, X_{i-1}\} - \mathbf{Pa}^G_{X_i} \mid \mathbf{Pa}^G_{X_i}) \\ &= \mathcal{I}(X_1, \ldots, X_n) - \mathcal{I}^G \end{aligned}$$

Thus we see that $D(P\|G)$ can be expressed as a sum of conditional information terms, where each term corresponds to a *Markov independence assumption* with respect to the order $X_1, \ldots, X_n$. Recall that the Markov independence assumptions (with respect to a given order) are necessary and sufficient to require the factored form of distributions consistent with G [5]. We see that $D(P\|G)$ is measured in terms of the extent these independencies are violated, since $I(X_i; \{X_1, \ldots, X_{i-1}\} - \mathbf{Pa}^G_{X_i} \mid \mathbf{Pa}^G_{X_i}) = 0$ if and only if X_i is independent of $\{X_1, \ldots, X_{i-1}\} - \mathbf{Pa}^G_{X_i}$ given its parents. Thus, $D(P\|G) = 0$ if and only if P is consistent with G.

An alternative representation of this distance measure is given in terms of multi-informations, since we can think of $D(P\|G)$ as the amount of information between the variables that cannot be captured by the dependencies of the structure G.

4 Multi-Information bottleneck principle

The multi-information allows us to introduce a simple "lift-up" of the original IB variational principle to the multivariate case, using the semantics of Bayesian networks of the previous section. Given a set of observed variables, $\mathbf{X} = \{X_1, \ldots, X_n\}$, instead of one partition variable T, we now consider a set $\mathbf{T} = \{T_1, \ldots, T_k\}$, which correspond to different partitions of various subsets of the observed variables. More specifically, we want to "construct" new variables. The relations between the observed variables and these new compression variables is specified using a DAG G_{in} over $\mathbf{X} \cup \mathbf{T}$. Since we assume that the new variables in $\mathbf{T}$ are functions of the original variables, we restrict attentions to DAGs where the variables in $\mathbf{T}$ are leafs. [1] Thus, each T_i is a stochastic function of a set of variables $\mathbf{Pa}^{G_{in}}_{T_i} \subseteq \mathbf{X}$. Once these are set, we have a joint distribution over the combined set of variables:

$$P(\mathbf{X}, \mathbf{T}) = P(\mathbf{X}) \prod_i P(T_i \mid \mathbf{Pa}^{G_{in}}_i) \qquad (1)$$

Similarly to the original IB formulation, the information that we would like to minimize is now given by $\mathcal{I}^{G_{in}}$. Minimizing this quantity attempts to make variables as independent of each other as possible. (Note that since we only modify conditional distributions of variables in $\mathbf{T}$, we cannot modify the dependencies among the original variables.)

The "relevant" information that we want to preserve is specified by another DAG, G_{out}. This graph specifies, for each T_i which variables it predicts. These are simply its children in G_{out}. Conversely, we want to predict each X_i (or T_i) by its parents in G_{out}. Thus, we think of $\mathcal{I}^{G_{out}}$ as a measure of how much information the variables in $\mathbf{T}$ maintain about their target variables. Thus, the term we wish to maximize is $\mathcal{I}^{G_{out}}$.

The generalized Lagrangian can be written as

$$\mathcal{L}^{(1)}[p(T_1 \mid \mathbf{Pa}^{G_{in}}_{T_1}), ..., p(T_k \mid \mathbf{Pa}^{G_{in}}_{T_k})] = \mathcal{I}^{G_{in}} - \beta \mathcal{I}^{G_{out}} \, , \qquad (2)$$

and the variation is done subject to the normalization constraints on the partition distributions. It leads to tractable self-consistent equations, as we henceforth show.

It is easy to see that the form of this Lagrangian is a direct generalization of the original IB principle. Again, we try to balance between the information $\mathbf{T}$ loses about $\mathbf{X}$ in G_{in} and the information it preserves with respect to G_{out}.

5 Analogous Variational Principle

We now describe another extension and interpretation of the new variational principle. This one is based on approximating distributions by a class defined by the Bayesian network G_{out}, rather than on preservation of multi-information.

