Sufficient Dimensionality Reduction - A novel analysis principle

Amir Globerson GAMIR@CS.HUJI.AC.IL
Naftali Tishby TISHBY@CS.HUJI.AC.IL
School of Computer Science and Engineering and
The Interdisciplinary Center for Neural Computation
The Hebrew University, Jerusalem 91904, Israel

Abstract

Data dimensionality reduction of empirical co-occurrence data is one of the most fundamental problems in unsupervised learning and complex data analysis. One principled approach to this problem is to represent the data in low dimension with minimal loss of information contained in the original data. In this paper we present a novel information theoretic principle and algorithm for extracting low dimensional representations, or feature-vectors, that approximately preserve the information in one variable about another. Unlike previous work in this direction, here we do not cluster or quantize the variables, but rather extract continuous feature functions directly from the co-occurrence matrix, using a converging iterative projection algorithm. Our approach is both simpler and more general than clustering techniques and proves effective in generating a very small feature set for document categorization.

1. Introduction

The problem of complex data analysis can be understood as the search for a most compact representation, or model, which *explains* a given set of observations. One explicit view of data *explanation* stems from the notion of sufficient statistics in parameter estimation. Sufficient statistics are functions of samples that capture *all* the information about the parameters of a distribution. Every bit of information that can be extracted from the sample about the parameters is captured by such finite dimensional statistics. One may say that sufficient statistics 'explain', or encode, the *relevant* part of the sample with respect to the parameters. This analogy between sufficient statistics and features in pattern recognition and learning has been noticed before. Here we take this analogy much further and propose a general information theoretic approach for extracting features as approximate sufficient statistics. Exact sufficient statistics exist, in general, only for very special distributions - exponential parametric families(Degroot, 1986). We first extend this narrow sense of sufficiency by considering general joint distributions of two variables. We then relax the requirement of *exact* sufficiency by introducing a variational approximation that enables us to calculate functions that approximately preserve the information, in any given dimension.

Consider a joint probability distribution, $p(x, y)$ over the variables x and y. We say that the d-dimensional vector function $\vec{\psi}(Y) = (\psi_1(y), ..., \psi_d(y))$ is *sufficient* for the variable x if $I(X;Y) = I(X;\vec{\psi}(Y))$, where $I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$ is Shannon's mutual information between X and Y (Cover & Thomas, 1991).[1] In this case the information about x in *a sample* of Y, for a given value of x, is completely captured by the empirical expectation of Y, namely, by $\langle\vec{\psi}(Y)\rangle_{p(y|x)}$. Similarly, one can say that the functions $\vec{\phi}(X) = (\phi_1(x), ..., \phi_d(x))$ are sufficient for Y if $I(X;Y) = I(\vec{\phi}(X);Y)$. In that case the dual problem of the information about y from a sample of X is captured by $\langle\vec{\phi}(X)\rangle_{p(y|x)}$. The interesting question we address in this work is if we can find such *dual* sets of functions, $\vec{\psi}(Y)$ and $\vec{\phi}(X)$, simultaneously.

Such complete dimensionality reduction, that preserves *all* the information in the original joint distribution, is possible in general only for special classes of $p(x, y)$, that are of an exponential form. However, these observations motivate a variational approach that finds such a reduction if one exists, but extracts functions $\vec{\psi}(Y)$ and $\vec{\phi}(X)$ that approximate such a reduction in a well defined information theo-

[1]In the conventional notation of parameter estimation X is a d dimensional parameter vector Θ, and Y is an n dimensional i.i.d sample from $p(y|\Theta)$.

retic sense. The duality of these two function sets is a key component of our approach. We further show that this variational principle has other suggestive interpretations which make it a viable candidate for a general principled dimensionality reduction or feature extraction technique.

1.1 Problem formulation

We assume that we are given a joint distribution of two variables $p(x, y)$. In most practical cases we are in fact given only a finite sample from such a distribution, as a co-occurrence matrix. This empirical joint distribution enables us to estimate expectation values of functions of x and y. We also assume that the marginal distributions, $p(x)$ and $p(y)$, are known or can be well estimated. This latter assumption simplifies our analysis and algorithm, but can be relaxed. We first deal with just one of the two dual problems: finding the set $\vec{\psi}(Y)$. We later show that the resulting algorithm in fact solves the two problems simultaneously, and this apparent asymmetry will be removed.

We look for a set of d functions (or features) of the variable y, denoted by $\psi_i(y)$ for $i = 1 \ldots d$, whose expectations can capture the information in the joint distribution in the following information theoretic sense.

