Chapter 5

Extracting relevant structures

A key problem in understanding auditory coding is to identify the acoustic features that neurons at various levels of the system code. If we can map the relevant stimulus features and trace how they change along the processing hierarchy, we can understand the processing properties of the system.

A principled approach for extracting relevant features was proposed by Tishby and co-authors [Tishby *et al.*, 1999], with the *Information Bottleneck* (IB) framework. This powerful approach aims at identifying structures of a variable X that have functional importance, by compressing X in a way that preserves information about another variable Y. In the context of the current problem it can be used to compress acoustic stimuli while preserving information about the distribution of neural responses, and use it to identify the stimulus aspects to which system is sensitive.

Unfortunately, the IB approach is insufficient for characterizing the processing that takes place in a brain region like the cortex. To understand the reason why, consider for example a case where one measures cortical activity in response to acoustic stimuli, and maps the acoustic features to which the cortical neurons respond. Such an analysis does not characterize cortical processing but rather the processing performed by the whole chain of processing stations that ends in the cortex. In fact, many of the features that are revealed this way do not reflect cortical processing, but processing that takes place at lower levels. For example, in the visual system, high level cells are often sensitive to the location of an object in the visual field (e.g. [Levy *et al.*, 2001]), but such spatial sensitivity is computed by the narrow spatial sensitivity of the receptors in the retina. Similarly, in the auditory system, cortical neurons are frequency sensitive, but the frequency content of a stimulus is already determined at the receptor level in the cochlea. The conclusion is that when we aim to identify the processing that occurs at the cortex, we should search for features relevant to cortical neurons and not lower level neurons.

This problem is in fact common to all unsupervised learning tasks, and is caused by an inherent problem with the definition of unsupervised learning and relevance. Namely, data contain several overlapping and competing structures, and the "right" structures depend on the task at hand. We will show here how this problem can be alleviated by using additional "non-interesting" data, such as neural activity in early processing stations. Using this type of irrelevant side data is the subject of the current chapter.

This chapter is organized as follows: Section 5.1 describes the information bottleneck approach. Section 5.2 explains how IB can be extended to use side information that allows to ignore irrelevant structures. Finally section 5.4 describes applications of IBSI in various domains.

5.1 Information bottleneck

5.1.1 Formulation

The information bottleneck method was first presented in [Tishby *et al.*, 1999], and later extended in a series of papers [Slonim and Tishby, 2000, Friedman *et al.*, 2001, Slonim *et al.*, 2001, Slonim *et al.*, 2002]. This framework has proven powerful for numerous applications, such as clustering the objects of sentences with respect to the verbs [Pereira *et al.*, 1993], documents with respect to their terms [Baker and McCallum, 1998, Hoffman, 1999, Slonim and Tishby, 2000], genes with respect to tissues [Friedman *et al.*, 2001, Sinkkonen and Kaski, 2001], and visual stimuli with respect to spike patterns [Schneidman *et al.*, 2002] and vice versa [Dimitrov and Miller, 2001, Gedeon *et al.*, 2002]. We provide here a short review of this general approach. The reader is referred to [Slonim *et al.*, 2002] for a fuller description.

In IB we are given a pair of variables X and Y, and their joint distribution $p(x, y)$. The goal is to compress X into another variable T, while preserving information about the variable Y. This compression is achieved via soft clustering; that is, a stochastic mapping of x's into t's, which we denote $p(t|x)$. This stochastic mapping determines the joint distribution of T, Y via the Markov relation $p(y|t) = \sum_x p(y|x)p(x|t)$. IB operates to learn a mapping that minimizes the information $I(X;T)$, while at the same time maximizing the information $I(T;Y)$ This is formally cast as a weighted tradeoff optimization problem

$$\min I(X;T) - \beta I(T;Y) \tag{5.1}$$

where β is the tradeoff parameter between compression and information preservation. It can also be shown [Gilad-Bachrach *et al.*, 2003] that this problem is equivalent to a constrained optimization problem of minimizing $I(X;T)$ under a lower bound on $I(T;Y)$. In the constrained optimization problem β takes the role of a Lagrange multiplier.

The full formal optimization problem is written as

$$\mathcal{L} = I(X;T) - \beta I(T;Y) - \sum_x \lambda_x (\sum_t p(t|x) - 1) \tag{5.2}$$

where λ_x are Lagrange multipliers and the last term is used to enforce the normalization constraint $1 = \sum_t p(t|x)\ \forall x$.

The target function 5.2 can be differentiated w.r.t. $p(t|x)$ to find an optimal mapping for any predefined value of the parameter β. Fortunately this yields a set of self-consistent equations that the optimum obeys

$$\begin{aligned} p(t|x) &= \frac{p(t)}{Z} e^{-\beta D_{KL}[p(y|x)||p(y|t)]} \\ p(t) &= \sum_x p(t|x)p(x) \\ p(y|t) &= \frac{1}{p(t)} \sum_x p(y|x)p(t|x)p(x) \end{aligned} \tag{5.3}$$

where

$$Z = \sum_t p(t) e^{-\beta D_{KL}[p(y|x)||p(y|t)]} \tag{5.4}$$

is a normalization factor, known in statistical mechanics as the partition function.

5.1.2 IB algorithms

A series of algorithms was developed for solving the IB variational problem of Eq. 5.2. The self consistent equations 5.3 were used in [Tishby *et al.*, 1999] to devise an iterative algorithm, in the spirit of the Blahut algorithm used in rate distortion theory [Blahut, 1972, Csiszar, 1974]. This iterative algorithm named iIB by [Slonim *et al.*, 2002] operates to optimize three free variables: $p(t|x)$, $p(t)$ and $p(y|t)$. It is based on the fact that the IB target function is convex w.r.t. each of the three variables independently but is not jointly convex. The iIB algorithm repeatedly fixes two of the free variables to their current values and optimizes over the third one. The pseudo code of iIB appears in figure 5.1.

Iterative IB

Input:
  Joint distribution $p(x, y)$.
  Trade-off parameter β .
  Cardinality parameter K and a convergence parameter ϵ .

Output:
  A (typically "soft") partition T of X into K clusters.

Initialization:
  Randomly initialize $p(t \mid x)$ and find the corresponding $p(t),\ p(y|t)$ through Eq. 5.3.

Main loop

repeat

$$p_{i+1}(t|x) \leftarrow \frac{p_i(t)}{Z_{i+1}(x,\beta)} e^{-\beta D_{KL}[p(y|x)||p(y|t)]} \ ,\ \forall\, t \in T,\ \forall\, x \in X$$

$$p_{i+1}(t) \leftarrow \sum_x p(x) p_{i+1}(t|x) \ ,\ \forall\, t \in T$$

$$p_{i+1}(y|t) = \frac{1}{p_{i+1}(t)} \sum_x p_{i+1}(t|x) p(x, y) \ ,\ \forall\, t \in T,\ \forall\, y \in Y$$

until (stopping criterion)

Figure 5.1: Pseudo-code of the iterative IB (iIB) algorithm. The stopping criterion used in (Slonim, 2002) is to stop if $\forall\, x \in X,\ JS_{\frac{1}{2},\frac{1}{2}}[p_{i+1}(t|x), p_i(t|x)] \le \epsilon$, where JS is the Jensen-Shannon divergence defined in **??**. In principle, this procedure only guarantees convergence to a local extremum, hence it should be repeated several times with different initializations, and the solution which minimizes the target function $L = I(T; X) - \beta I(T; Y)$ should be chosen.

In many applications, hard clustering solutions can be more easily interpreted. This is particularly true when dealing with categorical data where averaging elements is meaningless. A series of hard clustering algorithms that operate to optimize the IB functional were suggested based on widely known clustering methods such as hierarchical clustering and K-means.

Among these hard clustering methods, the one that was found to be specifically useful is *sequential-IB* (sIB) [Slonim *et al.*, 2001]. In sIB, one starts with a random assignment of elements to clusters, and then iteratively improves the clustering by almost-greedy steps. At each such step, an element is taken randomly out of its cluster and is assigned to a cluster such that the overall score will be maximally improved. A pseudo code of this algorithm appears in Figure 5.2.

Sequential IB

Input:
  Joint distribution $p(x, y)$.
  Trade-off parameter β.
  Cardinality value K.

Output:
  A partition T of X into K clusters.

Initialization:
  $T \leftarrow$ random partition of X into K clusters

Main Loop:
  while not *done*
    $done \leftarrow TRUE$
    for every $x \in X$:
      Remove x from current cluster $t(x)$
      $t^{new}(x) \leftarrow argmin_{t \in T} \Delta L(\{x\}, t)$
      if $t^{new}(x) \neq t(x)$
        $done \leftarrow FALSE$.
      Merge x into $t^{new}(x)$
    end for
  end while

Figure 5.2: Pseudo-code of the sequential-IB (sIB) algorithm. In principle, this procedure only guarantees convergence to a local extremum, hence should be repeated for different initializations, and the solution which maximizes the target function $L = I(X;T) - \beta I(T;Y)$ should be chosen.

5.2 Relevant and irrelevant structures

5.2.1 The problem

A fundamental goal of machine learning is to find regular structures in given empirical data, and use them to construct predictive or comprehensible models. This general goal, unfortunately, is very ill defined, as many data sets contain alternative, often conflicting, underlying structures. For example, documents may be classified either by subject or by writing style; spoken words can be labeled by their meaning or by the identity of the speaker; proteins can be classified by their structure or function - all are valid alternatives. Which of these alternative structures is "relevant" is often implicit in the problem formulation.

The problem of identifying "the" relevant structures is commonly addressed in supervised learning tasks, by providing a "relevant" label to the data, and selecting features that are discriminative with respect to this label. As described in the previous section, an information theoretic generalization of this supervised approach has been proposed in [Pereira *et al.*, 1993, Tishby *et al.*, 1999] through the "information bottleneck method" (IB).

