Sufficient Dimensionality Reduction - A novel analysis method

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

Dimensionality reduction of empirical co-occurrence data is a fundamental problem in unsupervised learning. One principled approach to this problem is to represent the data in low dimension with minimal loss of (mutual) information contained in the original data. In this paper we introduce a novel information theoretic nonlinear method for finding such most informative dimension reduction. In contrast with previously introduced clustering based approaches, here we extract *continuous feature functions* directly from the co-occurrence matrix. It is different from dimensionality reduction methods which are based on a specific, sometimes arbitrary, metric or embedding. We present an iterative projection algorithm for finding such features and prove its convergence. The algorithm is demonstrated here on text categorization and proves effective in selecting a small set of features for document categorization, without compromising performance.

1. Introduction

Many problems in machine learning involve structure discovery in co-occurrence data. It is often desirable to find simple features - functions of the variables - such that observation of the feature values captures all their mutual information. This is in fact the underlying idea behind the classical notion of *sufficient statistics* in parameter estimation theory. In this statistical context, the parameter θ is estimated from an i.i.d. sample $X^n = (X_1, \ldots, X_n)$ generated by $p_\theta(X)$. A statistic of the sample $T(X^n)$ is said to be sufficient if it captures all the information about θ available in the sample. In information theoretic terms this can be expressed by the requirement: $I(X^n;\theta) = I(T(X^n);\theta)$, where $I(X;Y)$ is the mutual information between X and Y (Cover & Thomas, 1991). Unfortunately, non-trivial sufficient statistics exist, in general, only for very special distributions, exponential parametric families (Pitman, 1936). For such families a sufficient statistic which is an additive function of the sample (i.e. $T(X^n) = \sum T(X_i)$) can always be found. The question we address in this paper is how to extract such most informative features, or statistics, in cases where no sufficient statistics are known to exist. We also extend the notion of sufficiency, or measurements that preserve information, beyond the original context of parameter estimation

To illustrate our motivation, consider a die with an unknown outcome distribution. Suppose we are given the mean outcome of a die roll. What information does it provide about the probability to obtain 6 in rolling this die? One possible answer to this question was suggested by Jaynes (Jaynes, 1957) using his "Maximum Entropy ($MaxEnt$) principle".

Denote by $X = \{1, \ldots, 6\}$ the possible outcomes of the die roll, and by $\vec{\phi}(x)$ an observation, feature, or function of X. In the example of the expected outcome of the die, the observation is $\vec{\phi}(x) = x$, the specific outcome. Given the empirical expected value $\vec{a} = \langle\vec{\phi}(x)\rangle$, $MaxEnt$ argues that the "most likely" outcome distribution $\tilde{p}(x)$ is the one with maximum entropy among all distributions satisfying the observation constraint, $\langle\vec{\phi}(x)\rangle_{\tilde{p}(x)} = \vec{a}$. This distribution depends, of course, on the actual value of the observed expectation, $\vec{a}$.

The $MaxEnt$ principle does not tell us, however, what are the observables whose expected values provide the *maximal* information about some unknown properties of the underlying distribution (the probability of 6 in this case). This question is meaningful only when there are many possible underlying distributions and some prior over these distributions is available. Let us denote the random variable representing the identity of the possible distribution (possible dice) by Y. Then a

good feature, or observable, $\vec{\phi}(x)$ would be one which allows for the best prediction of the required property of Y (e.g. outcome 6) from X, given only the set of expected values $\langle\vec{\phi}(x)\rangle_{p(x|y)}$, as well as the marginal distribution of X and Y. Clearly, for this example the best observation is simply the expected number of times 6 has occurred, but what is the general procedure for finding such features?

In this work, we provide a complete information theoretic answer to this question and present an algorithm for finding such optimal features, or observables. It exploits the fact that exponential forms are the only (non-trivial) joint distributions over X and Y whose mutual information is completely captured by the low-dimensional functions in the exponent, given the marginal distributions. We argue that the optimal features can be obtained by minimization of the Kullback-Leibler divergence, and provide a convergent iterative algorithm for solving this problem. The convergence of the algorithm is proved using information geometrical tools. We show that our problem is equivalent to approximating the co-occurrence data by a specific exponential form, whose natural sufficient statistics are the required optimal features, hence the name of our method: Sufficient Dimensionality Reduction (SDR).