The functions $\psi_i(y)$ should satisfy two complementary requirements. On the one hand their expectation should capture all the information about the values of x, thus no other information on y can be assumed. This means that we should *minimize* the information, subject to the known expected values, $\langle \psi_i(y) \rangle_{p(y|x)}$ for every x, as constraints. On the other hand, the functions $\psi_i(y)$ should be *selected* such that these expectations provide the maximum possible information on the variable X.

Given a candidate set of d functions ψ_i, we denote by $\tilde{p}(x, y)$ the distribution with *minimal* information under the expectation constraints. Namely

$$\tilde{p}(x, y) = \arg \min_{q(x,y) \in \mathcal{P}(\psi_i(y), p)} I[q(x, y)] \tag{1}$$

where $I[q(x, y)] = \sum_{x,y} q(x, y) \log \frac{q(x,y)}{p(x)p(y)}$ and the set $\mathcal{P}(\psi_i(y), p)$ is the set of distributions that satisfy the constraints, defined by

$$\mathcal{P}(\psi_i(y), p) = \left\{ \tilde{p}(x, y) : \begin{array}{cc} \langle \psi_i \rangle_{\tilde{p}(y|x)} = \langle \psi_i \rangle_{p(y|x)} & \forall x \\ \tilde{p}(x) = p(x) & \\ \tilde{p}(y) = p(y) & \end{array} \right\} . \tag{2}$$

The second requirement should select the best possible candidate functions ψ_i, by *maximizing* this information over all possible d functions ψ.

Together this can be written as a single Max-Min problem for the optimal functions $\psi_i^*(y)$,

$$\psi_i^*(y) = \arg \max_{\psi_i(y)} \min_{\tilde{p}(x,y) \in \mathcal{P}(\psi_i(y), p)} I(\tilde{p}(x, y)) . \tag{3}$$

Notice that this variational principle *does not* define a generative statistical model for the data and is in fact a model independent approach. As we show later, however, the resulting distribution $\tilde{p}(x, y)$ can be interpreted as a generative model in a class of exponential form. But there is no need to make any assumption about the validity of such a model for the data. The data distribution $p(x, y)$ is needed only to estimate the expectations $\langle \vec{\psi}(y) \rangle$ for every x, given the candidate features.

In what follows we present an iterative projection algorithm, which finds a local solution of this variational problem, and prove its convergence using information geometrical tools. We also show that this asymmetric variational problem for finding the functions $\psi_i(y)$ is in fact equivalent to the dual problem of finding the functions $\phi_i(x)$ and in fact the two problems are solved simultaneously. Our iterative algorithm uses any MaxEnt algorithm, such as "iterative scaling" (Darroch & Ratcliff, 1972), as its inner component. Using it we perform alternating projections between two (infinite) sets of linearly constrained distributions.

The problem as formulated provides a tool for data dimensionality reduction, and as such is applicable for a wide range of problems, from natural language processing to neural code analysis and bioinformatics. An illustrative document classification application will be presented. As we discuss further, this variational problem is similar in structure to the fundamental problem of the capacity of an unknown channel which suggests other interesting interpretations for our procedure. It is also related to the recently proposed *Information Bottleneck Method* (Tishby et al., 1999) as well as to other dimensionality reduction algorithms, such as LSA(Deerwester et al., 1990), LLE(Roweis & Saul, 2000) and non-negative matrix factorization (Lee & Seung, 1999).

2. The nature of the solution

We first show that the problem as formulated in Eq. 3 is equivalent to the problem of minimizing the KL divergence between the empirical distribution $p(x, y)$ and a large (parametric) family of distributions of an exponential form.

To simplify notation, we sometimes omit the suffix of (x, y) from the distributions. Thus $\tilde{p}_t$ stands for

$\tilde{p}_t(x,y)$ and p for $p(x,y)$

Minimizing the mutual information in Eq. 1 under the linear constraints on the expectations $\langle\vec{\psi}(y)\rangle$ is equivalent to maximizing the joint entropy, $H[\tilde{p}(x,y)] = -\sum_{x,y}\tilde{p}(x,y)\log\tilde{p}(x,y)$, under these constraints, with the additional requirement on the marginals, $\tilde{p}(x) = p(x)$ and $\tilde{p}(y) = p(y)$. Due to the concavity of the entropy and the convexity of the linear constraints, there exists a unique maximum entropy distribution (for compact domains of x and y) which has the exponential form,

$$\tilde{p}^*_{\vec{\psi}(y)}(x,y) = \frac{1}{Z}\exp\left(\sum_{i=1}^{d}\phi_i(x)\psi_i(y) + A(x) + B(y)\right) , \tag{4}$$

where
$Z = \sum_{x,y}\exp\left(\sum_{i=1}^{d}\phi_i(x)\psi_i(y) + A(x) + B(y)\right)$ is the normalization (partition) function [2].

