INTERNATIONAL COMPUTER SCIENCE INSTITUTE

1947 Center St. • Suite 600 • Berkeley, California 94704-1198 • (510) 643-9153 • FAX (510) 643-7684

Unsupervised Learning from Dyadic Data

Thomas Hofmann
International Computer Science Institute, Berkeley, CA &
Computer Science Division, UC Berkeley
hofmann@cs.berkeley.edu

and

Jan Puzicha
Institut für Informatik III
University of Bonn, Germany
jan@cs.uni-bonn.de

TR-98-042
December 1998

Abstract

Dyadic data refers to a domain with two finite sets of objects in which observations are made for *dyads*, i.e., pairs with one element from either set. This includes event co-occurrences, histogram data, and single stimulus preference data as special cases. Dyadic data arises naturally in many applications ranging from computational linguistics and information retrieval to preference analysis and computer vision. In this paper, we present a systematic, domain-independent framework for unsupervised learning from dyadic data by statistical *mixture models*. Our approach covers different models with flat and hierarchical latent class structures and unifies probabilistic modeling and structure discovery. Mixture models provide both, a parsimonious yet flexible parameterization of probability distributions with good generalization performance on sparse data, as well as structural information about data-inherent grouping structure. We propose an *annealed* version of the standard Expectation Maximization algorithm for model fitting which is empirically evaluated on a variety of data sets from different domains.

1 Introduction

Over the past decade, *learning from data* has become a highly active field of research distributed over many disciplines ranging from pattern recognition and neural computation to statistics, machine learning, and data mining. Most domain-independent learning architectures as well as the underlying theories of learning have been focusing on a feature-based data representation by vectors or points in an Euclidean space. For this restricted case, substantial progress has been achieved. However, the focus on metric data has disregarded a variety of important problems which do not fit into this setting. Examples that have recently received some attention are proximity data [HB97, BWD97] which replace metric distances by the weaker notion of pairwise similarities and ranked preference data [CSS98]. A variety of other types of non–metrical data can be found, for example, in the psychometric literature [Coo64, Kru78, CA80], in particular in the context of multidimensional scaling and correspondence analysis.

1.1 Dyadic Data

In this paper, we introduce a general framework for unsupervised learning from *dyadic data*. The notion *dyadic* refers to a domain with two (abstract) sets of objects, $\mathcal{X} = \{x_1, \ldots, x_I\}$ and $\mathcal{Y} = \{y_1, \ldots, y_J\}$ in which the observations are made for *dyads* (x, y). In the simplest case – on which we focus in the sequel – an elementary observation consists of (x, y) itself, i.e., a *co-occurrence* of x and y, while other cases may also provide a scalar value $w(x, y)$ (e.g., a strength of preference/association or a rating). In some cases, additional observations of features for objects in $\mathcal{X}$ and/or $\mathcal{Y}$ may be available, but we will restrict our attention to the *immanent* case of purely dyadic observations. Some exemplary application areas of dyadic data are:

- *Computer vision*, in particular in the context of image segmentation, where $\mathcal{X}$ corresponds to image locations, $\mathcal{Y}$ to discretized or categorical feature values, and a dyad denotes the occurrence of a feature at a particular image location [HPB98].
- *Text-based information retrieval*, where $\mathcal{X}$ corresponds to a document collection, $\mathcal{Y}$ to the vocabulary, and a dyad represents the occurrence of a token in the content of a document [HPJ99].
- *Computational linguistics* in the corpus-based statistical analysis of word co-occurences which has applications in probabilistic language modeling, word clustering [PTL93], word sense disambiguation [Hin90, DLP97] and discrimination [Sch98], and automated thesaurus construction [SP97b].
- *Preference analysis and consumption behavior* by identifying $\mathcal{X}$ with individuals and $\mathcal{Y}$ with objects. Dyads then correspond to single stimulus preferences. This type of data is the starting point for a machine learning technique known as *collaborative filtering* [GNOD92, KMMH97].

1.2 Modeling Goals and Principles

Across different domains, there are two main tasks which play a fundamental role in unsupervised learning from dyadic data: (i) probabilistic modeling, i.e., learning a joint or conditional probability model over $\mathcal{X} \times \mathcal{Y}$, and (ii) structure discovery, e.g., identifying clusters and data hierarchies.