We face the problem of choosing the conditional distributions $P(T_i \mid \mathbf{Pa}^{G_{in}}_i)$. Thus, we also need to specify what is exactly our target in constructing these variables. As with the original IB method, we are going to assume that there are two goals.

On the one hand, we want to compress, or partition, the original variables. As before the natural multivariate form of this, is to minimize the multi-information of P. Since P is consistent with G_{in}, we can denote this multi-information as $\mathcal{I}^{G_{in}}$.

While in the multi-information bottleneck principle of the previous section the second goal was to preserve the multi-information to other (target, relevant) variables, here we think of a target class of model distributions, specified by a target Bayesian network. In that interpretation the compressed variables should help us in describing the joint distribution in a different desired structured. We specify this structure by another DAG G_{out} that represents which dependencies and independencies we would like to impose.

[1] It will be convenient to think of cases where G_{in} restricted to $\mathbf{X}$ forms a complete graph. However, this is not crucial in the following development. To simplify the technical details, we do assume that $P(\mathbf{X})$ is consistent with G_{in}.

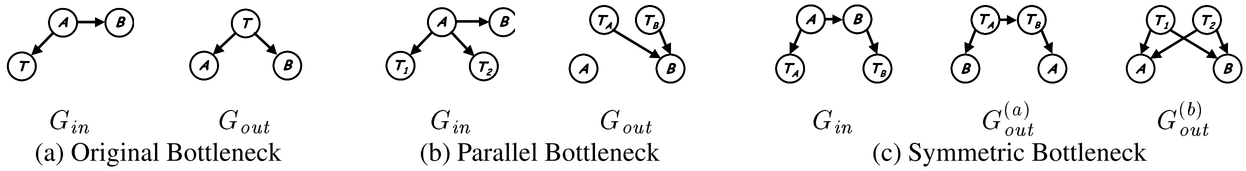


Figure 1: The source and target networks for the three examples we consider.

To make this more concrete consider the simple two-variable case shown in Figure 1(a). In this example, we are given the distribution of two variables A and B. The graph G_{in} specifies that T is a compressed version of A. On the other hand, we would like this T to make A and B as independent as possible. A way of formally specifying this desire, is to specify the DAG G_{out} of Figure 1(a) such that here T separates between A and B.

The question now is how to force the construction of $P(T \mid A)$ such that it will lead to the independencies that are specified in the target graph. Notice that these two graphs are incompatible. We cannot achieve both sets of independencies simultaneously. Instead, we aim to come close to achieving this by a tradeoff between the two. We formalize this, by requiring that P is closely *approximated* by a distribution consistent with G_{out}. As previously discussed, the natural information theoretic measure for this approximation is provided by Sanov theorem to be the $D(P\|G)$, or the minimal KL divergence from P to distributions consistent with G_{out}.

As before, we introduce a Lagrange multiplier, β, that controls the tradeoff between these objectives. The Lagrangian we want to minimize in this formulation is thus:

$$\mathcal{L}^{(2)} = \mathcal{I}^{G_{in}} + \beta D(P\|G_{out}) \tag{3}$$

where the parameters that we can change during the minimization is over the conditional distributions $P(T_i \mid \mathbf{Pa}_{T_i}^{G_{in}})$. The range of β is between 0, in which case we have a trivial solution in which the T_i's are independent of their parents, and ∞, in which we strive to make P as close as possible to G_{out}.

Example 5.1: Consider again the example of Figure 1(a). In this case, we have that $\mathcal{I}^{G_{in}} = I(A;B)+I(T;A)$. Using Proposition 3.2, we have that $D(P\|G_{out}) = \mathcal{I}^{G_{in}} - \mathcal{I}^{G_{out}}$. Putting, these together, we get the Lagrangian

$$\mathcal{L} = I(A;T) - \beta I(B;T) + (1+\beta)I(A;B)$$

Since, $I(A;B)$ is constant, we can ignore it, and we end up with the Lagrangian of the original IB method. Thus, we can think of the bottleneck as finding a compression T of A that results in a joint distribution that is as close as possible to the graph where A and B are independent given T.

Going back to the general case, we can apply Proposition 3.2 to rewrite the Lagrangian in terms of multi-informations:

$$\mathcal{L}^{(2)} = \mathcal{I}^{G_{in}} + \beta(\mathcal{I}^{G_{in}} - \mathcal{I}^{G_{out}})$$

which is equivalent to the Lagrangian $\mathcal{L}^{(1)}$ presented in the previous section, under the transformation $\beta' = \frac{\beta}{1+\beta}$.