The functions $\phi_i(x)$ are uniquely determined as Lagrange multipliers from the expectation values $\langle\psi_i(y)\rangle$. This exponential form is also directly linked to our interpretation of the functions ψ_i and ϕ_i as conjugate sufficient statistics.

The set of distributions of the exponential form in Eq. 4 can also be considered as a parametric family, parametrized by the infinite family of functions $\Theta = [\psi_i(y), \phi_i(x), A(x), B(y)]$ (note we treat ψ and ϕ symmetrically). We denote this family by P_Θ and a member of the family by p_Θ. Naturally, $\tilde{p}^*_{\vec{\psi}(y)} \in P_\Theta$.

We define the set of distributions $\mathcal{P}_\Psi \subset P_\Theta$:

$$\mathcal{P}_\Psi = \left\{\tilde{p} \in P_\Theta : \begin{array}{cc} \langle\psi_i\rangle_{\tilde{p}(y|x)} = \langle\psi_i\rangle_{p(y|x)} & \forall x \\ \tilde{p}(x) = p(x) & \\ \tilde{p}(y) = p(y) & \end{array}\right\} . \tag{5}$$

The problem presented above (Eq. 3) is then equivalent to finding the information maximizing distribution in $\mathcal{P}_\Psi$,

$$\tilde{p}^* = \arg\max_{\tilde{p}\in\mathcal{P}_\Psi} I[\tilde{p}] . \tag{6}$$

We now return to the information maximization problem, i.e. the selection of the feature candidates $\vec{\psi}(y)$. For every $\tilde{p} \in \mathcal{P}_\Psi$ one can easily show that

$$I[\tilde{p}] = I[p] - D_{KL}[p|\tilde{p}] , \tag{7}$$

where $D_{KL}[p|q] = \sum_{x,y} p\log\frac{p}{q}$. The above expression has two important consequences. The first is that maximizing $I[\tilde{p}]$ for $\tilde{p} \in \mathcal{P}_\Psi$ is equivalent to minimizing $D_{KL}[p|\tilde{p}]$ for $\tilde{p} \in \mathcal{P}_\Psi$:

$$\tilde{p}^* = \arg\min_{\tilde{p}\in\mathcal{P}_\Psi} D_{KL}[p|\tilde{p}] . \tag{8}$$

In addition, Eq. 7 shows that the information in $I[\tilde{p}^*]$ can not be larger than the information in the original data. This supports the intuition that the model $\tilde{p}^*$ maintains only properties present in the original distribution that are captured by the selected features $\psi_i^*(y)$, for any value of x.

The problem in Eq. 8 is a maximization of a function over a subset of P_Θ, namely $\mathcal{P}_\Psi$. The following proposition shows that this is in fact equivalent to maximizing the same function over all of P_Θ.

Proposition 1
$\arg\min_{\tilde{p}\in\mathcal{P}_\Psi} D_{\mathrm{KL}}[p|\tilde{p}] = \arg\min_{\tilde{p}\in P_\Theta} D_{\mathrm{KL}}[p|\tilde{p}]$

Proof: We need to show that the distribution which minimizes the right hand side is in $\mathcal{P}_\Psi$. Indeed, by taking the (generally functional) derivative of $D_{KL}[p(x,y)|\tilde{p}]$ w.r.t. the parameters Θ in $\tilde{p}$, one obtains the following conditions:

$$\begin{array}{ll} \forall x, i & \langle\psi_i(y)\rangle_{\tilde{p}(y|x)} = \langle\psi_i(y)\rangle_{p(y|x)} \\ \forall y, i & \langle\phi_i(x)\rangle_{\tilde{p}(x|y)} = \langle\phi_i(x)\rangle_{p(x|y)} \\ \forall x & \tilde{p}(x) = p(x) \\ \forall y & \tilde{p}(y) = p(y) . \end{array} \tag{9}$$

Clearly, this distribution satisfies the constraints in $\mathcal{P}(\psi_i(y), p)$ and is therefore in $\mathcal{P}_\Psi$. □

Our problem is thus equivalent to the minimization problem:

$$p^* = \arg\min_{\tilde{p}\in P_\Theta} D_{KL}[p|\tilde{p}] . \tag{10}$$

Equation 10 is symmetric with respect to ϕ and ψ, thus removing the asymmetry between X and Y in the formulation of Eq. 3.

Notice that this minimization problem can be viewed as a Maximum Likelihood fit to the given $p(x,y)$ in the class P_Θ, but since nothing on its own guarantees the quality of this fit, nor justifies the class P_Θ, we prefer the information theoretic, model independent, interpretation of our approach.