At a first glance, it may seem that statistical models for dyadic data are trivial. For object sets with a nominal scale the empirical co-occurrence frequencies are sufficient statistics, capturing all we know about the data. However, the intrinsic difficulty in modeling dyadic data is the *data sparseness*, also known as the *zero frequency problem* [Goo53, Goo65, Kat87, WB91]. When the product set $\mathcal{X} \times \mathcal{Y}$ is very large, a majority of pairs (x, y) only have a small probability of being observed in a given sample. Most of the empirical frequencies are typically zero or at least significantly

corrupted by sampling noise. The sparseness problem becomes even more urgent in the case of higher order co-occurrences where triplets, quadruples, etc. are observed as in n-gram language modeling. Typical state-of-the-art techniques in natural language processing apply smoothing to deal with zero frequencies of unobserved events. Prominent techniques are, for example, Katz's *back-off method* [Kat87], *model interpolation* with held–out data [JM80, Jel85], and similarity–based smoothing techniques [ES92, DLP97].

In contrast, we propose a model–based statistical approach and present a family of *latent class* [And97] or *finite mixture models* [TSM85, MB88] as a principled approach to deal with the data sparseness problem. Mixture and clustering models for dyadic data have been investigated before under the titles of class-based n-gram models [BdM$^+$92], distributional clustering [PTL93], and aggregate Markov models [SP97a] in natural language processing. All three approaches are recovered as special cases in our general learning framework. There is also a close relation to clustering methods for qualitative data like the *information clustering* approach (cf. [Boc74]).

The modeling principle of latent variable models is the specification of a joint probability distribution for latent and observable variables. This unifies *statistical modeling* and *structure detection*: a probabilistic model of the observables is obtained by marginalization, while Bayes' rule induces posterior probabilities on the latent space of structures w.r.t. given observations. The latter is crucial for exploratory data analysis where the extraction of natural structures such as data groups and data hierarchies are essential goals.

1.3 Latent Classes and Clusters

The common trait of mixture models is the assumption of an underlying *latent class* or cluster membership for observations or sets of observations. More formally, our starting point is an observation sequence $\mathcal{S} = (x^n, y^n)_{1 \leq n \leq N}$ [1] which is a realization of an underlying sequence of random variables $(X^n, Y^n)_{1 \leq n \leq N}$.

There are at least three different possibilities to define (flat) latent class models for dyadic data:

- The most direct way is to introduce an (unobserved) latent class for each *observation*, i.e., with each dyad (x^n, y^n) is associated a latent variable A^n over some finite set $\mathcal{A} = \{a_1, \ldots, a_K\}$. A realization $\mathbf{a} = (a^n)_{1 \leq n \leq N}$ effectively partitions the observations $\mathcal{S}$ into K classes. This type of model is called *aspect-based* and observations sharing the same class are simply referred to as an *aspect*.
- Alternatively, latent classes can be introduced for *objects* in *one* of the spaces, resulting in a model that is referred to as *one-sided* or asymmetric clustering model. Without loss of generality we consider latent variables $C(x)$ over $\mathcal{C} = \{c_1, \ldots, c_K\}$ and denote by $\mathbf{c} = (c(x))_{x \in \mathcal{X}}$ a realization which partitions $\mathcal{X}$. Reversing the role of $\mathcal{X}$ and $\mathcal{Y}$ results in a different model.
- Finally, latent classes can be defined for both sets simultaneously. Latent variables $C(x)$ over $\mathcal{C} = \{c_1, \ldots, c_K\}$ partition $\mathcal{X}$ while variables $D(y)$ over $\mathcal{D} = \{d_1, \ldots, d_L\}$ partition $\mathcal{Y}$. This type of model is called *two-sided* or symmetric clustering model.

The difference between aspect and clustering models has to be emphasized. While the latent class structure of an aspect model partitions the *observations*, clustering models provide a group structure on *object spaces*. As a consequence, identical observations (x, y) may have different latent classes in aspect models, whereas latent variables are shared by sets of observations in clustering models. Although the underlying latent class structures are very different in either case, we will show in the sequel how clustering models can be derived as *constrained aspect models*.