The problem as formulated provides a general procedure 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 is presented here. As we discuss further, this optimization problem is similar in structure to the fundamental information theoretic problem of finding the capacity of an unknown channel, which suggests other interesting interpretations for our procedure. It is also closely related to the recently introduced *Information Bottleneck Method* (Tishby et al., 1999) as well as to other embedding algorithms, such as latent semantic indexing (Deerwester et al., 1990), locally linear embedding (Roweis & Saul, 2000), and non-negative matrix factorization (Lee & Seung, 1999).

2. Problem formulation

Consider a joint probability distribution, $p(x,y)$ over the variables X and Y. A candidate feature vector for X is a d-dimensional function $\vec{\phi}(x) : X \to \Re^d$. We define the *measurement* of $\vec{\phi}(x)$ on $p(x,y)$ as the set of $|Y|$ expected values $\langle\vec{\phi}(x)\rangle_{p(x|y)}$ for all values of y. This can be understood as the value of the statistic $\vec{\phi}(X^n) = \frac{1}{n}\sum_{i=1\ldots n}\vec{\phi}(X_i)$ where X^n is an i.i.d. sample from $p(x|y)$ and n goes to infinity.

The question that concerns us is what fraction of the information between X and Y can be obtained from knowledge of *only* the measurement values of $\vec{\phi}(x)$ as well as the marginal distributions $p(x)$ and $p(y)$. In other words, if we restrict our knowledge of X and Y to their marginals and the measurement of $\vec{\phi}(x)$, how well can we predict Y from X? This question can be formalized using Shannon's mutual information (Cover & Thomas, 1991).

We define the *information in the measurement* of $\vec{\phi}(x)$ on $p(x,y)$ as:

$$I_M(\vec{\phi}(x),p) \equiv min_{\tilde{p}(x,y)\in\mathcal{P}(\vec{\phi}(x)),p)} I[\tilde{p}(x,y)] \qquad (1)$$

where $I[\tilde{p}(x,y)] = \sum_{x,y}\tilde{p}(x,y)\log\frac{\tilde{p}(x,y)}{\tilde{p}(x)\tilde{p}(y)}$ is the mutual information between X and Y under the distribution $\tilde{p}(x,y)$. The set $\mathcal{P}(\vec{\phi}(x),p)$ is the set of distributions that satisfy the constraints, defined by

$$\mathcal{P}(\vec{\phi}(x),p) \equiv \left\{\tilde{p}(x,y) : \begin{array}{r}\langle\vec{\phi}(x)\rangle_{\tilde{p}(x|y)} = \langle\vec{\phi}(x)\rangle_{p(x|y)} \\ \tilde{p}(x) = p(x) \\ \tilde{p}(y) = p(y)\end{array}\right\}. \qquad (2)$$

The above definition reflects the requirement that *only* the measurement values of $\vec{\phi}(x)$ are known, as represented in the constrained expectations.

The optimal features $\vec{\phi}^*(x)$ are those whose measurements provides maximal information. Thus:

$$\vec{\phi}^*(x) = \arg\max_{\vec{\phi}(x)} I_M(\vec{\phi}(x),p) \qquad (3)$$

Plugging in the definition of I_M we have the following Max-Min problem:

$$\vec{\phi}^*(x) = \arg\max_{\vec{\phi}(x)} \min_{\tilde{p}(x,y)\in\mathcal{P}(\vec{\phi}(x),p)} I(\tilde{p}(x,y)) \ . \qquad (4)$$

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 $\tilde{p}(x,y)$, is necessarily of a special exponential form and can be interpreted as a generative model in this class. However, 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 $\langle\vec{\phi}(x)\rangle_{p(x|y)}$ for every y, given the candidate features $\vec{\phi}(x)$. In practice these expectations are estimated from finite samples and empirical distributions.