An important condition for this approach to work is that the auxiliary variable indeed corresponds to the task. In many situations, however, such a "pure" variable is not available. The auxiliary variable may in fact contain alternative and even conflicting structures. We show here that this general and common problem can be alleviated by providing "negative information"; that is, information about "non-important", or irrelevant, aspects of the data that can interfere with the desired structure during learning.

As an illustration, consider a simple nonlinear regression problem. Two variables x and y are related through a functional form $y = f(x) + \xi$, where $f(x)$ is in some known function class and ξ is noise with some distribution that *depends on* x. When given a sample of (x, y) pairs with the goal of extracting the relevant dependence $y = f(x)$, the noise ξ - which may contain information on x and thus interfere with extracting y - is an irrelevant variable. Knowing the joint distribution of (x, ξ) can of course improve the regression result.

The problem of identifying stimulus features that are relevant for neural activity, presented at the beginning of this chapter is a more real-life example. Another real world example can be found in the analysis of gene expression data. Such data, as generated by DNA-chips technology, can be considered as an empirical joint distribution of gene expression levels and different tissues, where the tissues are taken from different biological conditions and pathologies. The search for expressed genes that testify to the existence of a pathology may be obscured by genetic correlations that also exist in other conditions. Here again a sample of irrelevant expression data, taken for instance from a healthy population, can enable clustering analysis to focus on the pathological features alone, and ignore spurious structures.

These examples, and numerous others, are all instantiations of a common problem: in order to better extract the relevant structures information about the irrelevant components of the data should be used. Naturally, various solutions have been suggested to this basic problem in many different contexts (e.g. spectral subtraction, weighted regression analysis). The section below presents a general unified information theoretic framework for such problems, extending the original information bottleneck variational problem to deal with discriminative tasks of this nature.

5.2.2 Information theoretic formulation

To formalize the problem of extracting relevant structures consider first three categorical variables X, Y^+ and Y^- whose co-occurrence distributions are known. Our goal is to uncover structures in $P(X, Y^+)$, that do not exist in $P(X, Y^-)$. The distribution $P(X; Y^+)$ may contain several conflicting underlying structures, some of which may also exist in $P(X, Y^-)$. These variables stand for example for a set of stimuli X, a set of neural responses Y^+ from a brain region whose code we wish to explore, and an additional set of neural responses Y^-. Other examples can be a set of terms and two sets of documents or a set of genes and two sets of tissues with different biological conditions. In all these examples Y^+ and Y^- are *conditionally independent* given X. We thus make the assumption that the joint distribution factorizes as $p(x, y^+, y^-) = p(x)p(y^+|x)p(y^-|x)$.

The relationship between the variables can be expressed by a Venn diagram (Figure 5.3), where the area of each circle corresponds to the entropy of a variable (see e.g. [Cover and Thomas, 1991] p.20 and [Csiszar and J.Korner, 1997] p.50 for discussion of this type of diagrams) and the intersection of two circles corresponds to their mutual information.

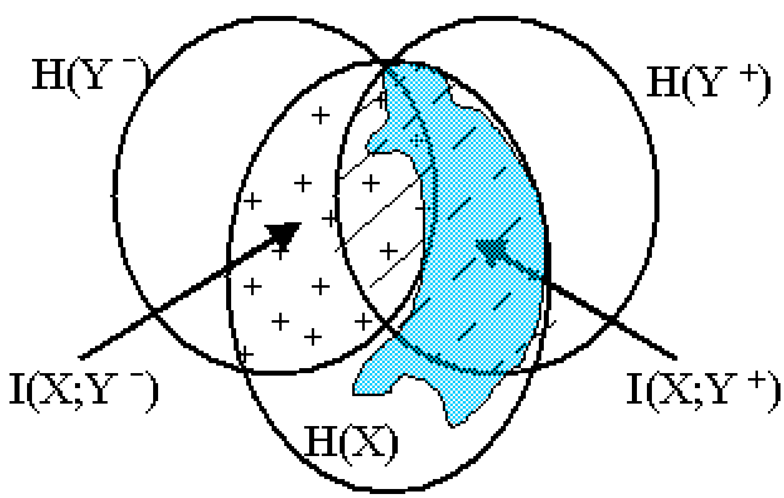

Figure 5.3: A Venn diagram illustrating the relations between the entropy and mutual information of the variables X,Y^+,Y^-. The area of each circle corresponds to the entropy of a variable, while the intersection of two circles corresponds to their mutual information. As Y^+ and Y^- are independent given X, their mutual information vanishes when x is known, thus all their overlap is included in the circle of X.

To identify the relevant structures in the joint distribution $p(x, y^+)$, we aim to extract a compact representation of the variable X with minimal loss of mutual information about the relevant variable Y^+, and *at the same time* with maximal loss of information about the irrelevance variable Y^-. The goal of *information bottleneck with side information* (IBSI) is therefore to find a stochastic map of X to a new variable T, $p(t|x)$, in a way that maximizes its mutual information about Y^+ and minimizes the mutual information about Y^-. In general one can only achieve this goal perfectly in the asymptotic case and the finite case leads to a sub optimal compression, an example of which is depicted in the blue region in figure 5.3. These constraints can be cast into a single variational functional that we aim to minimize

$$\mathcal{L} = I(X;T) - \beta \left[I(T;Y^+) - \gamma I(T;Y^-) \right] \ . \tag{5.5}$$

This functional consists of three terms that quantify *compression* ($I(T;X)$), information preservations ($I(T;Y^+)$), and information removal ($I(T;Y^-)$). The Lagrange multiplier β determines the tradeoff between compression and information extraction while the parameter γ determines the tradeoff between preservation of information about the relevant Y^+ variable and removal of information about the irrelevant one Y^-. In some applications, such as in communication, the value of γ may be determined by the relative cost of transmitting the information about Y^- by other means (see [Wyner, 1975]). In others, as shown for example in [Globerson *et al.*, 2003], real systems exhibit phase transitions phenomena at designated values of γ, which allow to identify the transition points in the parameter space and reveal the qualitatively different regimes of the system's states space

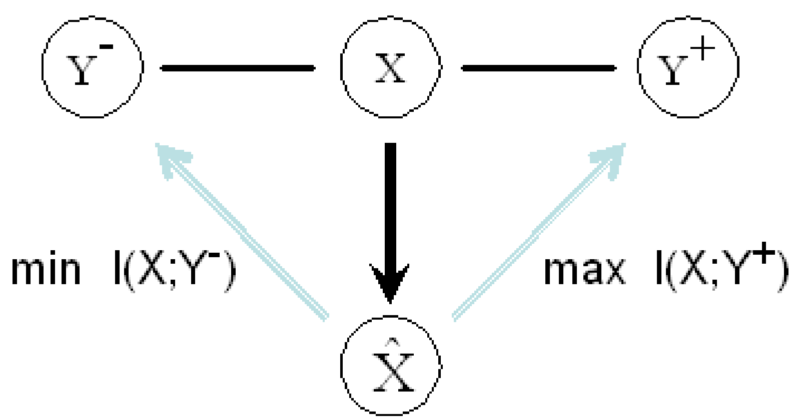

Figure 5.4: A graphic representation of discriminative IB. Given the three variables X,Y^+ and Y^-, we seek a compact stochastic representation T of X which preserves information about Y^+ but removes information about Y^-. In this graph Y^+ and Y^- are indeed conditionally independent given X.

The structure of the data in the original information bottleneck formulation is characterized by the information curve; namely, the values of the IB functional as a function of different β values. In IBSI the data are similarly characterized by the functional values but this time as a function of two free parameters, β and γ, creating a two dimensional manifold. Such a manifold is demonstrated in Figure 5.5.

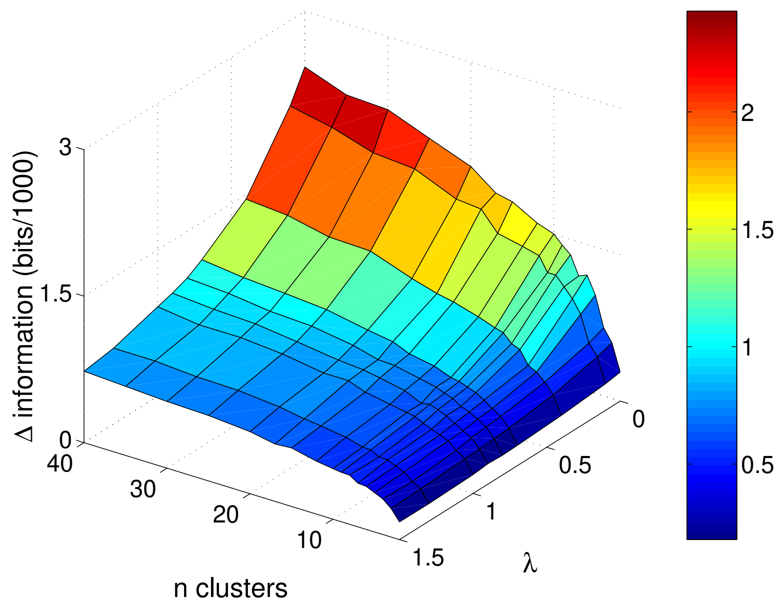

Figure 5.5: The information plane. The value of the IBSI functional as a function of the two Lagrange coefficient values β and γ. The information plane plotted in this figure was calculated using face images as described in section 5.4.4.

The IB variational problem of Eq. 5.1 is a special case of our current variational problem with $\gamma = 0$, namely, no side or irrelevant information is available. In this case only the distributions $p(t|x)$, $p(t)$ and $p(y^+|t)$ are determined.