3. An Iterative Projection Algorithm

Our main result is an iterative algorithm for solving the Max-Min variational problem, Eq. 3, which provably converges to a minimum of Eq. 10. We describe the algorithm using the information geometric notion of *I-projections* (Csiszar, 1975).

[2]Note that the unique distribution can actually be on the closure of such exponential forms. We do not address this detail here.

The *I-projection* of a distribution $q(x)$ on a set of distributions $\mathcal{P}$ is defined as the distribution $p^*(x)$ in $\mathcal{P}$ which minimizes the KL-divergence $D_{KL}[p|q]$.

$$p^*(x) = \arg\min_{p \in \mathcal{P}} D_{KL}[p|q] \ . \tag{11}$$

We shall use the notation $p^*(x) = IPR(q, \mathcal{P})$.

An important property of the *I-projection* is the, so called, Pythagorean property (Cover & Thomas, 1991). For every distribution p in $\mathcal{P}$ the following holds:

$$D_{KL}[p|q] \le D_{KL}[p|p^*] + D_{KL}[p^*|q] \ . \tag{12}$$

We now focus on the case where the set $\mathcal{P}$ is a convex set determined by expectation values. Given a set of d functions $f_i(x)$ and a distribution $p(x)$, we denote the set of distributions which agree with $p(x)$ on the expectation values of $f_i(x)$, by $\mathcal{P}(f_i(x), p(x))$. Namely,

$$\mathcal{P}(f_i(x), p(x)) = \left\{ \tilde{p}(x) : \langle f_i(x) \rangle_{\tilde{p}(x)} = \langle f_i(x) \rangle_{p(x)} \right\} \tag{13}$$

The *I-projection* in this case has the exponential form

$$IPR(q(x), \mathcal{P}(f_i(x), p(x))) = \frac{1}{Z^*} q(x) \exp \sum_i \lambda_i^* f_i(x) \ , \tag{14}$$

where λ_i^* is the Lagrange multiplier corresponding to $\langle f_i(x) \rangle_{p(x)}$. In addition, for this special set, the Pythagorean inequality actually becomes an equality (Csiszar, 1975).

Before describing the algorithm, we need some additional notations:

- $\tilde{p}_t(x, y)$ - the distribution after t iterations.
- $\psi_{i,t}(y)$ - the $\psi_i(y)$ functions for $\tilde{p}_t(x, y)$, Lagrange multipliers for $\langle \phi_{i,t-1}(x) \rangle$.
- $\phi_{i,t+1}(x)$ - the $\phi_i(x)$ functions for $\tilde{p}_{t+1}(x, y)$, Lagrange multipliers for $\langle \psi_{i,t}(y) \rangle$.
- Θ_t - the full parameter set for $\tilde{p}_t(x, y)$.

The iterative projection algorithm is outlined in figure 1. In this figure the algorithm is described through iterated *I-projections* of (exponential) distributions, once for fixed $\psi_i(y)$ and their expectations, and then for fixed $\phi_i(x)$ and their expectations. Interestingly, during the first projection, the functions $\phi_i(x)$ are modified as Lagrange multipliers for $\langle \psi_i(y) \rangle$, and vice-versa in the second projection. The iteration can thus be viewed as alternating mapping between the two sets of d-dimensional functions, ψ_i and ϕ_i. This is also the direct goal of the variational problem.

Input: Joint (empirical) distribution $p(x, y)$

Output: $2d$ feature functions: $\psi_i(y)$ $\phi_i(x)$ that result from $\tilde{p}^*$, a solution of the variational problem Eq. 3 and a (local) minimum of Eq. 8.

Initialization:

- Initialize $\tilde{p}_0(x, y) \in P_\Theta$ randomly

Iterate:

- For all x, set:

  $$\begin{aligned} \tilde{p}_{t+1}(y|x) &= IPR(\tilde{p}_t(y|x), \mathcal{P}(\psi_{i,t}(y), p(y|x))) \\ \tilde{p}_{t+1}(x, y) &= \tilde{p}_{t+1}(y|x) p(x). \end{aligned}$$

  The functions $\phi_{i,t+1}(x)$ are determined as the Lagrange multipliers

- For all y, set:

  $$\begin{aligned} \tilde{p}_{t+2}(x|y) &= IPR(\tilde{p}_{t+1}(x|y), \mathcal{P}(\phi_{i,t+1}(x), p(x|y))) \\ \tilde{p}_{t+2}(x, y) &= \tilde{p}_{t+2}(x|y) p(y). \end{aligned}$$

  The functions $\psi_{i,t+1}(y)$ are determined as Lagrange multipliers

- Halt on convergence (when small enough change in $\tilde{p}_t(x, y)$).