The rest of the paper is organized as follows: In Section 2 we will give an overview over different mixture models for dyadic data including extensions to hierarchical and more structured latent class

[1] We follow the convention to utilize subscript indices for unique names of members of sets, while superscript indices are used to number observations.

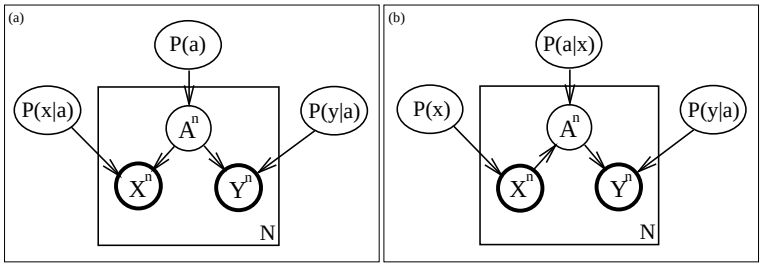

Figure 1: Graphical model representation of the aspect model: (a) in the symmetric parameterization and (b) in the asymmetric parameterization. Circles denote random variables (bold circles correspond to observed quantities), ellipses represent parameters, and rectangular frames are utilized for multiple instances. The directed graph structure represents conditional independence properties in the standard way.

models and derive Expectation Maximization (EM) algorithms for maximum likelihood model fitting. Section 3 discusses generalizations and modifications of the pure maximum likelihood framework by annealed EM, cross validation, as well as a number of acceleration techniques. A variety of experimental results from different application domains is presented in Section 4.

2 Mixture Models for Dyadic Data

2.1 Aspect Model

Model Specification The aspect model is built on the assumption that all co-occurrences in the sample $\mathcal{S}$ are i.i.d. and that the pairs of random variables X^n and Y^n are conditionally independent given the respective latent class A^n. The randomized data generation process can be described as follows:

1. choose an aspect a with probability $P(a)$,
2. select a $\mathcal{X}$–object $x \in \mathcal{X}$ with probability $P(x|a)$, and
3. select a $\mathcal{Y}$–object $y \in \mathcal{Y}$ with probability $P(y|a)$.

The corresponding complete data probability, i.e., the joint probability of the data and a hypothetical instantiation for the latent variables, is given by

$$P(\mathcal{S}, \mathbf{a}) = \prod_{n=1}^{N} P(x^n, y^n, a^n), \quad \text{where} \tag{1}$$

$$P(x^n, y^n, a^n) = P(a^n)\, P(x^n|a^n)\, P(y^n|a^n) \ . \tag{2}$$

A graphical model representation of the aspect model is depicted in Figure 1. By summing over all possible realizations of the latent variables and grouping identical dyads together, one obtains the usual mixture probability distribution on the observables

$$P(\mathcal{S}) = \prod_{x \in \mathcal{X}} \prod_{y \in \mathcal{Y}} P(x,y)^{n(x,y)} \quad \text{with} \tag{3}$$

$$P(x,y) = \sum_{a \in \mathcal{A}} P(a) P(x|a) P(y|a) \ . \tag{4}$$

Here $n(x,y) = |\{(x^n, y^n) : x^n = x \wedge y^n = y\}|$ represents the empirical co-occurrence frequencies. The corresponding marginal counts for objects x and y are denoted by $n(x)$ and $n(y)$, respectively.

Expectation Maximization Algorithm Maximum likelihood estimation requires to maximize the log-likelihood $\log P(\mathcal{S};\theta)$ with respect to the model parameters, which are concatenated in a vector θ for notational convenience. To overcome the difficulties in maximizing a log of a sum, we apply the Expectation Maximization (EM) framework [DLR77, MK97] and alternate two re-estimation steps:

- an Expectation (E)-step for estimating the posterior probabilities of the unobserved mapping $P(\mathbf{a}|\mathcal{S};\theta')$ for a given parameter estimate θ',
- a Maximization (M)-step, which involves maximization of the *expected complete data log–likelihood* $\mathcal{L}(\theta|\theta') = \sum_{\mathbf{a}} P(\mathbf{a}|\mathcal{S};\theta') \log P(\mathcal{S},\mathbf{a};\theta)$ for given posterior probabilities with respect to θ.

The EM algorithm is known to increase the observed likelihood in each step, and converges to a (local) maximum under mild assumptions [MK97].