In the next section we show that the problem of finding the optimal functions $\vec{\phi}(x)$ is equivalent to the dual problem of finding the optimal features for Y and in fact the two problems are solved simultaneously.

3. The nature of the solution

We first show that the problem as formulated in Eq. 4 is equivalent to the problem of minimizing the KL divergence between the empirical distribution $p(x,y)$ and a 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 $\langle\vec{\phi}(x)\rangle_{p(x|y)}$ 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 (Cover & Thomas, 1991),

$$\tilde{p}^*_{\vec{\phi}(x)}(x,y) = \frac{1}{Z}\exp\left(\vec{\phi}(x)\cdot\vec{\psi}(y) + A(x) + B(y)\right) , \tag{5}$$

where $Z = \sum_{x,y}\exp\left(\vec{\phi}(x)\cdot\vec{\psi}(y) + A(x) + B(y)\right)$ is the normalization (partition) function [1].

The functions $\vec{\psi}(y), A(x), B(y)$ are uniquely determined as Lagrange multipliers from the expectation values $\langle\vec{\phi}(x)\rangle_{p(x|y)}$ and the marginals.

One key property of the exponential form, Eq. 5, is that it is the only joint distribution for which the mutual information between X and Y is completely captured by the dual function vectors, $\vec{\psi}(y), \vec{\phi}(x)$. Namely, it is the only joint distribution $\tilde{p}(x,y)$ for which the following equality holds: $I[\tilde{p}(x,y)] = I(\psi;\phi)$. This interesting property will be discussed elsewhere.

The distributions of the form Eq. 5 can also be viewed as a distribution class parametrized by the infinite family of functions $\Theta = [\vec{\psi}(y), \vec{\phi}(x), A(x), B(y)]$ (note we treat ψ and ϕ symmetrically). We thus denote this class by P_Θ.

The above discussion shows that for every candidate feature $\vec{\phi}(x)$, the minimum information in Eq. 4 is the information in the distribution $\tilde{p}^*_{\vec{\phi}(x)}$. Let us define the set of information minimizing distributions $\mathcal{P}_\Phi \subset P_\Theta$:

$$\mathcal{P}_\Phi = \left\{\tilde{p}\in P_\Theta : \exists\vec{\phi}(x) : \quad \tilde{p} = \tilde{p}^*_{\vec{\phi}(x)}\right\} . \tag{6}$$

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

It can be shown that $\tilde{p}^*_{\vec{\phi}(x)}$ is the only distribution in P_Θ satisfying the constraints $\mathcal{P}(\vec{\phi}(x),p)$ (Della-Pietra et al., 1997). Thus, $\mathcal{P}_\Phi$ can alternatively be defined as:

$$\mathcal{P}_\Phi = \left\{\tilde{p}\in P_\Theta : \exists\vec{\phi}(x) : \quad \tilde{p}\in\mathcal{P}(\vec{\phi}(x),p)\right\} . \tag{7}$$

Finding the optimal feature $\vec{\phi}^*(x)$ now amounts to finding the distribution in $\mathcal{P}_\Phi$ with maximum information:

$$\tilde{p}^* = \arg\max_{\tilde{p}\in\mathcal{P}_\Phi} I[\tilde{p}] . \tag{8}$$

The optimal $\vec{\phi}^*(x)$ will be the ϕ parameters of $\tilde{p}^*$.

For every $\tilde{p}\in\mathcal{P}_\Phi$ one can easily show that

$$I[\tilde{p}] = I[p] - D_{KL}[p|\tilde{p}] , \tag{9}$$

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 $\tilde{p}\in\mathcal{P}_\Phi$ is equivalent to minimizing $D_{KL}[p|\tilde{p}]$ for $\tilde{p}\in\mathcal{P}_\Phi$:

$$\tilde{p}^* = \arg\min_{\tilde{p}\in\mathcal{P}_\Phi} D_{KL}[p|\tilde{p}] . \tag{10}$$

In addition, Eq. 9 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 $\vec{\phi}^*(x)$.