Figure 1. The iterative projection algorithm.

We proceed to prove the convergence of the algorithm. We first show that every step reduces $D_{KL}[p(x, y)|\tilde{p}_t(x, y)]$.

Proposition 2
$D_{KL}(p(x, y), \tilde{p}_{t+1}(x, y)) \le D_{KL}(p(x, y), \tilde{p}_t(x, y))$.

Proof: For each x, the following is true:

- $\tilde{p}_{t+1}(y|x)$ is the *I-projection* of $\tilde{p}_t(y|x)$ on the set $\mathcal{P}(\psi_{i,t}(y), p(y|x))$.
- $p(y|x)$ is also in $\mathcal{P}(\psi_{i,t}(y), p(y|x))$.

Using the Pythagorean property, (which is an equality here) we have that $D_{KL}[p(y|x)|\tilde{p}_t(y|x)]$ is equal to:

$$D_{KL}[p(y|x)|\tilde{p}_{t+1}(y|x)] + D_{KL}[\tilde{p}_{t+1}(y|x)|\tilde{p}_t(y|x)] \ . \tag{15}$$

Multiplying by $p(x)$ and summing over all x values, we obtain:

$$D_{KL}[p|\tilde{p}_t] = D_{KL}[p|\tilde{p}_{t+1}] + D_{KL}[\tilde{p}_{t+1}|\tilde{p}_t] \ . \quad (16)$$

where we have used $\tilde{p}_t(x) = p(x)$. Note that the summation resulted in the term $D_{KL}[p(x)|\tilde{p}_{t+1}(x)]$ on both sides of the equation.

Using the non-negativity of $D_{KL}[\tilde{p}_{t+1}|\tilde{p}_t]$ we have:

$$D_{KL}[p|\tilde{p}_t] \geq D_{KL}[p|\tilde{p}_{t+1}] \ . \quad (17)$$

Note that equality is obtained iff $D_{KL}[\tilde{p}_{t+1}|\tilde{p}_t]$. □

An analogous argument proves that:

$$D_{KL}[p|\tilde{p}_{t+2}] \leq D_{KL}[p|\tilde{p}_{t+1}] \ . \quad (18)$$

The following easily provable proposition states that the stationary points of the algorithm coincide with extremum points of the target function $D_{KL}[p|p_\Theta]$. Its proof uses the properties of the projections in the algorithm, and the characterization of the extremum point in Eq. 9.

Proposition 3 *If $\tilde{p}_t = \tilde{p}_{t+2}$ then the corresponding Θ_t satisfies $\frac{\partial}{\partial\Theta} D_{\mathrm{KL}}[p|p_\Theta] = 0$.*

It is now easy to prove convergence of the algorithm to a local minimum of $D_{KL}[p|p_\Theta]$ using continuity arguments and the improvement at each iteration given in Eq. 16. We will give a detailed proof in a separate paper.

3.1 Implementation - Partial I-projections

The description of the iterative algorithm assumes the existence of a module which calculates I-projections on linear constraints. Because no closed form solution for such a projection is known, it is found by successive iterations which asymptotically converge to the solution. It is straightforward to show that even if our projection algorithm uses such a "Partial I projection" algorithm as its IPR module, it still converges to a minimum.

In this work we use as an IPR algorithm the Generalized Iterative Scaling (GIS) procedure (Darroch & Ratcliff, 1972), described in figure 2. Alternative algorithms such as Improved Iterative Scaling (Pietra & Lafferty, 1997) or conjugate gradient methods, can also be used and may improve convergence rate.

Input: Distributions $q(x), p(x)$, d functions $f_i(x)$

Output: $IPR(q(x), \mathcal{P}(f_i(x), p(x)))$

Initialize: $p_0(x) = q(x)$

Iterate:

- $p_{t+1}(x) = \frac{1}{Z_t} p_t(x) \exp \sum_{i=1}^{d} f_i(x) \log \frac{\langle f_i(x)\rangle_{q(x)}}{\langle f_i(x)\rangle_{p_t(x)}}$
- Z_t is a normalization constant

Figure 2. The Generalized Iterative Scaling Algorithm.

4. Example of application: Document analysis

The analysis of large databases of documents presents a challenge for machine learning and has been addressed by many different methods (see (Yang & Liu, 1999) for a recent review). Important applications in this domain include text categorization, information retrieval and document clustering.