In the aspect model, θ contains all continuous parameters, namely $P(a)$, $P(x|a)$ and $P(y|a)$. The E-step equations for the class posterior probabilities in the aspect model can be derived from Bayes' rule and are given by

$$\mathrm{P}\{A^n = a | X^n = x, Y^n = y; \theta\} = \frac{P(a)P(x|a)P(y|a)}{\sum_{a'} P(a)P(x|a')P(y|a')} \quad . \tag{5}$$

In the sequel the latter is simply denoted by $P(a|x,y;\theta)$. It is straightforward to derive the M-step re-estimation formulae by differentiating $\mathcal{L}$ w.r.t. the different components of θ. After introducing appropriate Lagrange multipliers to ensure the correct normalization one obtains

$$P(a) = \frac{1}{N} \sum_{n=1}^{N} P(a|x^n, y^n; \theta') \quad , \tag{6}$$

$$P(x|a) = \frac{\sum_{n:x^n=x} P(a|x^n, y^n; \theta')}{\sum_{n=1}^{N} P(a|x^n, y^n; \theta')} \quad , \tag{7}$$

$$P(y|a) = \frac{\sum_{n:y^n=y} P(a|x^n, y^n; \theta')}{\sum_{n=1}^{N} P(a|x^n, y^n; \theta')} \quad . \tag{8}$$

Notice, that it is unnecessary to store all $N \cdot K$ posteriors, as the E- and M-step can be efficiently interleaved.

Asymmetric Parameterization and Cross Entropy Minimization To achieve a better understanding of the aspect model, it is elucidating to switch to an equivalent asymmetric parameterization (cf. Figure 1 (b))

$$P(x,y) = P(x)P(y|x) = P(x) \sum_{a \in \mathcal{A}} P(a|x)P(y|a) \quad . \tag{9}$$

From (9 one arrives at an interpretation of the aspect model in terms of a low-dimensional representation of conditional probabilities $P(y|x)$ (or $P(x|y)$, by symmetry). The class-conditionals $P(y|a)$ can be interpreted as non-negative J-dimensional vectors $P(\cdot|a)$ which span a subspace of dimensionality $\leq K$ (more precisely, they span a sub-simplex in the simplex of multinomial distributions over $\mathcal{Y}$). The determination of $P(a|x)$ is then equivalent to an optimal approximation of $P(y|x)$ in that subspace. In this view, the parameters $P(a|x)$ correspond to coordinates in the subspace spanned by $\{P(\cdot|a_k)\}_{1 \leq k \leq K}$. By reversing the role of $\mathcal{X}$ and $\mathcal{Y}$ an equivalent dual formulation is obtained. In fact, this shows that for the aspect model all conditional distributions $P(y|x)$ for a fixed x are obtained by convex combination of the prototypical class-conditionals $P(y|a)$. Notice

that the estimation of $P(x) = n(x)/N$ is decoupled from the other parameters. Hence, maximizing the joint and predictive likelihood is equivalent.

It is well–known that, in general, maximum likelihood estimation can be equivalently stated as the minimization of the cross entropy (or Kullback–Leibler divergence) between the empirical distribution and the model distribution. In the case of co-occurrence data, the empirical distribution is given by $\hat{P}(x,y) \equiv \frac{1}{N} n(x,y)$ and in the asymmetric parameterization one may rewrite the log–likelihood as

$$\frac{1}{N}\mathcal{L}(\theta;\mathcal{S}) = \sum_{x,y} \hat{P}(x,y) \log[P(x) \sum_{a} P(a|x)P(y|a)] \tag{10}$$

$$= \sum_{x} \hat{P}(x) \left[\log P(x) + \sum_{y} \hat{P}(y|x) \log \sum_{a} P(a|x)P(y|a) \right]. \tag{11}$$

Since the estimation of $P(x)$ can be carried out independently, maximum likelihood estimation for the remaining parameters is equivalent to minimizing a sum over cross entropies between empirical conditional distributions $\hat{P}(y|x)$ and model distributions $P(y|x) = \sum_a P(a|x)P(y|x)$ weighted with $n(x) = N\hat{P}(x)$. Because of the symmetry of the model, an equivalent decomposition is obtained by interchanging the role of the sets $\mathcal{X}$ and $\mathcal{Y}$.