The problem in Eq. 10 is a maximization of a function over a subset of P_Θ, namely $\mathcal{P}_\Phi$. The following proposition shows that this is in fact equivalent to maximizing the same function over all of P_Θ.

Proposition 1

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

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

$$\begin{array}{lll} \forall x & \langle\vec{\phi}(x)\rangle_{\tilde{p}(x|y)} = \langle\vec{\phi}(x)\rangle_{p(x|y)} \\ \forall y & \langle\vec{\psi}(y)\rangle_{\tilde{p}(y|x)} = \langle\vec{\psi}(y)\rangle_{p(y|x)} \\ \forall x & \tilde{p}(x) = p(x) \\ \forall y & \tilde{p}(y) = p(y) . \end{array} \tag{11}$$

Clearly, this distribution satisfies the constraints in $\mathcal{P}(\vec{\phi}(x),p)$, and from Eq. 7 we conclude that it is in $\mathcal{P}_\Phi$. □

Our problem is thus equivalent to the minimization problem:

$$p^* = \arg\min_{\tilde{p}\in P_\Theta} D_{KL}[p|\tilde{p}] \ . \tag{12}$$

Equation 12 is symmetric with respect to ϕ and ψ, thus removing the asymmetry between X and Y in the formulation of Eq. 4.

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.

4. An Iterative Projection Algorithm

We now present an iterative algorithm for solving the Max-Min variational problem, Eq. 4, which provably converges to a local minimum of Eq. 12. We describe the algorithm using the information geometric notion of *I-projections* (Csiszar, 1975).

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 $D_{KL}[p|q]$.

$$p^*(x) = \arg\min_{p\in\mathcal{P}} D_{KL}[p|q] \ . \tag{13}$$

We 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:

$$D_{KL}[p|q] \leq D_{KL}[p|p^*] + D_{KL}[p^*|q] \ . \tag{14}$$

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 $\vec{\phi}(x)$ and a distribution $p(x)$, we denote the set of distributions which agree with $p(x)$ on the expectation values of $\vec{\phi}(x)$, by $\mathcal{P}(\vec{\phi}(x), p(x))$. Namely,

$$\mathcal{P}(\vec{\phi}(x), p(x)) = \left\{\tilde{p}(x) : \langle\vec{\phi}(x)\rangle_{\tilde{p}(x)} = \langle\vec{\phi}(x)\rangle_{p(x)}\right\} \tag{15}$$

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

$$IPR(q(x), \mathcal{P}(\vec{\phi}(x), p(x))) = \frac{1}{Z^*} q(x) \exp\left(\vec{\lambda}^* \cdot \vec{\phi}(x)\right) \ , \tag{16}$$

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

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

Output: $2d$ feature functions: $\vec{\psi}(y)$ and $\vec{\phi}(x)$ that result from $\tilde{p}^*$, a solution of the variational problem Eq. 4 and a (local) minimum of Eq. 10.

Initialization:

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

Iterate:

- For all x, set:

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

  The functions $\vec{\phi}_{t+1}(x)$ are determined as Lagrange multipliers

- For all y, set:

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

  The functions $\vec{\psi}_{t+1}(y)$ are determined as Lagrange multipliers

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

Figure 1. The iterative projection algorithm.

Before describing the algorithm, we need some additional notations:

- $\tilde{p}_t(x,y)$ - the distribution after t iterations.
- $\vec{\psi}_t(y)$ - the $\vec{\psi}(y)$ functions for $\tilde{p}_t(x,y)$.
- $\vec{\phi}_{t+1}(x)$ - the $\vec{\phi}(x)$ functions for $\tilde{p}_{t+1}(x,y)$.
- Θ_t - the full parameter set for $\tilde{p}_t(x,y)$.