We conclude with two examples of the principle applied to different situations.

Example 5.2: We first consider a simple extension of the original IB. Suppose, we introduce two variables T_1 and T_2. As specified in G_{in} of Figure 1(b), both of these variables are stochastic functions of A. In addition, similarly to the original IB, G_{out} specifies that T_1 and T_2 should predict B. We can think of this requirement as an attempt to decompose the information A contains about B into two "orthogonal" components. We call this example the *parallel* bottleneck.

We can make this intuition more precise by examining the Lagrangian defined by the principle. In this case, $\mathcal{I}^{G_{in}} = I(A;B) + I(A;T_2) + I(A;T_1)$ and $\mathcal{I}^{G_{out}} = I(B;T_1,T_2)$. After dropping the constant term $I(A;B)$, the Lagrangian $\mathcal{L}$ can be written as

$$\mathcal{L} = I(A;T_1) + I(A;T_2) - \beta(I(B;T_1,T_2))$$

Thus, we attempt to minimize the information between T_1 and T_2 while maximizing the information they *together* preserve about B.

Example 5.3: We now consider the *symmetric* bottleneck. In this case, we want to compress A into T_A and B into T_B so that T_A extracts the information A contains about B, and at the same time T_B extracts the information B contains about A. The DAG G_{in} of Figure 1(c) captures the form of the compression. The choice of G_{out} is less obvious.

One alternative, shown as $G_{out}^{(a)}$ in Figure 1(c), attempts to make each of T_A and T_B sufficient to separate A from B. As we can see, in this network A is independent of B and T_B given T_A. Similarly, T_B separates B from the other variables. The structure of the network states that T_A and T_B are dependent of each other. Developing the Lagrangian defined by this network, we get:

$$\mathcal{L}^{(a)} = I(A;T_A) + I(B;T_B) - \beta I(T_A;T_B)$$

Thus, on one hand we attempt to compress, and on the other hand we attempt to make T_A and T_B as informative about each other as possible. (Note that if T_A is informative about T_B, then it is also informative about B.)

Alternatively, we might argue that T_A and T_B should each compress different "aspects" of the connection between A and B. This intuition is specified by the target network $G_{out}^{(b)}$ of Figure 1(c). In this network T_A and T_B are independent of each other, and both are needed to make A and B conditionally independent. In this sense, our aim is to find T_A and T_B that capture independent attributes of the connection between A and B. Indeed, we can write the Lagrangian as:

$$\begin{aligned}\mathcal{L}^{(b)} &= I(A;T_A) + I(B;T_B) + \\ &\quad \beta(2I(T_A;T_B) - I(A;T_B) - I(B;T_A))\end{aligned}$$

That is, we attempt to maximize the information T_A maintains about B and T_B about A, and at the same time try to minimize the information between T_A and T_B.

6 Characterization of the Solution

In the previous section we stated a variational principle. In this section we consider the form of the solutions of the principle. More precisely, we assume that G_{in}, G_{out}, and β are given. We now want to describe the properties of the distributions $P(T_i \mid \mathbf{Pa}_{T_i}^{G_{in}})$.

In the presentation of this characterization, we need some additional notational shorthands. We denote by $\mathbf{U}_i = \mathbf{Pa}_{T_i}^{G_{in}}$, $\mathbf{V}_{T_i} = \mathbf{Pa}_{T_i}^{G_{out}}$, and $\mathbf{V}_{X_i} = \mathbf{Pa}_{X_i}^{G_{out}}$. We also denote $\mathbf{V}_{T_j}^{-T_i} = \mathbf{V}_{T_j} - \{T_i\}$ and similarly for $\mathbf{V}_{X_j}^{-T_i}$. Finally, we denote by $\mathbf{W}_i = \mathbf{U}_i \cap \mathbf{V}_{T_i}$ the common parents of T_i in G_{in} and G_{out}.

The main result of this section is as follows.