Probabilistic Factor Analysis for Discrete Data The dimensionality reduction obtained by the aspect model is similar in spirit to *factor analysis* [Bar87]. The factor analysis of co-occurrence data is also known as *Latent Semantic Analysis* (LSA) [LS89, DTGL90]. In LSA, the co-occurrence matrix of counts $\mathbf{C} = (n(x_i, y_j))_{i,j}$ is decomposed by *Singular Value Decomposition* (SVD) into $\mathbf{C} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^t$ with orthogonal matrices $\mathbf{U}$ and $\mathbf{V}$ and a diagonal matrix $\mathbf{\Sigma}$ containing the singular values of $\mathbf{C}$. The approximation obtained by thresholding all but the largest K singular values to zero is rank K optimal in the sense of the L_2-matrix norm.

Similarly, the aspect model can be represented as a matrix decomposition, where $\hat{\mathbf{U}} = (P(x_i|a_k))_{i,k}$, $\hat{\mathbf{V}} = (P(y_j|a_k))_{j,k}$, and $\hat{\mathbf{\Sigma}} = \mathrm{diag}(P(a_k))_k$. The joint probability model $\mathbf{P}$ in matrix notation can be written as the product $\mathbf{P} = \hat{\mathbf{U}}\hat{\mathbf{\Sigma}}\hat{\mathbf{V}}^t$. The weighted sum over outer products between rows of $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ reflects the conditional independence assumption and the K factors correspond to the mixture components. The mixing proportions substitute the singular values of the SVD.

The most important advantages of the aspect model compared to LSA are: (i) The aspect model utilizes a statistically meaningful divergence measure, namely the cross–entropy between probability distributions, to measure the quality of a low-dimensional approximation. In contrast, LSA is based on a least squares principle for the approximation of co-occurrence counts. This is equivalent to an additive Gaussian noise assumption which is not adequate for frequency tables. (ii) The mixture approximation yields a well-defined probability distribution and factors have a clear probabilistic meaning in terms of component distributions. In contrast, the LSA–model does not define a properly normalized probability distribution and, even worse, may contain negative entries. In this regard, the aspect model can be understood as a novel factor analysis approach for count data which utilizes the likelihood as an objective function and does not make additional assumptions about non-intrinsic loss functions or association measures.

2.2 One-Sided Clustering Model

Model Specification The *asymmetric* or *one-sided clustering model* differs from the aspect model in that latent class variables are associated with each object x and not with single observations (x^n, y^n). It is thus a clustering model in the strict sense as it assumes a definite partition of the $\mathcal{X}$–space into groups. To make the connection with the aspect model explicit, we modify the latter by introducing additional latent clustering variables $C(x) \in \{c_1, \ldots, c_K\}$, where latent variable

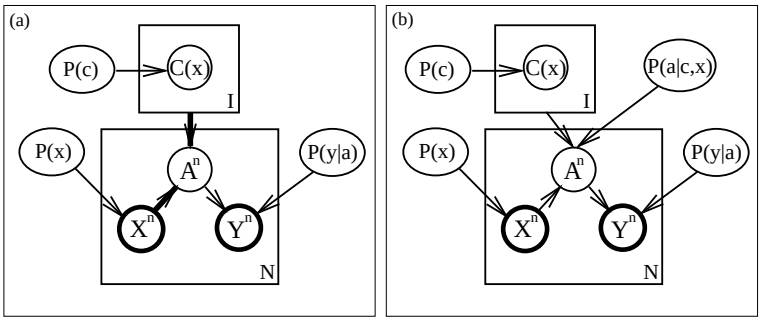

Figure 2: Graphical model representation of (a) the one-sided clustering model and the (b) hierarchical clustering model.