The iterative projection algorithm is outlined in figure 1. *I-projections* of (exponential) distributions are applied iteratively: Once for fixed $\vec{\psi}(y)$ and their expectations, and then for fixed $\vec{\phi}(x)$ and their expectations. Interestingly, during the first projection, the functions $\vec{\phi}(x)$ are modified as Lagrange multipliers for $\langle\vec{\psi}(y)\rangle$, and vice-versa in the second projection. The iteration can thus be viewed as alternating mapping between the two sets of d-dimensional functions, $\vec{\psi}(y)$ and $\vec{\phi}(x)$. This is also the direct goal of the variational problem.

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 $D_{\mathrm{KL}}[p|\tilde{p}_{t+1}] \leq D_{\mathrm{KL}}[p|\tilde{p}_t]$.

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 $\mathcal{P}(\vec{\psi}_t(y), p(y|x))$.
- $p(y|x)$ is also in $\mathcal{P}(\vec{\psi}_t(y), p(y|x))$.

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:

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

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

$$D_{KL}[p|\tilde{p}_t] = D_{KL}[p|\tilde{p}_{t+1}] + D_{KL}[\tilde{p}_{t+1}|\tilde{p}_t] \ . \tag{18}$$

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

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

$$D_{KL}[p|\tilde{p}_t] \geq D_{KL}[p|\tilde{p}_{t+1}] \ . \tag{19}$$

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

An analogous argument proves that:

$$D_{KL}[p|\tilde{p}_{t+2}] \leq D_{KL}[p|\tilde{p}_{t+1}] \ . \tag{20}$$

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. 11.

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. 18. We will give a detailed proof in a separate paper.

4.1 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 (Della-Pietra et al., 1997) or conjugate gradient methods, can also be used and may improve convergence rate.

Input: Distributions $q(x), p(x)$, d functions $\vec{\phi}(x)$

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

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

Iterate:

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

Figure 2. The Generalized Iterative Scaling Algorithm.

5. 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 $doc \in doc_1, \ldots, doc_m$ is denoted $p(w|doc)$ and is obtained from the normalized word count vector for this document:

$$p(w|doc) = \frac{n(doc, w)}{\sum_w n(doc, w)} \ , \tag{21}$$

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 $\vec{\phi}(w)$ such that using their mean values alone we can infer information about document identity. This approach differs 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.

Our approach extends the works on maximum entropy in NLP (Berger et al., 1996; Della-Pietra et al., 1997). In these works, the set of features $\vec{\phi}(w)$ 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 $\vec{\phi}(w)$.

5.1 Indexing and classification using SDR

The output of the SDR algorithm can be used to represent every document by an index in $\Re^d$. This transformation reduces the representation of a document from $|W|$ to d dimensions. Such a dimensionality reduction 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. The resulting indices can be used for document retrieval or categorization.

A natural index is the expected value of $\vec{\phi}(w)$ for the given document: $\hat{\phi}(doc) \equiv \langle \vec{\phi}(w) \rangle_{p(w|doc)}$.

Another possible index is the set of Lagrange multipliers , associated with a given document. These denote the relative weights of the different features for that specific document. They are given by the Lagrange multipliers resulting from I-projecting $\frac{1}{Z}\exp A(w)$ on the set of constrains $\mathcal{P}(\vec{\phi}(w), p(w|doc))$. We denote this index by $\hat{\psi}(doc)$.

5.2 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 SDR 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 shows $D_{KL}[p|p^*]$, $I[p^*]$

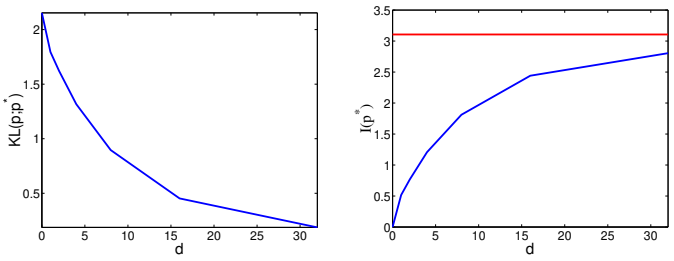

Figure 3. Left: KL between the original distribution and the model, for different d values. Right: Information in the model. The horizontal line marks the information in the original distribution.