Theorem 6.1: *Assume that $P(\mathbf{X})$, G_{in}, G_{out}, and β are given. The conditional distributions $\{P(T_i \mid \mathbf{U}_i)\}$ are a stationary point of $\mathcal{L} = \mathcal{I}^{G_{in}} - \beta \mathcal{I}^{G_{out}}$ if an only if*

$$P(t_i \mid \mathbf{u}_i) = \frac{P(t_i)}{Z_{T_i}(\mathbf{u}_i, \beta)} e^{-\beta d(t_i, \mathbf{u}_i)} \qquad (4)$$

where $Z_{T_i}(\mathbf{u}_i, \beta)$ is a normalization term, and $d(t_i, \mathbf{u}_i)$ is given by

$$\begin{aligned}&\sum_{j:T_i \in \mathbf{V}_{X_j}} E_{P(\cdot|\mathbf{u}_i)}[D(P(X_j \mid \mathbf{V}_{X_j}^{-T_i}, \mathbf{u}_i) \| P(X_j \mid \mathbf{V}_{X_j}^{-T_i}, t_i))] \\ &+ \sum_{j:T_i \in \mathbf{V}_{T_j}} E_{P(\cdot|\mathbf{u}_i)}[D(P(T_j \mid \mathbf{V}_{T_j}^{-T_i}, \mathbf{u}_i) \| P(T_j \mid \mathbf{V}_{T_j}^{-T_i}, t_i))] \\ &\qquad + D(P(\mathbf{V}_{T_i}^{-\mathbf{U}_i} \mid \mathbf{u}_i) \| P(\mathbf{V}_{T_i}^{-\mathbf{U}_i} \mid t_i, \mathbf{w}_i)) \,.\end{aligned}$$

All marginal and conditional probabilities in this term are derived from the definition of the model in Eq. (1). [2]

The essence of this theorem is that it defines $P(t_i \mid \mathbf{u}_i)$ in terms of the *distortion* $d(t_i, \mathbf{u}_i)$. This distortion measures how close are the conditional distribution in which t_i is involved into these where we replace t_i with $\mathbf{u}_i$. In other words, we can understand this as measuring how well t_i performs as a "representative" of the particular assignment $\mathbf{u}_i$. The conditional distribution $P(T_i \mid \mathbf{u}_i)$ depends on the differences in the distortion for different values of T_i.

The theorem also allows us to understand the role of β. When β is small, the conditional distribution is diffused, since β reduces the differences between the distortions for different values or T_i. On the other hand, when β is large, the exponential term acts as "softmax" gate, and most of the conditional probability mass will be assigned to the value t_i with the smallest distortion. This behavior matches the intuition that when β is small, most of the emphasis is on compressing the input variables $\mathbf{U}_i$ into T_i and when β is large, most of the emphasis is on predicting the "outputs" variables of T_i, as specified by G_{out}.

Example 6.2: To see a concrete example, we reconsider the parallel bottleneck of Example 5.2. Applying the theorem, we get that the distortion term for T_1 is

$$d(t_1, a) = E_{P(\cdot|a)}[D(P(B \mid a, T_2) \| P(B \mid t_1, T_2))]$$

This term corresponds to the information of B and T_1, T_2. We see that $P(t_1 \mid a)$ increases when the predictions of B given t_1 are similar to those given a (when averaging over T_2). The distortion for T_2 is defined analogously.

Example 6.3: Consider now the symmetric bottleneck case of $G_{out}^{(a)}$ in Example 5.3. Applying the theorem, we get two self-consistent equations:

$$\begin{aligned}P(t_A \mid a) &= \frac{P(t_A)}{Z_{T_A}(a, \beta)} e^{-\beta D(P(T_B|a) \| P(T_B|t_A))} \\ P(t_B \mid b) &= \frac{P(t_B)}{Z_{T_B}(b, \beta)} e^{-\beta D(P(T_A|b) \| P(T_A|t_B))}\end{aligned}$$

Thus, T_A attempts to make predictions as similar to these of A about T_B, and similarly T_B attempts to make predictions as similar to these of B about T_A.