5.2.3 Solution characterization

The complete Lagrangian of this constrained optimization problem is given by

$$\begin{aligned}\mathcal{L}[p(t|x)] &= I(X;T) - \beta\left[I(T;Y^+) - \gamma I(T;Y^-)\right] \\ &\quad - \sum_x \lambda(x) \sum_t p(t|x)\end{aligned} \tag{5.6}$$

where $\lambda(x)$, are the normalization Lagrange multipliers that enforce the constraint $\sum_t p(t|x) = 1$. Here, the minimization is performed with respect to the stochastic mapping $p(t|x)$, taking into account its probabilistic relations to $p(z)$, $p(y^+|z)$ and $p(y^-)$.

Theorem 5.2.1 : *The extrema of $\mathcal{L}$ obey the following self consistent equations*

$$\begin{aligned} p(t|x) &= \frac{p(t)}{Z} e^{-\beta\left(D_{KL}\left[p(y^+|x)||p(y^+|t)\right]-\gamma D_{KL}\left[p(y^-|x)||p(y^-|t)\right]\right)} \\ p(t) &= \sum_x p(t|x)p(x) \\ p(y^+|t) &= \frac{1}{p(t)}\sum_x p(y^+|x)p(t|x)p(x) \\ p(y^-|t) &= \frac{1}{p(t)}\sum_x p(y^-|x)p(t|x)p(x) \end{aligned} \tag{5.7}$$

where

$$Z = \sum_t p(t) e^{-\beta\left(D_{KL}\left[p(y^+|x)||p(y^+|t)\right]-\gamma D_{KL}\left[p(y^-|x)||p(y^-|t)\right]\right)} \tag{5.8}$$

is a normalization factor.

Proof: Following the Markov relation $p(y|x,t) = p(y|x)$, we write

$$\begin{aligned} p(y,t) &= \sum_x p(y,t|x)p(x) \\ &= \sum_x p(y|t,x)p(t|x)p(x) = \sum_x p(y|x)p(t|x)p(x) \end{aligned} \tag{5.9}$$

where $p(x) = \sum_{y^+}\sum_{y^-} p(y^+,y^-,x)$, and obtain for the second term of Eq. 5.6

$$\begin{aligned} \frac{\delta I(T;Y^+)}{\delta p(t|x)} &= \frac{\delta}{\delta p(t|x)} \sum_{t'}\sum_{y^+}\sum_{x'} p(y^+|x')p(t'|x')p(x') \times \\ &\quad \times \log\left(\frac{p(y^+|t')}{p(y^+)}\right) \\ &= p(x)\sum_{y^+} p(y^+|x)\log\left(\frac{p(y^+|t)}{p(y^+|x)}\frac{p(y^+|x)}{p(y^+)}\right) \\ &= -p(x)D_{KL}[p(y^+|x)||p(y^+|t)] \\ &\quad +p(x)D_{KL}[p(y^+|x)||p(y^+)] \end{aligned} \tag{5.10}$$

a similar differentiation for the remaining terms yields

$$\begin{aligned} \frac{\delta}{\delta p(t|x)}\mathcal{L} &= p(x)\log\left(\frac{p(t|x)}{p(t)}\right) \\ &\quad -p(x)\beta D_{KL}[p(y^+|x)||p(y^+|t)] \\ &\quad -p(x)\beta\gamma D_{KL}[p(y^-|x)||p(y^-|t)] \\ &\quad +p(x)\lambda(x,y^+,y^-) \end{aligned} \tag{5.11}$$

where

$$\begin{aligned}\lambda(x, y^+, y^-) &= \frac{\lambda(x)}{p(x)} + \\ &\quad \beta \left(D_{KL}[p(y^+|x)||p(y^+)] - \gamma D_{KL}[p(y^-|x)||p(y^-)] \right)\end{aligned} \tag{5.12}$$

holds all terms independent of t. Equating the derivative to zero then yields the first equation of proposition 5.2.1. The remaining equations hold due to the Markov relations $p(y^+|t,x) = p(y^+|x)$, $p(y^-|t,x) = p(y^-|x)$. $\square$

The formal solutions of the above variational problem have an exponential form which is a natural generalization of the solution of the original IB problem (Eq. 5.2).

5.2.4 IBSI and discriminative models

The side information formalism naturally raises the question of its relation to discriminative models.

Two types of discriminative models come to mind in this context. First, one may consider a learning task whose goal is to discriminate between samples from Y^+ and Y^-. This task, widely studied in the context of binary classification is inherently different from the task we are interested in here. In the second type of tasks, we are interested in discriminative training of mixture models for $p(X, Y^+)$ and $p(X, Y^-)$. This is the subject of the section here.

Consider a mixture model in which we generate an observation y for a given x, by the following procedure: At first, every x is assigned to a cluster t according to a prior distribution $\pi(t)$. This cluster determines a distribution $p(y|t(x))$, parametrized by $\theta(y|t(x))$. Then, samples of (x, y) pairs are generated as follows: First, a value of x is chosen according to a prior distribution $p_x(x)$. Then, multiple samples of y are drawn from the distribution $p(y|t(x))$. We assume that y takes discrete values.

Let us denote the empirical count generated by this process by $n^+(x_i, y_j)$. We also assume that an additional count was generated by a similar process, and denote it by $n^-(x_i, y_j)$. The cardinality $|X|$ is the same for the two counts, but the remaining parameters are not necessarily equivalent.

In *generative training*, one aims at finding the parameters $\theta(y|t)$ (clusters' centroids) and $t(x)$ (cluster assignments), such that the likelihood of the count given the parameters is maximized. The likelihood of the n^+ count is formally written as

$$\begin{aligned}&L(n^+ \mid \pi, \theta^+, {p_x}^+, t(x)) \\ &\quad = \prod_{i=1}^{|X|} \pi(t(x_i)) \prod_{l=1}^{n} {p_x}^+(x_l) \theta(y|t(x_l))\end{aligned} \tag{5.13}$$

$$= \prod_{i=1}^{|X|} \pi(t(x_i)) \prod_{j=1}^{|X|} \prod_{j=1}^{|Y|} \left[{p_x}^+(x_i)\theta(y_j|t(x_i)) \right]^{n(x_i,y_j)}$$

and similarly for the n^- count. Searching for the parameters that maximize this likelihood can be achieved using standard maximum likelihood algorithms for mixture models, such as Expectation Maximization (EM) [Dempster *et al.*, 1977]. These are applied *separately* on the counts n^+ and n^- to estimate the parameters that maximize each of the likelihood terms.

In *discriminative training*, one aims at finding the parameters that maximize the log likelihood ratio

$$\max_{\pi,\theta^+,\theta^-,{p_x}^+,{p_x}^-,t(x)} \log \left(\frac{L(n^+|\pi, \theta^+, {p_x}^+, t(x))}{L(n^-|\pi, \theta^-, {p_x}^-, t(x))} \right) \tag{5.14}$$

where maximization allows to fit the distribution parameters p_x and θ separately for each of the counts, **but the prior $\pi(t)$ and the assignment $t(x)$ is common to both counts**.

This log likelihood ratio equals

$$\begin{aligned} R &= \log \left(\frac{L(n^+|\pi, \theta^+, {p_x}^+, t(x))}{L(n^-|\pi, \theta^-, {p_x}^-, t(x))} \right) = \\ &= \sum_{i=1}^{|X|} \pi(t(x_i)) + \sum_{i=1}^{|X|} \sum_{j=1}^{|Y^+|} n^+(x_i, y_j^+) \log \left({p_x}^+(x_i)\theta^+(y_j^+|t(x_i)) \right) \\ &\quad - \sum_{i=1}^{|X|} \pi(t(x_i)) - \sum_{i=1}^{|X|} \sum_{j=1}^{|Y^-|} n^-(x_i, y_j^-) \log \left({p_x}^-(x_i)\theta^-(y_j^-|t(x_i)) \right) \end{aligned} \tag{5.15}$$

In the limit of large counts $n \equiv \sum_{i=1}^{|X|} \sum_{j=1}^{|Y|} n(x_i, y_j) \to \infty$ we have $\frac{1}{n} n(x_i, y_j) \to p(x_i, y_j)$, and the likelihood ratio converges to

$$\begin{aligned} R &\to n \sum_{i=1}^{|X|} \sum_{j=1}^{|Y^+|} p^+(x_i, y_j) \log \left({p_x}^+(x_i)\theta^+(y_j|t(x_i)) \right) \\ &\quad -n \sum_{i=1}^{|X|} \sum_{j=1}^{|Y^-|} p^-(x_i, y_j) \log \left({p_x}^-(x_i)\theta^-(y_j|t(x_i)) \right) \\ &= nE \left[\log \left(\frac{{p_x}^+(x_i)\theta^+(y_j|t(x_i))}{{p_x}^-(x_i)\theta^-(y_j|t(x_i))} \right) \right] \end{aligned} \tag{5.16}$$

where expectation is over the joint distribution $p(x, y+, y-)$. We conclude that this discriminative training of mixture models aims at maximizing the expected log likelihood ratio, which has two terms: the priors $\frac{{p_x}^+(x_i)}{{p_x}^-(x_i)}$ and the conditionals $\frac{\theta^+(y_j|t(x_i))}{\theta^-(y_j|t(x_i))}$.