Many works use a probabilistic and information theoretic framework for this analysis (Slonim & Tishby, 2000; Hofmann, 2001), whereby documents and terms are considered stochastic variables. The probability of a word $w \in w_1, \ldots, w_n$ appearing in a document $doc \in doc_1, \ldots, doc_n$ is denoted $p(w|doc)$ and is obtained from the normalized word count vector for this document:

$$p(w|doc) = \frac{n(doc, w)}{\sum_w n(doc, w)} \ , \quad (19)$$

where $n(doc, w)$ is the number of occurrences of word w in document doc. The joint distribution $p(w, doc)$ is then calculated by multiplying by a document prior $p(doc)$ (e.g. uniform). This stochastic relationship is then analyzed using an assumed underlying generative model (e.g. a linear mixture model in (Hofmann, 2001) , or a maximum entropy model in (Berger et al., 1996)). An optimal model is found, and used for predicting properties of unseen documents.

In the current work, we use our feature space analysis for finding word features $\psi_i(w)$ such that using their mean values alone we can infer information about document identity. This approach is conceptually different from linear based approaches such as LSI (Deerwester et al., 1990) or PLSI (Hofmann, 2001) in that it focuses on properties of the feature space, rather than of the probability space.

It is related to, but extends, the work on maximum entropy in NLP (Berger et al., 1996; Pietra & Lafferty, 1997). In these works, the set of features $\psi_i(w)$ was predetermined, or was algorithmically chosen from a large set of predefined features. Our approach does not assume anything about the feature space (e.g positivity, range etc') and performs a completely unsupervised search for the optimal $\psi_i(w)$.

One of the advantages of our approach is that it allows an easy interpretation of the resulting model, which can also be used as an exploratory data analysis tool. The functions $\psi_i(w)$ describe different features (hopefully few) of words which are relevant for deciding upon document identity.

4.1 Using features for indexing and classification

The output of our algorithm is a set of features $\psi_i(w)$ whose expected value provides information about document identity. It is therefor natural to take the expected values $\langle\psi_i(w)\rangle_{p(w|doc)}$ as a d dimensional index vector for document doc . We denote this vector by $\hat{\psi}(doc)$. We have thus reduced the representation of a document from $|W|$ dimensions to d dimensions. This serves two purposes. First, to extract relevant information and eliminate noise. Second, low dimensional vectors can allow faster, more efficient information retrieval, an important feature in real world applications.

Another possible index is the set of Lagrange multipliers $\phi_i(doc)$, denoted by $\hat{\psi}(doc)$, associated with a given document. $\hat{\psi}(doc)$ is obtained by I projecting $p(w|doc)$ on the set of constrains $\mathcal{P}(\psi_i(w), p(w|doc))$. The resulting coefficients determine the relative importance of the $\psi_i(w)$ features for the given document.

The indices $\hat{\psi}(doc)$ and $\hat{\phi}(doc)$ are obtained without using the classes (subjects) of the training documents. In this important sense the algorithm is unsupervised. However, they can still be used for classifying documents into underlying topics, using a supervised learning scheme.

In the experiments described below, we used both $\hat{\psi}(doc)$ and $\hat{\phi}(doc)$ as inputs to a support vector machine classifier (SVM-Light (Joachims, 1998)). Baseline results were obtained by using an SVM classifier trained on the original $p(w|doc)$ vectors.

4.2 Experiments

The performance of our method was tested on the 20Newsgroups database collected by (Lang, 1995).

We start with an illustrative example, which shows how features capture information about document content. We chose two different newsgroups with subjects : "alt.atheism" and "comp.graphics", and preselected 300 terms and 250 documents per subject, using the information gain criterion (Yang & Liu, 1999). The projection algorithm was then run on the resulting $p(w, doc)$ matrix, for values of $d = 1, 2, 4, 8, 16, 32$.

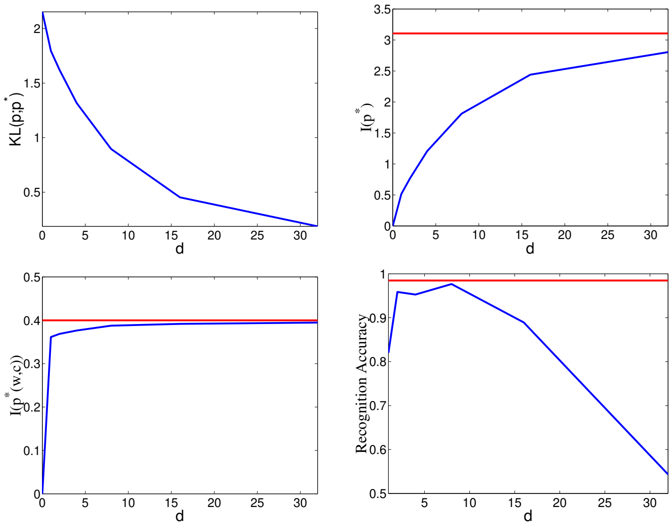

Figure 3. Upper left: KL between the original distribution and the model, for different d values. Upper right: Information in the model. The horizontal line marks the information in the original distribution. Bottom left: Information between document class and words according to p^*. The horizontal line marks this information for the original distribution. Bottom right: Recognition accuracy for different values of d. The horizontal line marks baseline classification accuracy.