7 Iterative Optimization Algorithm

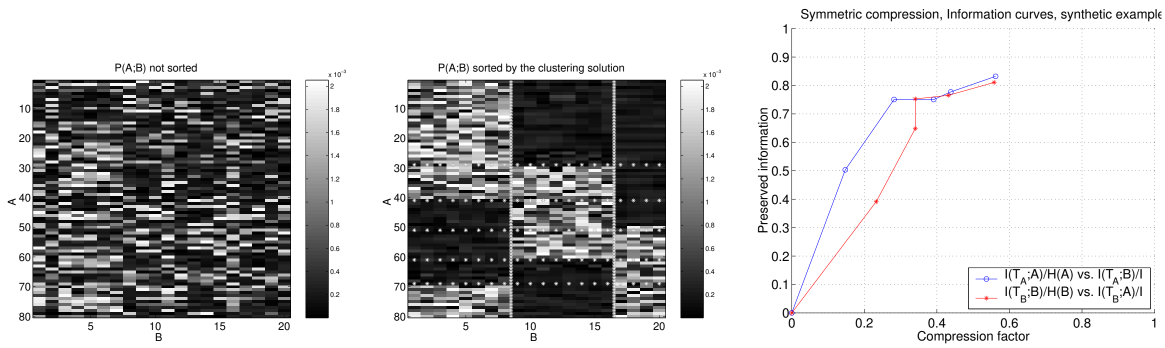
We now consider algorithms for constructing solutions of the variational principle.

We start with the case where β is fixed. In this case, following standard strategy in variational methods, we simply apply the self-consistent equations. More precisely, we use an iterative algorithm, that at the m'th iteration maintains the conditional distributions $\{P(T_i \mid \mathbf{U}_i)^{(m)} : i = 1, \ldots, k\}$. At the $m+1$'th iteration, the algorithm applies an *update step*

$$P(t_i \mid \mathbf{u}_i)^{(m+1)} \leftarrow \frac{P(t_i)^{(m)}}{Z_{T_i}(\mathbf{u}_i, \beta)^{(m+1)}} e^{-\beta d(t_i, \mathbf{u}_i)^{(m)}} \qquad (5)$$

where $P(t_i)^{(m)}$ and $d(t_i, \mathbf{u}_i)^{(m)}$ are computed with respect to the conditional probabilities $\{P(T_i \mid \mathbf{U}_i)^{(m)} : i = 1, \ldots k\}$.

[2] This can be done by standard variable elimination procedures.



(a) (b) (c)

Figure 2: Application of the symmetric bottleneck on a simple synthetic example. (a) input joint distribution, (b) the same joint distribution where rows and columns were permuted to match the clustering found; dotted lines show cluster boundaries. (c) information curves showing the progression along the information tradeoff graph. The x-axis is the fraction of the information about the original variable that is maintained by the compressed variable, and the y-axis is the fraction of the information between A and B that is captured by $I(T_A; B)$ or $I(T_B; A)$. Circles denote bifurcation events.

There are two main variants of this algorithm. In the *synchronous* variant, we apply the update step for all the conditional distributions in each iteration. That is, each conditional probability $P(T_i \mid \mathbf{U}_i)$ is updated by computing the distortion based on the conditional probabilities of the previous iterations. In the *asynchronous* variant, we choose one variable T_i, and perform the update only for this variable. For all $j \neq i$, we set $P(T_j \mid \mathbf{U}_j)^{(m+1)} = P(T_j \mid \mathbf{U}_j)^{(m)}$. The main difference between the two variants is that the update of T_i in the asynchronous update incorporates the implications of the updates of the previous T_j's.

The main question is whether these iterations converge.

Theorem 7.1 : *Asynchronous iterations of the self-consistent equations converge to a stationary point of the optimization problem.*

At the current stage we do not have a proof of convergence for the synchronous case, although in our experiments, synchronous update converges as well.