For the purpose of comparison, we now rewrite the information preservation and removal terms of IBSI (i.e. taking $\beta \to \infty$)

$$\begin{aligned}
I(T;Y^+) \quad - & \quad \gamma I(T;Y^-) = \\
= & \quad \sum_t \sum_{y^+} \sum_{y^-} p(t, y^+, y^-) \log\left(\frac{p(y^+|t)}{p(y^+)}\right) \\
& \quad -\gamma \sum_t \sum_{y^+} \sum_{y^-} p(t, y^+, y^-) \log\left(\frac{p(y^-|t)}{p(y^-)}\right) \\
= & \quad \left\langle \log\left(\frac{p(y^+|t)}{p(y^-|t)^\gamma} \frac{p(y^-)^\gamma}{p(y^+)}\right)\right\rangle_{p(t,y^+,y^-)}
\end{aligned} \tag{5.17}$$

This means that when $\gamma = 1$, and the marginal distributions are equal ${p_x}^+(x) = {p_x}^-(x)$, the two expressions become similar. Moreover, when the cardinality of Y^+ and Y^- is the same and the marginal distributions of Y^+ and Y^- are uniform, the last equation becomes identical to the discriminative training of the mixture model. In the general case however such an equivalence does not hold.

5.2.5 Multivariate extensions

The above setup can be extended to the case of multiple variables on which multi-information should be preserved about $\{y_1^+, ..., y_{N^+}^+\}$ and variables on which multi-information should be removed about $\{y_1^-, ..., y_{N^-}^-\}$, as discussed in [Friedman *et al.*, 2001]. This yields

$$\begin{aligned}
\log \frac{p(t|x)}{p(t)} \quad \propto & \quad -\log(Z) - \sum_i \gamma_i^+ D_{KL}\Big[p(y_i^+|x)||p(y_i^+|t)\Big] \\
& \quad + \sum_i \gamma_i^- D_{KL}\Big[p(y_i^-|x)||p(y_i^-|t)\Big]
\end{aligned} \tag{5.18}$$

which can be solved together with the other self-consistent conditions, similarly to Eq. 5.8.

Iterative IBSI

Input:
Joint distributions $P(x, y^+)$, $P(x, y^-)$
Trade-off parameters β,γ
Number of clusters K
Stopping parameter ϵ

Output:
A soft partition $P(T|X)$ of X into T using K clusters.

Main:
Initialize $p(t|x)$ and set i=0.
Calculate $p_i(t)$,$p_i(y^+|t)$ and $p_i(y^- \mid t)$ through Eq. 5.7.
repeat

$i \leftarrow i + 1$

$p_i(y^+|t) \leftarrow \frac{1}{p_{i-1}(t)} \sum_x p_{i-1}(t|x)p(x, y^+)$, $\forall t \in T$, $\forall y^+ \in Y^+$

$p_i(y^-|t) \leftarrow \frac{1}{p_{i-1}(t)} \sum_x p_{i-1}(t|x)p(x, y^-)$, $\forall t \in T$, $\forall y^- \in Y^-$.

$p_i(t|x) \leftarrow \frac{p_{i-1}(t)}{Z^i} \times$

$\exp\left(-\beta(D_{KL}[p(y^+|x)||p_i(y^+|t)] - \gamma D_{KL}[p(y^-|x)||p_i(y^-|t)])\right)$

$\forall t \in T$, $\forall x \in X$

$p_i(t) \leftarrow \sum_x p(x)p_{i+1}(t|x)$, $\forall t \in T$

until $\forall\, x \in X$, $JS[p_i(t|x), p_{i-1}(t|x)] \leq \epsilon$.

Figure 5.6: Pseudo-code of the iterative IBSI algorithm. JS denotes the Jensen-Shannon divergence defined in **??**, and Z the partition (normalization) function. This procedure only guarantees convergence to a local extremum, hence should be repeated for different initializations and a solution that maximizes $\mathcal{L} = I(T; X) - \beta I(T; Y^+) + \beta\gamma I(T; Y^-)$ should be chosen.

5.3 IBSI algorithms

The common formulation of IB and IBSI allows us to adopt a series of algorithms that were originally developed for IB. A detailed description of these original algorithms can be found in [Slonim *et al.*, 2002], together with a comparison of their performance on text categorization data. Our main interest in this section is the novel difficulties that the IBSI formulation poses. We therefore focus on two algorithms. The first is an iterative Blahut-Arimoto style algorithm that iterates between the self consistent equations 5.7 derived above. The second is a heuristic sequential K-means algorithms, that was shown by [Slonim *et al.*, 2002] to achieve good empiric results for the IB case ($\gamma = 0$).

5.3.1 Iterating the fix point equations

The optimization problem defined in Eq. 5.5 requires finding the optimum over four parameters $p(t|x)$, $p(t)$, $p(y^+|t)$ and $p(y^-|t)$. Fortunately, even though the Lagrangian 5.6 is not jointly convex with respect to these parameters, it is convex w.r.t. $p(t|x)$ and $p(t)$ separately under some conditions. This convexity allows us to use the fixed points equations of 5.7 in an iterative manner to approach the optimum. This is achieved by iteratively fixing all the equations but one, and optimizing over the non freezed parameter. This calculation is performed for all values of $x \in X$, $t \in T$, $y^+ \in Y^+$ and $y^- \in Y^-$, and is repeated until convergence.

This algorithm, *iterative-IBSI*, whose pseudo code is given in figure 5.6, is similar in form but inherently different from the iterative-IB (iIB) algorithm described in [Slonim *et al.*, 2001]. This is due to the additional term in the IBSI target function, that makes the problem concave rather than convex with respect to $p(y^-|t)$. Instead of iteratively optimizing over the four free parameters $\{p(t|x), p(t), p(y^+|t), p(y^-|t)\}$, we use the Markovian relation for $p(y|t)$ explicitly, and optimize over $p(t|x)$ and $p(t)$. While convergence is always achieved with the original IB algorithm, convergence for IBSI is only guaranteed for a limited range of γ values. The following theorem establishes that such a range of parameters indeed exists.

Theorem 5.3.1 : Convergence of iterative IBSI
There exists a range of γ values for which iterative-IBSI converges to a stationary fixed point of the IBSI functional.

The proof follows the following line: We define an auxiliary (Lyapunov) function $\mathcal{F}$, and show that it is bounded from below and convex w.r.t. the relevant parameters for some non-empty region of γ. We then show that $\mathcal{F}$ decreases at each step of the algorithm unless it has reached a fixed point.

We begin the proof by defining the free energy functional for the IBSI

Definition 5.3.2 : IBSI Free energy
The free energy for a given $p(x,t)$ and $p(t)$ is defined as

$$\mathcal{F} \equiv -\langle \log(Z) \rangle = -\sum_{x,t} p(x)p(t \mid x) \log Z(x,\beta,\gamma) \ , \tag{5.19}$$

where $Z(x, \beta, \gamma)$ is defined by

$$
\begin{aligned}
-\log Z(x,\beta,\gamma) &= \log\frac{p(t|x)}{p(t)} + \beta D_{KL}[p(y^+|x)||p(y^+|t)] \\
&\quad -\beta\gamma D_{KL}[p(y^-|x)||p(y^-|t)] \quad .
\end{aligned}
\tag{5.20}
$$

At the extremum points, when all four equations of 5.7 hold, Z is the normalization (partition) function of $p(t|x)$, as follows from equation 5.7. Substituting Z into 5.19 we have

$$
\begin{aligned}
\mathcal{F} &= \sum_{x,t} p(x,t)\log\frac{p(t|x)}{p(t)} + \beta\sum_{x,t} p(x,t) D_{KL}[p(y^+|x)||p(y^+|t)] \\
&\quad -\beta\gamma\sum_{x,t} p(x,t) D_{KL}[p(y^-|x)||p(y^-|t)]
\end{aligned}
\tag{5.21}
$$

Lemma 5.3.3 : *$\mathcal{F}$ equals the IBSI functional $\mathcal{L}$ up to an additive constant.*

Proof: We rewrite the average of $D_{KL}[p(y|x)||p(y|t)]$ using the Markov relation $p(y|t) = \sum_x p(y|x)p(x|t)$ to obtain

$$
\begin{aligned}
\sum_{x,t} p(x,t) D_{KL}[p(y|x) \; || \; p(y|t)] &= \sum_{x,t} p(x,t) \sum_y p(y|x)\log\left(\frac{p(y|x)}{p(y|t)}\right) \\
&= \sum_x p(x)\sum_y p(y|x)\log(p(y|x)) \\
&\quad - \sum_t p(t)\sum_y \log(p(y|t))\sum_x p(y|x)p(x|t) \\
&= -H(Y|X) - \sum_t p(t)\sum_y \log(p(y|t))p(y|t) \\
&= -H(Y|X) + H(Y|T) \\
&= -H(Y|X) - I(Y;T) + H(Y) \\
&= I(X;Y) - I(Y;T) \quad .
\end{aligned}
\tag{5.22}
$$

Therefore

$$
\begin{aligned}
\mathcal{F} &= I(T;X) + \beta D_{KL}[p(y^+|x)||p(y^+|t)] \\
&\qquad -\beta\gamma D_{KL}[p(y^-|x)||p(y^-|t)] \\
&= I(T;X) + \beta(I(X;Y^+) - I(T;Y^+)) \\
&\qquad -\beta\gamma(I(X;Y^-) - I(T;Y^-)) \\
&= \mathcal{L} + \beta I(X;Y^+) - \beta\gamma I(X;Y^-) \quad ,
\end{aligned}
\tag{5.23}
$$

Thats is, $\mathcal{F}$ equals $\mathcal{L}$ up to an additive constant that depends on the empirical distributions $p(X,Y^+)$ and $p(X,Y^-)$. $\square$

Lemma 5.3.4 : *$\mathcal{F}$ is bounded from below.*

Proof: Following the lemma 5.3.3, we denote $c \equiv \beta I(X;Y^+) - \beta\gamma I(X;Y^-)$ and write

$$\begin{aligned} \mathcal{F} &= c + \mathcal{L} \qquad (5.24) \\ &= c + I(T;X) - \beta I(T;Y^+) + \beta\gamma I(T;Y^-) \\ &\geq c + 0 - \beta I(T;Y^+) + 0 \\ &\geq c - \beta H(Y^+) \end{aligned}$$

and the last term is a constant. □

Two observations should be made at this point. First, $\mathcal{F}$ is not convex w.r.t. $p(y^-|t)$, but concave. Secondly, $\mathcal{F}$ is not jointly convex w.r.t. all its parameters. With all this, we now characterize a weaker form of convexity that guarantees convergence of the algorithm.