states for cluster and aspects are identified, $c_k \cong a_k$. The clustering variables impose the following consistency constraints on the aspect variables:

$$P(a|x,c) \equiv \mathrm{P}\{A^n = a | X^n = x, C(x) = c\} = \delta_{ac}, \tag{12}$$

with the notation $\delta_{ac} = 1$, if a and c are identified and $\delta_{ac} = 0$, otherwise. These constraints restrict the latent variable space and guarantee that all observations for a particular x share the same latent aspect which is identified with $c(x)$. The corresponding graphical model representation is shown in Figure 2 (a). Notice that compared to the asymmetric parameterization of the aspect model, the parameters for conditional distributions $P(a|x)$ are replaced by $P(a|x,c)$ which are no free parameters, since they are determined by the consistency constraints (12) and are thus omitted in the graphical model representation. From Figure 2 (a) one can directly read off the complete data distribution along with the parameterization of the one-sided clustering model

$$P(\mathcal{S}, \mathbf{c}) = \prod_{x \in \mathcal{X}} P(c(x)) \prod_{y \in \mathcal{Y}} \left[P(x)P(y|c(x))\right]^{n(x,y)} \quad . \tag{13}$$

In (13), aspect variables have been replaced by the respective clustering variable according to the imposed constraints. Summing over the latent space of clustering variables, this effectively defines the mixture

$$P(\mathcal{S}) = \prod_{x \in \mathcal{X}} P(\mathcal{S}_x), \quad P(\mathcal{S}_x) = \sum_{c \in \mathcal{C}} P(c) \prod_{y \in \mathcal{Y}} \left[P(x)P(y|c)\right]^{n(x,y)} \quad . \tag{14}$$

Here, $\mathcal{S}_x$ denotes the set of all observations involving x. Notice that co-occurrences in $\mathcal{S}_x$ are not independent for given parameters (as they are in the aspect model), but are coupled by the shared latent variable $C(x)$.

Expectation Maximization Algorithm As before, it is straightforward to derive an EM algorithm for the one–sided clustering model. The update equations for the posterior probabilities (E-step) are given by

$$\mathrm{P}\{C(x) = c | \mathcal{S}_x, \theta\} \quad = \quad \frac{P(c) \prod_{y \in \mathcal{Y}} P(y|c)^{n(x,y)}}{\sum_{c'} P(c') \prod_{y \in \mathcal{Y}} P(y|c')^{n(x,y)}} \quad . \tag{15}$$

In contrast to the E–step equation for the aspect model in (5), all likelihood contributions for observations in $\mathcal{S}_x$ are combined in the calculation of posterior probabilities for the shared variable

$C(x)$. For the parameter update equation (M-step) one obtains

$$P(y|c) = \frac{\sum_{x\in\mathcal{X}} n(x,y)\,\mathrm{P}\{C(x)=c|\mathcal{S}_x,\theta\}}{\sum_{x\in\mathcal{X}} n(x)\,\mathrm{P}\{C(x)=c|\mathcal{S}_x,\theta\}}\ , \tag{16}$$

$$P(c) = \frac{1}{I}\sum_{x\in\mathcal{X}} \mathrm{P}\{C(x)=c|\mathcal{S}_x,\theta\}, \tag{17}$$

and $P(x) = n(x)/N$. Alternating posterior calculations (15) and parameter re-estimations (16,17) defines a convergent maximum likelihood procedure.

Cross–entropy Clustering and Naive Bayes' Classification It is illuminating to rewrite the class posterior probabilities in (15) as

$$\mathrm{P}\{C(x)=c|\mathcal{S}_x,\theta\} \propto P(c)\exp\left\{-n(x)\left(-\sum_{y\in\mathcal{Y}}\hat{P}(y|x)\log P(y|c)\right)\right\}. \tag{18}$$

Comparing (18) with standard models like the Gaussian mixture model [MB88] or probabilistic vector quantization [RGF92] demonstrates that the cross entropy between the empirical conditional probability $\hat{P}(y|x)$ and the class–conditional $P(y|c)$ serves as a distortion measure in the one-sided clustering model. Notice that with an increasing number of observations, $n(x)\to\infty$, posteriors (almost surely) approach extremal values $\{0,1\}$, an asymptotic behavior which differs from the aspect model as is obvious by comparing (15) with (5).

Eq. (18) also allows us to give the one-sided clustering model an interpretation in terms of information theory by identifying objects x with *sources* over the alphabet $\mathcal{Y}$. The conditional probabilities $P(y|c)$ correspond to K source codes with asymptotic expected codeword length $l(y) = \mathbf{E}[-\log P(y|c)]$ and the data are encoded in two stages by first specifying the code $c(x)$ to utilize for each x and then representing all co–occurring y in that code. The asymptotic average codeword length of a code based on $P(y|c(x))$ is exactly the cross entropy which governs the latent class posterior probabilities (cf. [CT91]).