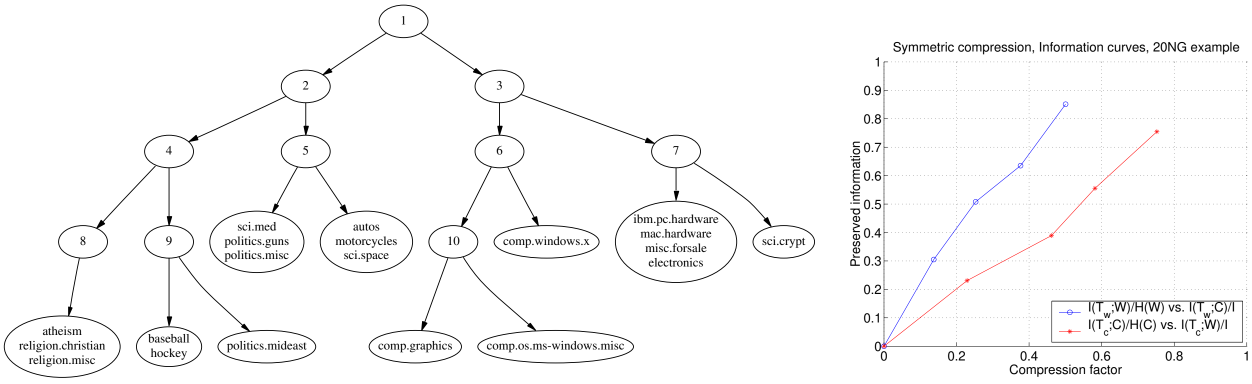
A key question is how to initialize this procedure, as different initialization points can lead to different solutions. We now describe a *deterministic annealing* like procedure [7, 11]. This procedure works by iteratively increasing the parameter β and then adapting the solution for the previous value of β to the new one. This allows the algorithm to "track" the changes in the solution as the system shifts its preferences from compression to prediction.[3]

Recall that when $\beta \to 0$, the optimization problem tends to make T_i independent of its parents. At this point the solution consists of essentially only one cluster for each T_i which is not predictive about any other variable. As we increase β, we suddenly reach a point where the values of T_i diverge and show two different behaviors. This phenomena is a phase-transition of the system. Successive increases of β will reach additional phase transitions in which additional splits of values of T_i emerge. The general idea of this annealing procedure is to identify these *bifurcations* of clusters. At each step of the procedure, we maintain a set of values for each T_i. Initially, when $\beta = 0$, each T_i has a single value. We then progressively increase β and try to detect bifurcations. At the end of the procedure we record for each T_i a bifurcating tree that traces the sequence of solutions at different values of β (see for example Figure 3(a)).

The main technical problem is how to detect such bifurcations. We adopt the methods of Tishby et al [11] to multiple variables. At each step, we take the solution from the previous step (i.e., for the previous value of β we considered) and construct an initial problem in which we duplicate each value of each T_i. To define such an initial solution we need to specify the conditional probabilities of these "doubled" values given each value $\mathbf{U}_i$. Suppose that $t_{i,a}$ and $t_{i,b}$ are the two copies of the value t_i. Then we set $P^*(t_{i,a} \mid \mathbf{u}_i) = P(t_i \mid \mathbf{u}_i)\left(\frac{1}{2} + \gamma\epsilon(\mathbf{u}_i)\right)$ and $P^*(t_{i,b} \mid \mathbf{u}_i) = P(t_i \mid \mathbf{u}_i)\left(\frac{1}{2} - \gamma\epsilon(\mathbf{u}_i)\right)$ where $\epsilon(\mathbf{u}_i) \sim U[-\frac{1}{2}, \frac{1}{2}]$ is a noise term and $\gamma \leq 1$ is a scale parameter. Thus, each copy $t_{i,a}$ and $t_{i,b}$ is a *perturbed* version of t_i. If β is high enough, this random perturbation suffices to allow the two copies of t_i to diverge. If β is too small to support such bifurcation, both perturbed versions will collapse to the same solution.

After constructing this initial point, we iteratively perform the update equations of (5) until convergence. If at the convergence point the behavior of $t_{i,a}$ and $t_{i,b}$ is identical, then we declare that the value t_i has not split. On the other hand, if the distribution $P(t_{i,a} \mid \mathbf{u}_i)$ is sufficiently different from $P(t_{i,b} \mid \mathbf{u}_i)$ for some values of $\mathbf{U}_i$, then we declare that the value t_i has split, and incorporate $t_{i,a}$ and $t_{i,b}$ into the bifurcation we construct for T_i. Finally, we

[3]In deterministic annealing terminology, $\frac{1}{\beta}$ is the "temperature" of the system, and thus increasing β corresponds to "cooling" the system.