Lemma 5.3.5 : *Under the Markov relations $p(y^+|t) = \sum_x p(y|x)p(x|t)$, $p(y^-|t) = \sum_x p(y^-|x)p(x|t)$, there exists a range of γ values for which $\mathcal{F}$ is convex with respect to $p(t)$ and to $p(t|x)$ independently.*

Proof: The convexity of $\mathcal{F}$ w.r.t. $p(t)$ stems directly from the concavity of the log function, and holds for all values of γ.

From the definition of the $\mathcal{F}$ (equation 5.21), it is a weighted linear combination of D_{KL} terms, where all the terms with negative coefficients are weighted by γ. We can therefore use the convexity of D_{KL} to infer the convexity of this linear combination.

Since $D_{KL}[p||q]$ is a convex function of q (see e.g. [Cover and Thomas, 1991], Chapter 2), $D_{KL}[p(y^+|x)||p(y^+|t)]$ is convex with respect to $p(y^+|t)$ for all y and t. Thus, under the Markov relation $p(y^+|t) = \sum_x p(y^+|x)p(x|t)$, $D_{KL}[p(y^+|x)||p(y^+|t)]$ is a convex function of $p(x|t)$, and its second derivative w.r.t. $p(x|t)$ is strictly negative. So is its derivative w.r.t $p(t|x)$ Similarly, $-D_{KL}[p(y^-|x)||p(y^-|t)]$ is concave w.r.t. $p(x|t)$, and its second derivative is positive. Since all the terms with positive second derivatives are weighted by γ, there exist a critical value $\gamma^{max} > 0$ for which the second derivative of the linear combination is strictly negative, thus the functional is convex w.r.t. $p(x|t)$. □

We are now ready to prove theorem 5.3.1.

Proof: Convergence of iterative-IBSI

At each step of the algorithm, we first use the Markov relation to set $p(y^+|t)$ and $p(y^-|t)$, and then set $p(x|t)$ and $p(t)$ such that the first derivative of $\mathcal{L}$ vanishes. Since $\mathcal{F}$ equals $\mathcal{L}$ up to a constant, it also zeroes the first derivative of $\mathcal{F}$. Moreover, since $\mathcal{F}$ is convex w.r.t. $p(t|x)$ and $p(t)$, it decreases the value of $\mathcal{F}$. In addition, since $\mathcal{F}$ is bounded from below, the algorithm must converge, and reach the stopping criterion. Finally, consider the case where iterative-IBSI reached a fixed point of $\mathcal{F}$; that is, subsequent steps of the algorithm do not change $\mathcal{F}$ any longer. This only happens when the fixed points equations are satisfied, i.e. the algorithm has reached a local minima of $\mathcal{L}$. $\square$

Note however that this only proves the convergence of the algorithm to a stationary point w.r.t the target function $\mathcal{L}$ and not w.r.t. $p(t|x)$. If several local minima exist with the same value it does not rule out the possibility that the distribution cycles between them on consecutive steps of the algorithm.

5.3.2 Hard clustering algorithms

As in the case of IB, various heuristics can be applied, such as deterministic annealing - in which increasing the parameter β is used to obtain finer clusters; greedy agglomerative hard clustering [Slonim and Tishby, 1999]; or a sequential K-means like algorithm (sIB) [Slonim *et al.*, 2002]. The latter provides a good compromise between top-down annealing and agglomerative greedy approaches and achieves excellent performance, and its pseudo-code is given in Figure 5.7.

Sequential IBSI

<u>**Input:**</u>
Joint distributions $p(X, Y^+)$, $p(X, Y^-)$
Trade-off parameters β,γ
Number of clusters K

<u>**Output:**</u>
A hard partition $P(T|X)$ of X into T using K clusters.

<u>**Main:**</u>
Randomly initialize $p(t|x)$ and set i=0.
while not *Done*
  $Done \leftarrow TRUE$.
  for every $x \in X$:
    Remove x from current cluster, $t(x)$.
    $t^{new}(x) = argmin_{t\in T} \Delta L_{max}(\{x\}, t)$
    if $t^{new}(x) \neq t(x)$,
      $Done \leftarrow FALSE$.
    Merge x into $t^{new}(x)$
  end for
end while

Figure 5.7: Pseudo-code of sequential IBSI algorithm. The hard clustering is represented in the mapping $p(t|x)$, whose values are either 0 or $1/|X|$. JS denotes the Jensen-Shannon divergence defined in **??**, and Z the partition (normalization) function. This procedure should in principle be repeated for different initializations and then the solution that optimizes the target function $\mathcal{L} = I(T; X) - \beta I(T; Y^+) + \beta\gamma I(T; Y^-$. should be chosen

5.4 Applications

We describe several applications of IBSI. First we illustrate its operation using a simple synthetic example. Then we apply it to two "real world" problems: hierarchical text categorization in the domain of information retrieval (section 5.4.3) and feature extraction for face recognition (section 5.4.4). Finally its application to auditory neural coding is described in section 5.4.5.

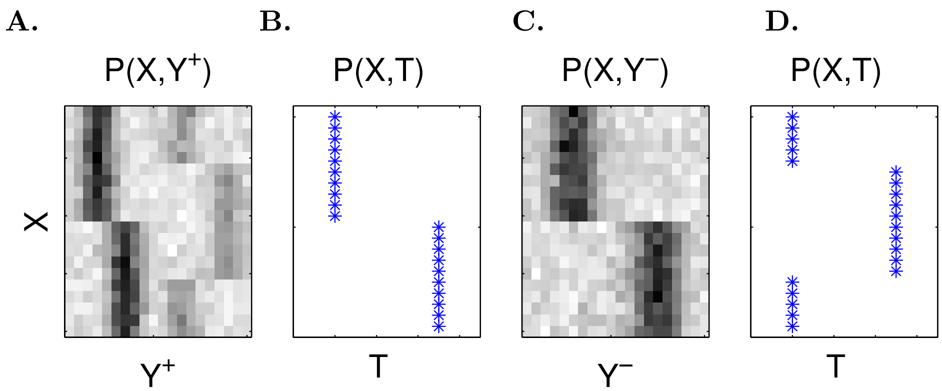

Figure 5.8: Demonstration of IBSI operation. **A.** A joint distribution $P(X, Y^+)$ that contains two distinct and conflicting structures. **B.** Clustering X into two clusters using the information bottleneck method separates upper and lower values of X, according to the stronger structure. **C.** A joint distribution $P(X, Y^-)$ that contains a single structure, similar in nature to the stronger structure $P(X, Y^+)$. **D.** Clustering X into two clusters using IBSI successfully extracts the weaker structure in $P(X, Y^+)$.

5.4.1 A synthetic illustrative example

To demonstrate the ability of our approach to uncover weak but interesting hidden structures in data, we designed a co-occurrences matrix contains two competing substructures (see figure 5.8A). For demonstration purposes, the matrix was created such that the stronger structure can be observed on the left and the weaker structure on the right. Compressing X into two clusters while preserving information on Y^+ using IB ($\gamma = 0$), yields the clustering of figure 5.8C, in which the first half of the x's are all clustered together. This clustering follows from the strong structure on the left of 5.8A.

We now create a second co-occurrence matrix, to be used for identifying the relevant structure, in which each half of X yield similar distributions $P(y^-|x)$. Applying our discriminative clustering algorithm now successfully ignores the strong but irrelevant structure in $P(Y^+, X)$ and retrieves the weak structure. Importantly, this is done in an unsupervised manner, without explicitly pointing to the irrelevant structure.

5.4.2 Model complexity identification

A fundamental problem in modeling data is to identify model complexity levels that well describe given data. This question, already raised in the 12^{th} century by William of Occam, has large effect on models' accuracy. In chapter 1, we described its consequences for density and mutual information estimation, and explained its consequences in terms of accuracy and reliability of estimation. This issue is exemplified by the well known problem of overfitting models to data, and has prompted numerous attempts to elucidate the relation between training error and generalization error (see e.g. chap 7 in [Hastie *et al.*, 2001]). The most widely known approaches are *Minimal description length* (MDL [Rissanen, 1978] and its related *Bayesian information criterion*, BIC, and *Akaike information criterion*, AIC) and VC theory [Vapnik, 1982, Vapnik, 1995]. These approaches focus on the supervised learning scenario, where the basic tradeoff between a good description of the data and a good generalization can be naturally quantified. In the context of unsupervised learning, I claim that there is not necessarily a single level of complexity that is the "correct" one for a system. For example, many natural data sets have several scales and multi resolutions, as illustrated in figures 5.9.

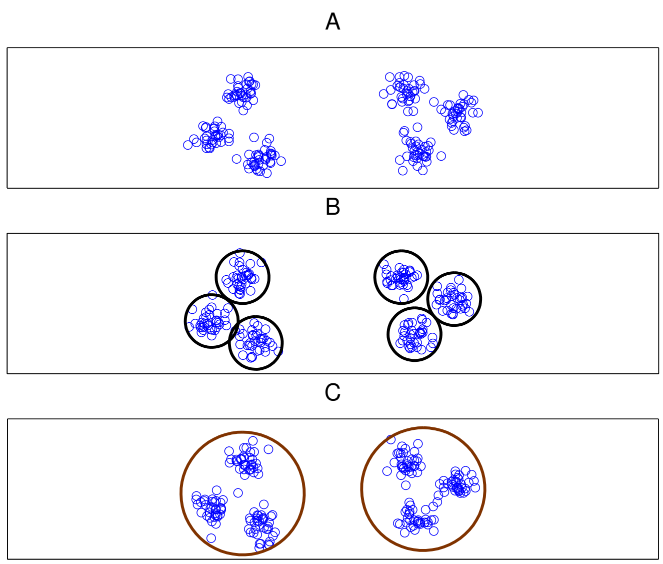

Figure 5.9: Multi resolution in clustering data. Raw data on the upper panel reveals interesting structures both at high resolution (black circles, panel B) and low resolution (brown circles, panel C).