Figure 3 shows $D_{KL}[p|p^*]$, $I[p^*]$ as a function of d. It can be seen that as larger values of d are used, the original distribution is approached in the KL sense and in the mutual information sense (as in equation 7).

The magnitude of $I[p^*]$ is not directly correlated with the ability to classify documents into underlying topics. A more relevant measure is the mutual information, $I[p(c, w)]$, between words w and topics c. As shown in figure 3, this measure, calculated for $p^*(c, w)$ approaches $I[p(c, w)]$ already at low values of d. This is indeed reflected in the classification accuracy as seen in the bottom right figure of 3. The drop in recognition accuracy for high values of d is a result of over-fitting.

To gain insight into the nature of the chosen features, we look at $\psi_1(w)$ obtained for the $d = 1$ case. A histogram of the values of $\psi_1(w)$ is shown in figure 4. It can be seen that the values are roughly symmetrical about 0. Figure 5 shows the 5 words with the highest and lowest $\psi_1(w)$ values. Clearly, the words with high $\psi_1(w)$ correspond to the "comp.graphics" subject, and the ones with low $\psi_1(w)$ correspond to "alt.atheism". This single feature thus creates a *ternary* code through which it assigns positive weights to one class, negative weights to the other, and low weights to words which are possibly irrelevant to the classes.

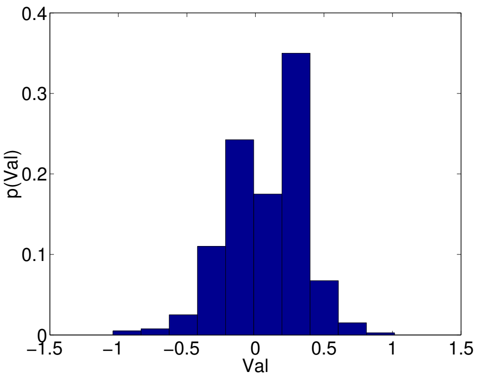

Figure 4. Histogram of $\psi_1(w)$ values.

universe	format
laws	pov
behavior	formats
moral	files
islamic	gif

Figure 5. Words with high (right) and low (left) $\psi_1(w)$ values

We next performed classification on eight different pairs of newsgroups. For each newsgroup, we evaluated classification based on the $\hat{\phi}(doc)$ and $\hat{\psi}(doc)$ indices, and the baseline performance. We experimented with values $d = 1, 2, 4, 8, 16$ to test how well we can classify with relatively few features. For each newsgroup pair, we found the optimal value of d for the two indices, and the corresponding optimal classification accuracy. Our results demonstrate that, on average, both indices give an accuracy that is $95 \pm 3\%$ of the baseline accuracy with an average optimal feature size of $d = 6 \pm 5$ for the $\hat{\phi}(doc)$ index and $d = 9 \pm 6$ for the $\hat{\psi}(doc)$ index.

Thus, classification based on either index achieves performance comparable to that of the baseline, even though it uses significantly less features (compared to the 500 features used by the baseline classifier). Importantly, the features were obtained in a wholly unsupervised manner, and have still discovered the document properties relevant for classification.

5. Discussion

Our proposed method is a new dimensionality reduction technique. It is nonlinear, unlike PCA or ICA, and it aims directly at preserving mutual information in a given empirical co-occurrence matrix. We achieved that through an information variation principle that enables us to calculate simultaneously informative feature functions for *both* random variables. In addition we obtain an exponential model approximation to the given data which has precisely these features as *dual sets* of sufficient statistics. We described an alternating projection algorithm for finding these features and proved its convergence to a (local) optimum. This is in fact an algorithm for extracting optimal sets of constraints from statistical data. We briefly address now several other important issues that our procedure raises.

5.0.1 Finite samples

The basic assumption behind our problem formulation is that we have the joint distribution of the variables x and y, $p(x, y)$. For the machine learning community this assumption may look strange, but one should remember that the goal is *not* to learn the mapping from X to Y, but rather to extract good features of X w.r.t. Y. In fact, $p(x, y)$ is not needed explicitly, but only to estimate the expectation values $\langle \vec{\psi}(y) \rangle$ and $\langle \vec{\phi}(x) \rangle$. These can be estimated *uniformly well* from a finite sample under the standard uniform convergence conditions. In other words, standard learning theoretical techniques can give us the sample complexity bounds, given the dimensions of X and Y and the reduced dimension d. For continuous x and y further assumptions must be made, such as the VC dimension of the features, etc.