(a) (b)

Figure 3: Application of the symmetric bottleneck to the 20 newsgroup data set with 300 informative words. (a) The learned cluster hierarchy of categories. (b) information curves showing the progression along the information tradeoff graph. Note that with 16 word clusters we preserve most of the information present in the data.

increase β and repeat the whole process.

8 Examples

To illustrate the ideas described above, we now examine few applications of symmetric and parallel versions of the bottleneck.

As a simple synthetic example we produced a joint probability matrix $P(A, B)$ (see Figure 2(a)) where $|A| = 80$ and $|B| = 20$. Using the symmetric compression algorithm we find 6 natural clusters for A and 3 natural clusters for B. Sorting the joint probability matrix by this solution illustrates this structure (Figure 2(b)). It is also interesting to see the fraction of information preserved by our clusters. One way of presenting these results is by considering the fraction of information preserved by T_A about B, and analogously, the fraction of information preserved in T_B about A. This amount of preserved information should be plotted w.r.t the compression factor, i.e., how compact is the new clusters representation. This is given of course by $I(T_A; A)$ and $I(T_B; B)$ respectively. In Figure 2(c) we present these two information curves. In both figures we see that splitting the current clusters set increase the amount of information preserved about the relevant variable, and simultaneously reduces the compression (since more clusters induces less compression). In the first split we find 2 clusters in T_A and 2 in T_B. The second split results with 4 clusters in T_A and 3 in T_B. The next split finds 6 clusters in T_A and leave T_B with 3 clusters. This is indeed the "real" structure of this data. Interestingly, due to this last split, the information T_B preserve about A is increased, though there was no split in T_B. The reason, of course, is that T_B predict A *through* T_A, thus the split in T_A increase $I(T_B; A)$. On the other hand, since there was no split (at this step) in T_B, $I(T_A; B)$ remains unchanged. The next splits are practically overfitting effects and accordingly there is no real information gain due to these splits.

As a more realistic example we used the standard "20 newsgroups" corpus. This natural language corpus contains about $20,000$ articles evenly distributed among 20 USENET discussion groups [4] and has been employed for evaluating text classification techniques (e.g., [10]). Many of these groups have similar topics. Five groups discuss different issues concerning computers, three groups discuss religion issues, etc. Thus, there is an inherent hierarchy among these groups.

To model this domain in our setting, we introduce two random variables. We let W denote words, and C denote a category (i.e., a newsgroup). The joint probability $P(w, c)$ is the probability that a random word-position (e.g., word 218 in document 1255) in this collection is equal to w and at the same time the category of the document is c. To obtain such a joint distribution we performed several preprocessing steps: We removed file headers, we transformed all words to lower case, and removed stop words and words containing digits or non-alphanumeric characters. We then sorted all words by the contribution to the mutual information about the category variable. More formally, we sorted all words by $I(w) \equiv P(w) \sum_{c \in C} P(c \mid w) \log \frac{P(c|w)}{P(c)}$, and used the subset of the top 300 most "informative" words. After re-normalization, we had a joint probability matrix with $|W| = 300$ and $|C| = 20$.

We first used the symmetric clustering algorithm to cluster both dimensions of this matrix into two sets of clusters: clusters of words, T_w, and clusters of categories, T_c. The hierarchy found in T_c, shown in Figure 3(a) is in high agreement with the natural hierarchy one would construct. Additionally, each of the word clusters is in high correlation with one of these category clusters. For example, for the second word cluster, the 5 most probable words (i.e. the 5 words that maximize $P(t_{w_2} \mid W)$), were *'islamic', 'religious', 'homosexual', 'peace'* and *'religion'*. Accordingly $P(T_c \mid t_{w_2})$ was maximized for the "religion" cluster in T_c

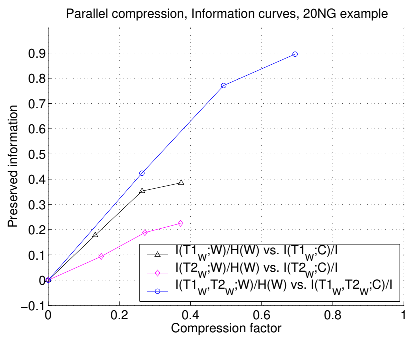


Figure 4: Information curves for parallel compression of the 20 newsgroup data set showing the progression along the information tradeoff graph.