In the context of modeling data with side information, the question of identifying the best resolution of the data, becomes a question of identifying the resolution that provides a good characterization of $P(X, Y^+)$, but not $P(X, Y^-)$. To test whether IBSI can discover such salient resolutions, a synthetic example was created, in which Y^+ contained two relevant resolutions (for 2 and 4 clusters), while Y^- only contained structure at the higher resolution of 4 clusters. The results of running IBSI on this data are demonstrated in figure 5.10

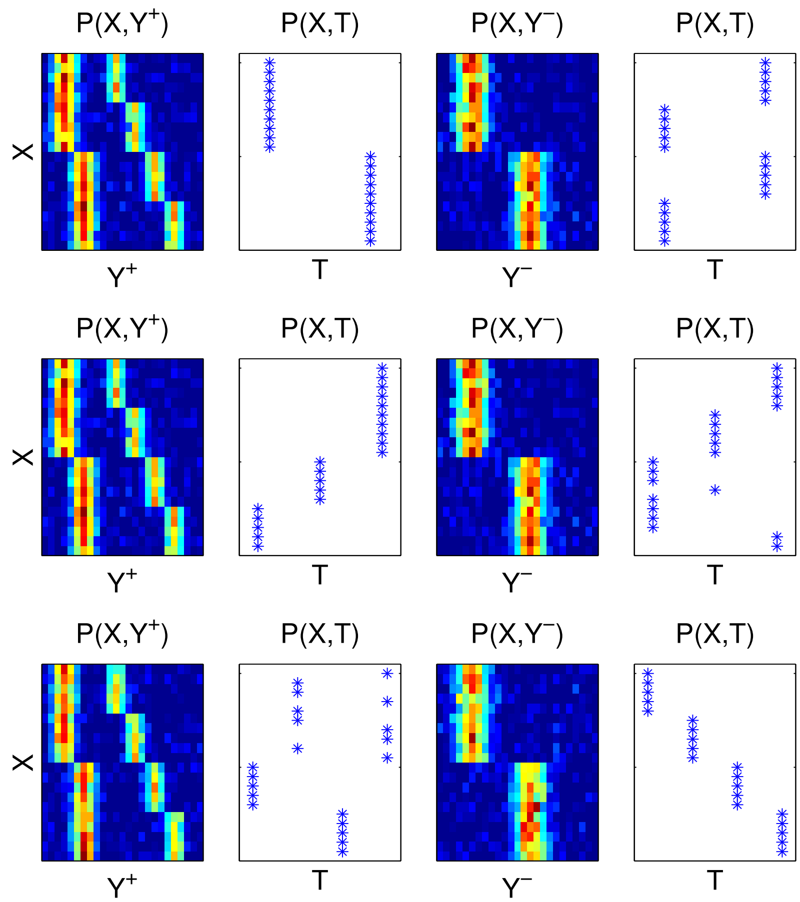

Figure 5.10: Demonstration of IBSI operation, for 2,3 and 4 clusters. Panels as in the previous figure.

Indeed the saliency of each resolution can be measured by the value of the IBSI functional as a function of number of clusters. Figure 5.11 shows that it can identify the relevant resolution in the current data.

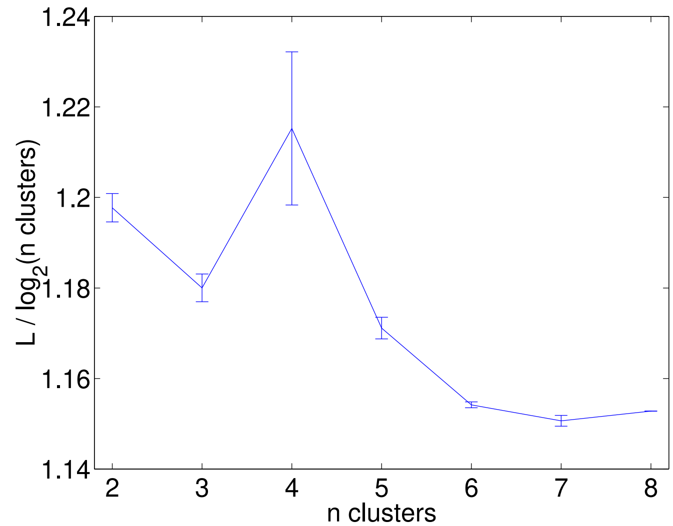

Figure 5.11: Saliency of IBSI operation, for 2 to 8 clusters. Y-axis measures the value of the function $L = I(T;Y^+) - I(T;Y^-)$, normalized by $H(T)$ which is an upper bound on the information. The resolution obtained with 4 clusters is found to best characterize the data.

The last two examples were designed for demonstration purposes, thus the irrelevant structures is strongly manifested in $P(X;Y^-)$. The next examples shows that our approach is also useful for real data, in which structures are much more covert.

5.4.3 Hierarchical text categorization

Text categorization is a fundamental task in information retrieval. Typically, one has to group a large set of texts into sets of homogeneous subjects. Recently, Slonim and colleagues showed that the IB method achieves categorization that predicts manually predefined categories with great accuracy, and largely outperforms competing methods [Slonim *et al.*, 2002]. Clearly, this unsupervised task becomes more difficult when the texts have similar subjects, because alternative categories are extracted instead of the "correct" one.

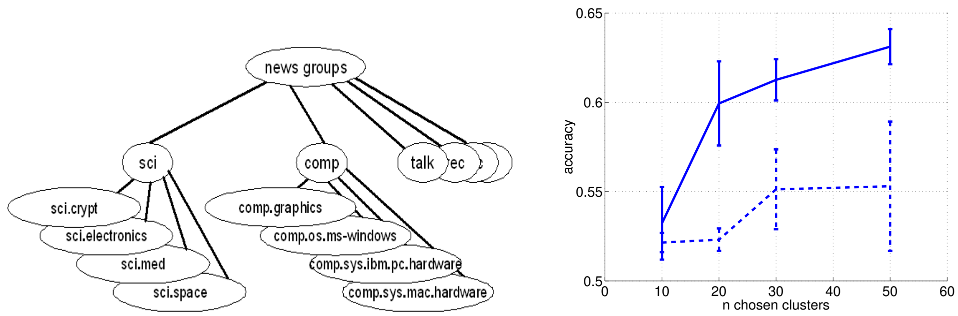

Figure 5.12: **A.** An illustration of the 20 newsgroups hierarchical data we used. **B.** Categorization accuracy vs. number of word clusters k. $N = 100$.

This problem can be alleviated by using side information in the form of additional documents from other categories. This is specifically useful in hierarchical document categorization, in which known categories are refined by grouping documents into sub-categories. [Dumais and Chen, 2000, Vinokourov and Girolani, 2002]. IBSI can be applied to this problem by operating on the terms-documents co-occurrence matrix while using the other top-level groups to focus on the relevant structures. To this end, IBSI is used to identify clusters of terms that will be used later to cluster a group of documents into its subgroups,.

While IBSI is targeted at learning structures in an unsupervised manner, we have chosen to apply it to a labelled dataset of documents to asses its results compared with manual classification. Labels are not used by our algorithms during learning and serve only to quantify performance. We used the *20 Newsgroups database* collected by [Lang, 1995] preprocessed as described in [Slonim *et al.*, 2002]. This database consists of 20 equal sized groups of documents, hierarchically organized into groups according to their content (figure 5.12). We aimed at clustering documents that belong to two newsgroups from the super-group of computer documents and have very similar subjects *comp.sys.ibm.pc.hardware* and *comp.sys.mac.hardware*. As side information we used all documents from the super-group of science (*sci.crypt*, *sci.electronics*, *sci.med*, *sci.space*).

To demonstrate the power of IBSI we used double clustering to separate documents into two groups. The goal of the first clustering phase is to use IBSI to identify clusters of terms that extract the relevant structures of the data. The goal of the second clustering phase is simply to provide a quantitative measure for the quality of the features extracted in the first phase. We therefore performed the following procedure. First, the 2000 most frequent words in these documents were clustered into N clusters using IBSI. Then, word clusters were sorted by a single-cluster score

$D_{KL}[p(y^+|t)||p(y^+)] - \gamma D_{KL}[p(y^-|t)||p(y^-)]$, and the k clusters with the highest score were chosen. These word-clusters were then used for clustering documents. The performance of this process is evaluated by measuring the overlap of the resulting clusters with the manually classified groups. Figure 5.12 plots document-clustering accuracy for $N = 100$, as a function of k. IBSI ($\gamma = 1$) is compared with the IB method (i.e. $\gamma = 0$). Using IBSI successfully improves mean clustering accuracy from about 55 to 63 percents.