5.0.2 Uniqueness of the solution

Interestingly, while the goal of the algorithm are features that preserve information, the information in the functions $\vec{\psi}(y)$ and $\vec{\phi}(x)$ is not estimated directly at any point. Furthermore, there is a freedom in selecting the features that stems from the fact that only the dot-product $\vec{\phi}(x) \cdot \vec{\psi}(y)$ appears in the distribution $\tilde{p}(x, y)$. Any invertible matrix R can be applied such that $\vec{\phi}(x) R^{-1}$ and $R\vec{\psi}(y)$ are equally good features. One can remove this ambiguity by orthogonalization and scaling of the feature functions, for example by applying SVD to $\log \tilde{p}(x, y)$. Notice, however, that our procedure is very different from direct application of SVD to $\log p(x, y)$. These two coincide *only* when the original joint distribution is already of the exponential form of Eq.(4). In all other cases SVD based approximations (LSA included) will not preserve information as well as our features at the same dimension reduction. The resulting functions $\vec{\psi}(y)$ and $\vec{\phi}(x)$ thus depend on the initial point of the iterations, but the information extracted does not (for the same optimum).

5.0.3 Information theoretic interpretation

Our information MaxMin principle is close in its formal structure to the problem of channel capacity with some channel uncertainty (see e.g. (Lapidoth & Narayan, 1998)). This suggests the interesting interpretation for the features as channel characteristics. If the channel only enables the reliable transmission of d expected values, then our $\vec{\psi}(y)$ exploit this channel in an optimal way. The channel decoder of this case is provided by the dual vector $\vec{\phi}(x)$ and the decoding is performed through a dot-product in of these two vectors. This intriguing interpretation of our algorithm obviously requires further analysis.

5.1 Relations to other algorithms

Dimension reduction algorithms have become a fundamental component in unsupervised large scale data analysis, together with clustering. While linear methods, as PCA and ICA, provide very useful first steps, they do not provide a principled method for statistical co-occurrence data, where such linear assumptions about the matrix are unjustified. Several interesting non-linear methods have been proposed in the past years. Of particular interest are LLE(Roweis & Saul, 2000) and non-negative matrix factorization (Lee & Seung, 1999). We feel that none of these new algorithms directly address the question of information preserving as we suggest here. Furthermore, preliminary comparisons of our algorithm with LSA (Deerwester et al., 1990) for document indexing are very encouraging and justify our non-linear approach. A closely related idea is the *Information Bottleneck Method* (Tishby et al., 1999) which aims at a clustering that preserves information. However, clustering may not be the correct answer for many problems where the relationship between the variables comes from some hidden low dimensional continuous structures. In such cases clustering tends to quantize the data in a rather arbitrary way, while low dimensional features are simpler and easier for interpretation. The resulting algorithm is also computationally simpler, with no need for complicated splitting or merging of clusters.

Acknowledgement

We thank Noam Slonim and Gal Chechik for helpful discussions and comments and for the help with experimental data. This work is partly supported by a grant from the Israeli Academy of Science. A.G. is supported by the Eshkol Foundation.

References

Berger, A., Pietra, S. D., & Pietra, V. D. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, *22*, 39–71.

Cover, T., & Thomas, J. (1991). *Elements of information theory*. Wiley.

Csiszar, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, *3*, 146–158.

Darroch, J., & Ratcliff, D. (1972). Generalized iterative scaling for log-linear models. *Ann. Math. Statist.*, *43*, 1470–1480.

Deerwester, S., Dumais, S., Landauer, T., Furnas, G., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, *41*, 391–407.

Degroot, M. (1986). *Probability and statistics*. Addison-Wesley.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, *42*, 177–196.

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. *10th European Conference on Machine Learning*.

Lang, K. (1995). Newsweeder: Learning to filter news. *ICML 95* (pp. 331–339).

Lapidoth, A., & Narayan, P. (1998). Reliable communication under channel uncertainty. *IEEE Transactions on Information Theory*, *44*, 2148–2177.

Lee, D., & Seung, H. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*, 788–791.

Pietra, S. D., & Lafferty, V. D. P. J. (1997). Inducing features of random fields. *IEEE Transactions on PAMI*, *19*, 380–393.

Roweis, S., & Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*, 2323–6.

Slonim, N., & Tishby, N. (2000). Document clustering using word clusters via the information bottleneck method. *ACM SIGIR 2000* (pp. 208–215).

Tishby, N., Pereira, F., & Bialek, W. (1999). The information bottleneck method. *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing* (pp. 368–377).

Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. *ACM SIGIR 1999* (pp. 42–49).