(left cluster in Figure 3(a)).

As already explained, the general mapping scheme we use is a “soft” one. That is, each object could be assigned to each cluster with some normalized probability. The clustering of C into T_c was typically “hard” (for every $c \in C$, $P(T_c \mid c)$ was approximately 1 for one cluster and 0 for the others). Nevertheless, the clustering of W into T_w utilized the “soft” aspect of the clustering to deal with words that are relevant to several category clusters. Thus, some of the words were assigned to more than one cluster. For example, the word *’Clinton’* was assigned to two different word clusters dealing with politics. The word *’sexual’* was assigned to the same two clusters, and also (with lower probability) to a cluster of words dealing with religious issues.

For the same data we used also the parallel compression scheme. In this case we have two compression variables, $T_{1,w}, T_{2,w}$, that try simultaneously (and independently), to cluster the set of words W in a way that will preserve the information about the category variable C, as high as possible. In Figure 4, we present the information curves for these two cluster sets. Clearly, using the combination of the compression variables is much more informative than using each one of them independently. For example, after the second split, $|T_{1,w}| = |T_{2,w}| = 4$, and $I(T_{1,w}; C)$ and $I(T_{2,w}; C)$ preserve 38% and 20% of the original information $I(W; C)$, respectively. On the other hand, at the same stage, $I(T_{1,w}, T_{2,w}; C)$ preserves almost 80% of $I(W; C)$. Thus, only 8 word-clusters are enough to preserve most of the information about the category variable. For comparison, using the (original) single-sided bottleneck, where there is only one compression variable T_w, we found that to preserve the same amount of information we need 16 word clusters instead of 8.

9 Discussion

We presented a novel general framework for data analysis. This new framework provides a natural generalization of the information bottleneck method. Moreover, as we have shown, it immediately suggests new bottleneck like constructions, and provides generic tools to implement them. This allows us to extract structure from data in numerous ways. In our examples, we explored only few relatively simple cases, but clearly this is the tip of the iceberg. We are currently working on several additional applications under this framework. These include analysis of gene expression data, neural coding, document clustering, and computational linguistic applications.

Many connections with other data analysis methods should be explored. The general structure of the iterative procedure is reminiscent of EM and k-means procedures. Other connections are, for example, to dimensionality reduction techniques, such as ICA. The parallel bottleneck construction provides an ICA-like decomposition with an important distinction. In contrast to ICA, it is aimed at preserving information about specific aspects of the data, defined by the user.

The variational formulation of the problem, as a tradeoff between two graphical multi-informations allow us to extend the analogy of the original method with statistical mechanics and study the structure of the data through an hierarchy of phase transitions. It also provide us with the Lyapunov function needed for the convergence of the emerged iterative algorithm.

References

[1] L. D. Baker and A. K. McCallum. Distributional clustering of words for text classification. In *ACM SIGIR 98*. 1998.

[2] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.

[3] T. Hofmann. Probabilistic latent semantic indexing. In *ACM SIGIR 99*, pages 50–57. 1999.

[4] K. Lang. Learning to filter netnews. In *12th Int. Conf. on Machine Learning*, pages 331–339. 1995.

[5] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kauffman, San Francisco, 1988.

[6] F.C. Pereira, N. Tishby, and L. Lee. Distributional clustering of english words. In *30th Annual Meeting of the Association for Computational Linguistics*, pages 183–190. 1993.

[7] K. Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of the IEEE*, 86:2210–2239, 1998.

[8] N. Slonim, R. Somerville, N. Tishby, and O. Lahav. Objective spectral classification of galaxies using the information bottleneck method. *Monthly Notices of the Royal Astronomical Society*, 2001. to appear.

[9] N. Slonim and N. Tishby. Document clustering using word clusters via the information bottleneck method. In *ACM SIGIR 2000*, pages 208–215. 2000.

[10] N. Slonim and N. Tishby. The power of word clusters for text classification. In *23rd European Colloquium on Information Retrieval Research*. 2001. to appear.

[11] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In *Proc. 37th Allerton Conference on Communication and Computation*. 1999.

[12] N. Tishby and N. Slonim. Data clustering by markovian relaxation and the information bottleneck method. In *Neural Information Processing Systems (NIPS-00)*. 2000.