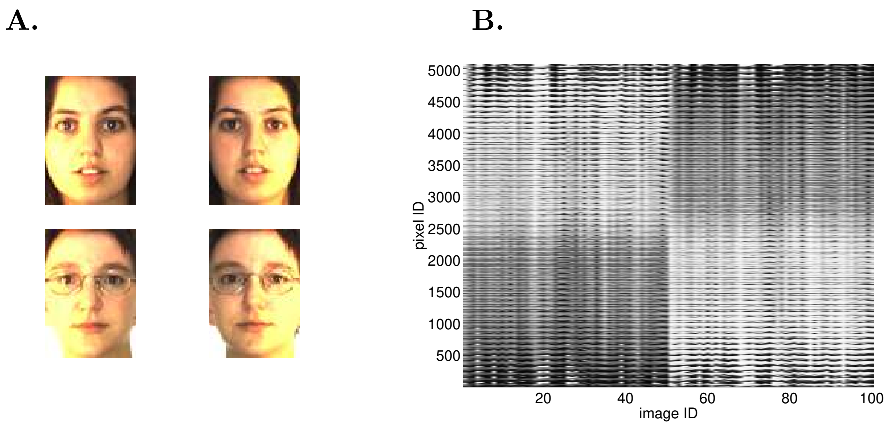

Figure 5.13: **A.** Samples of four face images, each woman face with two different light sources. **B.** Matrix of joint distribution $P(X, Y^+)$ used for IBSI. Each column of the matrix corresponds to a different image in the bank. The 50 left columns correspond to images with a left source light.

5.4.4 Face images

To further demonstrate the applicability of IBSI to diverse types of real-world data, we applied it to the problem of extracting features of face images. We used the AR database of faces [Martinez and Benavente, 1998], and focused on images that contained a strong light source either from the right or the left. These illumination conditions imply strong statistical structures in the set of images. The four examples of images we used are shown in figure 5.13A. Figure 5.13B shows the matrix of all the face images $P(X; Y^+)$, where each column vector of the matrix corresponds to a different picture in the database. The effects of light source are apparent: The columns on the left half of the matrix (corresponding to images with a left light source) appear similar. Moreover, the rows (corresponding to pixels) in the upper part of the matrix are similar (and so are the rows in its lower part). This suggests that clustering pixels according to their distribution across images is expected to reflect the effects of light source.

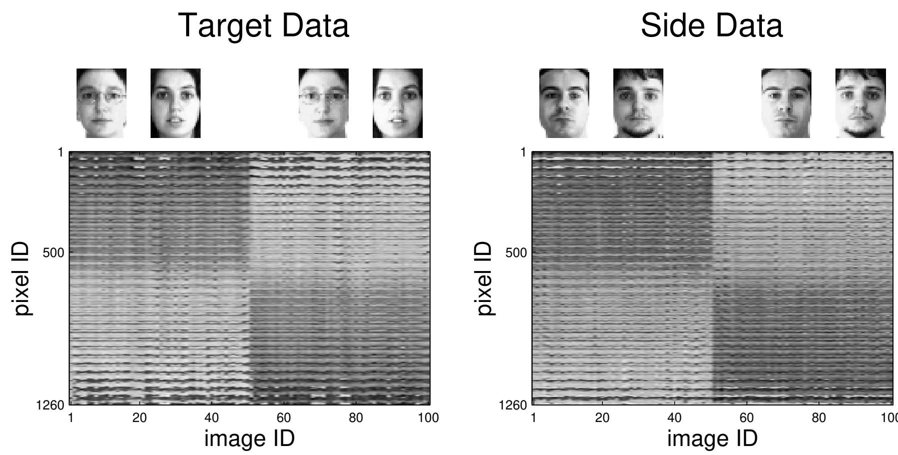

Figure 5.14: Face images of women (main data, left panel) and men (side data, right panel). Each image was reshaped to a column vector and all vectors were concatenated to a matrix with n_{images} columns and n_{pixels} rows.

We applied IBSI to this problem using an additional set of men's faces as side information data $P(X;Y^-)$. This choice of side data actually makes it harder for our method to extract relevant features because of the numerous structures common to men's and women's faces, and images of other illuminated objects would have been more appropriate to the task. Figure 5.15 depicts the results obtained when clustering pixels into 4 clusters, for varying levels of the parameter γ. On each panel clusters appear at different colors, where the colors code the level of differential information $I(T;Y^+) - \gamma I(T;Y^-)$ for each cluster (see the color-bar on the right).

When side data are ignored ($\gamma = 0$), clustering pixels extract a structure that reflects illumination conditions. However, for positive γ values, the effects of light direction is diminished and more salient features are enhanced such as the eyes and the area around the mouth.

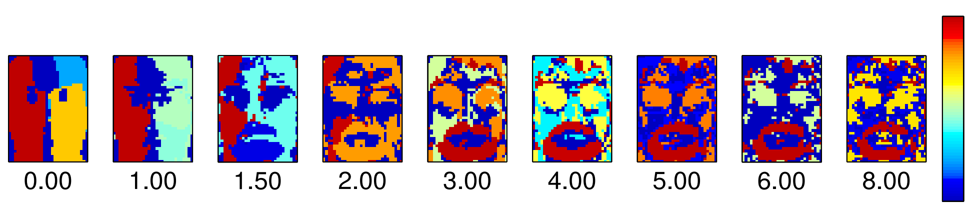

Figure 5.15: Clusters obtained with IBSI for various values of γ. In each panel, four clusters are depicted in different colors, where the colors code the level of differential information $I(T;Y^+) - \gamma I(T;Y^-)$ for each cluster. Similar results are obtained with different number of clusters

5.4.5 Auditory coding

IBSI was used to characterize the processing performed by cortical neurons while filtering out the processing that takes place in lower processing stages.

We calculated the empirical joint distributions of 15 stimuli and neural responses (spike counts) for all cortical neurons. For each cortical neuron we also calculated the same joint distributions for an ANF model neuron that has the same BF as the cortical neuron. We then applied IBSI using positive information terms for all cortical cells and negative ones for the ANF model cells, using an equal $\gamma = 1$ parameter for all irrelevant data and $\beta^{-1} = 0$. This yields a target function in the form of Eq. 5.18

$$\begin{aligned} \mathcal{L} &= -\left[I(T;Y_1^{AI}) + \ldots + I(T;Y_{45}^{AI})\right] \\ &\quad +\gamma\left[I(T;Y_1^{ANF}) + \ldots + I(T;Y_{45}^{ANF})\right] \end{aligned} \tag{5.25}$$

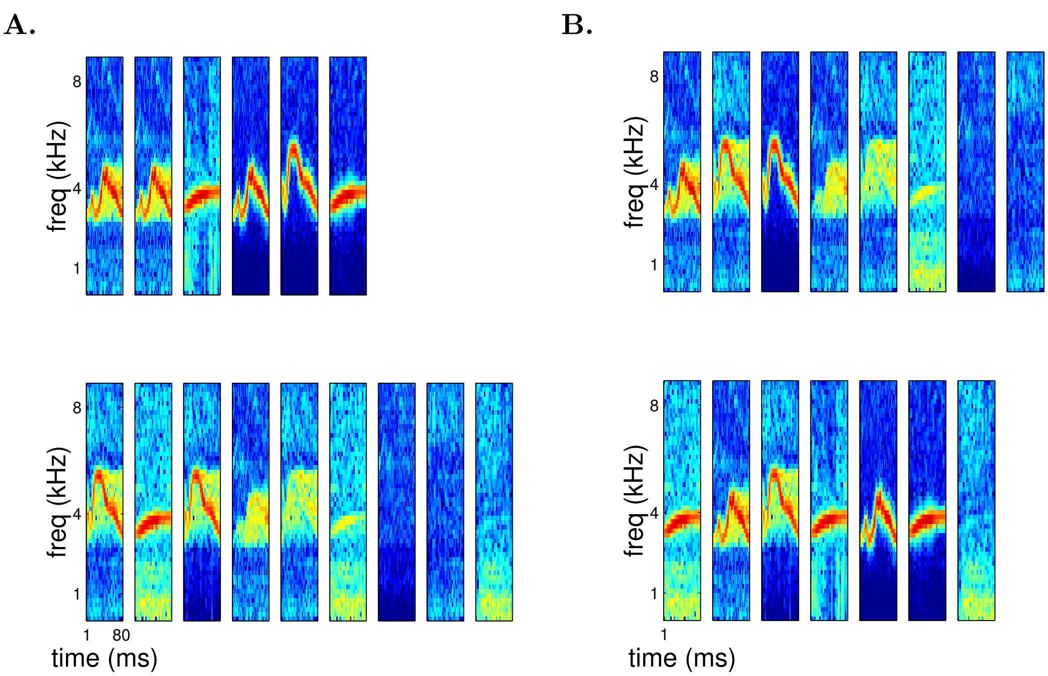

Figure 5.16: Clusters of stimuli obtained with IBSI with 45 AI cells as relevant variables and 45 ANF cells as irrelevant ones. **A.** $\gamma = 0$ **B.** $\gamma = 1$.

An example of the results for 2 clusters is given in figure 5.16. Adding the irrelevant data affects stimuli clusters in the following way. While with $\gamma = 0$ the stimuli that have low energy (noise and background) are all clustered together (see the second cluster in A), this is no longer true when ANF cells are used as irrelevant variables Fig. 5.16B. Unfortunately with this data, it is difficult to determine which are the common relevant features in the stimuli that are clustered together. The major limitations of the current data are two folds: First, the total number of stimuli is small, which makes it more difficult to perform a stability analysis or estimate clusters significance. More importantly, it is highly complex and of high dimensionality. Therefore, even if we know that several stimuli are considered similar by some neurons, it does not provide enough hints about the features to which these neurons respond. In this sense, the IBSI framework is expected to be more beneficial in settings where a centroid of a cluster can be defined in a meaningful way, for example when stimuli belong to some parametric family. Therefore we purposely refrain from drawing conclusions about auditory coding from this experiment, and use it only to demonstrate the potential of IBSI to neural coding investigation.

5.5 Extending the use of side information

This chapter has described how non interesting data can be used to unlearn the structure of the noise in an unsupervised task that uses the information bottleneck framework. While the method described here relies on clustering as the method of dimensionality reduction, the use of side information in the form of irrelevance variables is not limited to clustering but can be extended to various dimensionality reduction methods. This can be formalized by again considering the target function

$$\mathcal{L}(T) = I(X;T) - \beta\left(I(T;Y^+) - \gamma I(T;Y^+)\right) \tag{5.26}$$

which is to be maximized over all the (possibly stochastic) mappings $X \to T$.