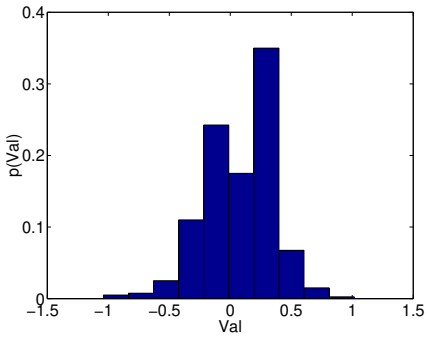

Figure 4. Histogram of $\vec{\phi}(w)$ values.

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 9).

To gain insight into the nature of the SDR features, we look at $\vec{\phi}(w)$ obtained for the $d = 1$ case. A histogram of the values of $\vec{\phi}(w)$ 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 $\vec{\phi}(w)$ values. Clearly, the words with high $\vec{\phi}(w)$ correspond to the "comp.graphics" subject, and the ones with low $\vec{\phi}(w)$ correspond to "alt.atheism". This single feature thus maps the words into a continuous scale through which it assigns positive weights to one class, negative weights to the other, and negligible weights to words which are possibly irrelevant to the classes.

We next performed classification on eight different pairs of newsgroups. The $\hat{\psi}(doc)$ and $\hat{\phi}(doc)$ indices were used as input to a support vector machine (SVM-Light (Joachims, 1998)), which was trained to classify documents according to their subjects. Baseline results were obtained by running the SVM classifier on the original $p(w|doc)$ vectors. The training and testing

universe	format
laws	pov
behavior	formats
moral	files
islamic	gif

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

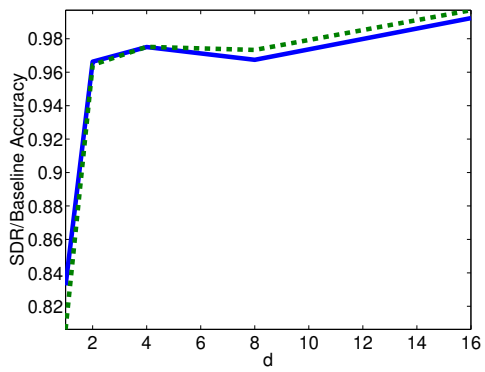

Figure 6. Fraction of baseline performance achieved by using a given number of SDR features. Results are averaged over eight newsgroup pairs. The solid line corresponds to the $\hat{\phi}(doc)$ index and the dashed line to the $\hat{\psi}(doc)$ index.

sets consisted of 500 documents per subject, and 500 words. We experimented with values $d = 1, 2, 4, 8, 16$ to test how well we can classify with relatively few features. Figure 6 shows the fraction of the baseline performance that can be achieved using a given number of features. It can be seen that even when using only four features, 97% of the baseline performance can be achieved.

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.

6. 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. This is achieved 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.

6.1 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 seem 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 fat-shattering dimension of the features, etc.

6.2 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. The optimal features are not unique in the following sense: since 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 equivalent to $\vec{\phi}(x)$ and $\vec{\psi}(y)$. 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.(5). 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).

6.3 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 vector $\vec{\phi}(x)$ and the decoding is performed through a dot-product of these two vectors. This intriguing interpretation of our algorithm obviously requires further analysis.

6.4 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, prove very useful for continuous data, they do not provide a principled method for statistical co-occurrence data, where linearity assumptions are not necessarily justified. 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.

Acknowledgement

We thank Noam Slonim and Gal Chechik for helpful discussions and comments and for the help with experimental data. This research was supported by a grants from the Israeli Academy of Sciences and from the Ministry of Science. A.G. is supported by the Eshkol Foundation.

References

Berger, A., Della-Pietra, S., & Della-Pietra, V. (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.

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

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

Jaynes, E. (1957). Information theory and statistical mechanics. *Physical Review*, *106*, 620.

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.

Pitman, E. (1936). Sufficient statistics and intrinsic accuracy. *Proc. of the Cambridge Phil. Soc.*, *32*, 567–579.

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).