We have extended the IB framework to handle continuous variables, and described its complete formal solution for the case of Gaussian variables [Chechik *et al.*, 2003]. For Gaussian-IBSI the problem reduces to one of finding generalized eigenvector of the covariance matrices $\Sigma_{X|Y^+}$ and $\Sigma_{X|Y^-}$ [Chechik and Globerson, 2003]. This result shows the connection between Gaussian-IB and Gaussian IBSI to the well studied problems of *canonical correlations* [Thompson, 1984, Borga, 2001] and the lesser studied *generalized canonical correlation*.

Using side information in the form of irrelevant data, can be further generalized to use information measures other than $I(X;Y)$. Consider the following formalism suggested by [Globerson, 2003]. Let $\vec{f}(x_1,\ldots,x_n)$ be a reduced representation of the data $X = (x_1, ..., x_n)$, and let $Q(f(X), Y)$ be some quality function that measures how well the dimensionality reduction f preserves the structures in the joint relation

X, Y. The goal of relevant feature extraction is then to maximize

$$Q(f(X), Y^+) - \gamma\, Q(f(X), Y^+) \ . \tag{5.27}$$

When f is a clustering of x, $f(X) \equiv T$ and $Q(f(X), Y)$ is the mutual information $I(T;Y)$ this formalism reduces to IBSI with infinite β. However this formalism allows us to extend the side information idea to other settings by choosing other quality functions and features. One such extension was presented in [Globerson *et al.*, 2003], extending an approach presented in [Globerson and Tishby, 2003] and named *sufficient dimensionality reduction*. In this approach continuous statistics are extracted which can be thought of as a set of weighting functions. The quality of these statistics is quantified using the notion of *information in measurement*, which replaces the Shannon mutual information used in IBSI.

5.6 Summary

We presented an information theoretic approach for extracting relevant structures from data, by utilizing additional data known to share irrelevant structures with the relevant data. Naturally, the choice of side data may considerably influence the solutions obtained with IBSI, simply because using different irrelevant variables is equivalent to asking different questions about the data analyzed. In practice, side data can be naturally defined in numerous applications, in particular in exploratory analysis of scientific experiments. For example, it is most suitable to investigating changes in the neural code following learning, or coding differences between brain regions both in terms of the difference between the stimuli that are coded and the code words that are used. It is expected that the use of irrelevance data as side information will be extended to additional dimensionality reduction methods and neural coding applications.

Bibliography

[Baker and McCallum, 1998] L.D. Baker and A. K. McCallum. Distributional clustering of words for text classification. In *Proceedings of the 21th Conference on Research and Development in Information Retrieval (SIGIR-98)*, New York, 1998. ACM press.

[Blahut, 1972] R.E. Blahut. Computation of channel capacity and rate-distortion function. *IEEE Transactions on Information Theory*, 18(4):460–473, July 1972.

[Borga, 2001] M. Borga. Canonical correlation: A tutorial. On line tutorial http://people.imt.liu.se/magnus/cca, January 2001.

[Chechik and Globerson, 2003] G. Chechik and A. Globerson. Information bottleneck and linear projections of Gaussian processes. Technical Report 4, Leibniz Center for Research, School of Engineering and Computer Science, The Hebrew University of Jerusalem, May 2003.

[Chechik *et al.*, 2003] G. Chechik, A. Globerson, Naftali Tishby, and Yair Weiss. Information bottleneck for Gaussian variables. Submitted, 2003.

[Cover and Thomas, 1991] T.M. Cover and J.A. Thomas. *The elements of information theory.* Plenum Press, New York, 1991.

[Csiszar and J.Korner, 1997] I. Csiszar and J.Korner. *Information theory: Coding Theorems for Discrete Memoryless Systems.* Academic Press, New York, 2nd edition, 1997.

[Csiszar, 1974] I. Csiszar. On the computation of rate-distortion functions. *IEEE Transactions on Information Theory*, 20(1):122–124, January 1974.

[Dempster *et al.*, 1977] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.

[Dimitrov and Miller, 2001] A.G. Dimitrov and J.P. Miller. Neural coding and decoding: Communication channels and quantization. *Network: Computation in Neural Systems*, 12(4):441–472, 2001.

[Dumais and Chen, 2000] S. Dumais and H. Chen. Hierarchical classification of web content. In N.J. Belkin, P. Ingwersen, and M-K. Leong, editors, *Proceedings of the 23th Conference on Research and Development in Information Retrieval (SIGIR-00)*, pages 256–263, New York, 2000. ACM press.

[Friedman *et al.*, 2001] N. Friedman, O. Mosenzon, N. Slonim, and N. Tishby. Multivariate information bottleneck. In J.S. Breese and D. Koller, editors, *Uncertainty in Artificial Intelligence: Proceedings of the Seventeenth Conference (UAI-2001)*, pages 152–161, San Francisco, CA, 2001. Morgan Kaufmann.

[Gedeon *et al.*, 2002] T. Gedeon, A.E. Parker, and A.G. Dimitrov. Information distortion and neural coding. *Canadian Applied mathematics Quarterly*, 2002.

[Gilad-Bachrach *et al.*, 2003] R. Gilad-Bachrach, A. Navot, and N. Tishby. An information theoretic tradeoff between complexity and accuracy. In *Proceedings of the Conference on Learning Theory*, Washington, 2003.

[Globerson and Tishby, 2003] A. Globerson and N. Tishby. Sufficient dimensionality reduction. *Journal of Machine Learning Research*, 3:1307–1331, 2003.

[Globerson *et al.*, 2003] A. Globerson, G. Chechik, and N. Tishby. Sufficient dimensionality reduction with irrelevant statistics. In *Uncertainty in Artificial Intelligence: Proceedings of the Nineteenth Conference (UAI-2003)*, San Francisco, CA, 2003. Morgan Kaufmann.

[Globerson, 2003] A. Globerson. Sufficient dimensionality reduction with irrelevant statistics. A talk at the computer science department, the Hebrew University, 2003.

[Hastie *et al.*, 2001] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer series in statistics. Springer, 2001.

[Hoffman, 1999] T. Hoffman. Probabilistic latent semantic indexing. In *Proceedings of the 21th Conference on Research and Development in Information Retrieval (SIGIR-98)*, pages 50–57, New York, 1999. ACM press.

[Lang, 1995] K. Lang. NewsWeeder: learning to filter netnews. In *Proceedings of the 12th International Conference on Machine Learning*, pages 331–339, San Mateo, CA, 1995. Morgan Kaufmann.

[Levy *et al.*, 2001] I. Levy, U. Hasson, G. Avidan, T. Hendler, and R. Malach. Center-periphery organization of human object areas. *Nature Neuroscience*, 4:533–539, 2001.

[Martinez and Benavente, 1998] A.M. Martinez and R. Benavente. The AR face database. Technical report, CVC Technical Report, June 1998.

[Pereira *et al.*, 1993] F.C. Pereira, N. Tishby, and L. Lee. Distributional clustering of English words. In *31st Annual Meeting of the Association for Computational Linguistics*, pages 183–190, Columbus, Ohio, 1993. Morgan Kaufman.

[Rissanen, 1978] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.

[Schneidman *et al.*, 2002] E. Schneidman, N. Slonim, N. Tishby, R. deRuyter van Steveninck, and W. Bialek. Analyzing neural codes using the information bottleneck method. Technical report, Leibniz Center for Research, School of Engineering and Computer Science, The Hebrew University of Jerusalem, 2002.

[Sinkkonen and Kaski, 2001] J. Sinkkonen and S. Kaski. Clustering based on conditional distribution in an auxiliary space. *Neural Computation*, 14:217–239, 2001.

[Slonim and Tishby, 1999] N. Slonim and N. Tishby. Agglomerative information bottleneck. In S. Solla, D.A. Cohn, and T. Lynn, editors, *Advances in Neural Information Processing Systems*, volume 12, Cambridge, MA, 1999. MIT Press.

[Slonim and Tishby, 2000] N. Slonim and N. Tishby. Document clustering using word clusters via the information bottleneck method. In N.J. Belkin, P. Ingwersen, and M-K. Leong, editors, *Proceedings of the 23th Conference on Research and Development in Information Retrieval (SIGIR-00)*, pages 208–215, New York, 2000. ACM press.

[Slonim *et al.*, 2001] N. Slonim, N. Friedman, and N. Tishby. Agglomerative multivariate information bottleneck. In *Advances in Neural Information Processing Systems*, Cambridge, MA, 2001. MIT Press.

[Slonim *et al.*, 2002] N. Slonim, N. Friedman, and N. Tishby. Unsupervised document classification using sequential information maximization. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR-02)*, New York, 2002. ACM Press.

[Thompson, 1984] B. Thompson. *Canonical correlation analysis: Uses and interpretation.*, volume 47. Sage publications, Thousands Oak, CA, 1984.

[Tishby *et al.*, 1999] N. Tishby, F.C. Pereira, and W. Bialek. The information bottleneck method. In *Proceedings of 37th Allerton Conference on communication, control and computation*, Allerton House, Illinois, 1999.

[Vapnik, 1982] V.N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, Berlin, 1982.

[Vapnik, 1995] V.N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag, Berlin, 1995.

[Vinokourov and Girolani, 2002] A. Vinokourov and M. Girolani. A probabilistic framework for the hierarchic organization and classification of document collections. *Journal of intelligent information systems*, 18(23):153–172, 2002. Special Issue on automated text categorization.

[Wyner, 1975] A. Wyner. On source codding with side information at the decoder. *IEEE Transactions on Information Theory*, 21:294–300, 1975.