The one-sided clustering model can also be viewed as an unsupervised version of the naive Bayes' classifier, if we give $\mathcal{Y}$ the interpretation of a feature space for $x\in\mathcal{X}$. Each sample set $\mathcal{S}_x$ then correspond to a set of feature values which are assumed to be conditionally independent. In the supervised setting, the M–step estimation is performed once over the training data and class posteriors for new data with unknown labels are computed according to the E–step equations. There are further possibilities to weaken the conditional independence assumption both for the aspect and the clustering model, e.g., by substituting the class-conditional multinomial distributions $P(y|c)$ by log–linear Markov models [WL90, Whi87] or tree dependency models [MJ98]. This direction, however, is not further pursued in this paper.

2.3 Two-sided Clustering Model

Model Specification In the two-sided clustering model, the latent structure consists of cluster partitionings $\mathbf{c}$ and $\mathbf{d}$ defined over $\mathcal{X}$ and $\mathcal{Y}$, respectively. The conditional model we propose, is defined by

$$P(x,y|c,d) \equiv P\{X^n=x, Y^n=y|C(x)=c, D(y)=d) = P(x)\,P(y)\,\phi(c,d) \tag{19}$$

where $\phi(c,d)\in\mathbb{R}_0^+$ are $K\cdot L$ *cluster association* parameters. Intuitively, these weights increase or decrease the probability of observing a dyad with associated cluster pair (c,d) relative to the unconditional independence model. In order for (19) to define a proper probabilistic model, we have to ensure a correct global normalization, which constrains the choice of possible cluster association

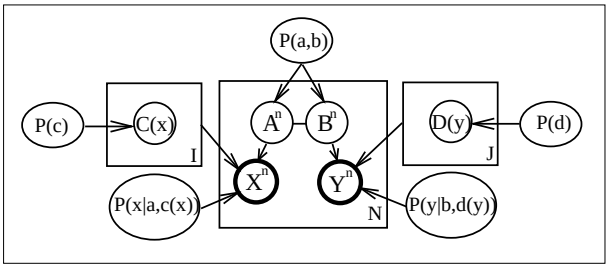

Figure 3: Graphical model representation of the two-sided clustering model.

parameters. Prior probabilities $P(\mathbf{c},\mathbf{d}) = [\prod_x P(c(x))] \cdot [\prod_y P(d(y))]$ complete the model specification. The two-sided clustering model enforces a strict partition in $\mathcal{X}$ as well as in $\mathcal{Y}$ and thus provides a principled approach to infer clustering structures simultaneously in both sets.

On the level of single co-occurrence observations, the simultaneous clustering in $\mathcal{X}$– and $\mathcal{Y}$–space implies that each observation has an associated latent aspect pair (A^n, B^n), the first/second component being identified with the latent class of the $\mathcal{X}/\mathcal{Y}$–object, respectively. Similar to the one–sided clustering model, we may interpret the two–sided clustering model in terms of a constrained aspect model by identifying $a_k \cong c_k$, $b_l \cong d_l$ and imposing two sets of compatibility conditions

$$P(a|x,c) \equiv \mathrm{P}\{A^n = a | X^n = x, C(x) = c\} = \delta_{ac}, \quad \text{and} \tag{20}$$

$$P(b|y,d) \equiv \mathrm{P}\{B^n = b | Y^n = y, D(y) = d\} = \delta_{bd}\,. \tag{21}$$

From the graphical model representation of the two–sided clustering model in Figure 3 we obtain[2]

$$P(x,y|a,b,C(x)=c,D(y)=d) = P(x|a,c)P(y|b,d)\,, \tag{22}$$

and by Bayes' rule together with the conditional independences implied by the graph in Figure 3 one may rewrite

$$P(x|a,C(x)=c) = \frac{P(a|x,c)P(x)}{P(a|C(x)=c)} = \delta_{ac}\frac{P(x)}{P(a)}, \tag{23}$$

$$P(y|b,D(y)=d) = \frac{P(b|y,d)P(y)}{P(b|D(y)=d)} = \delta_{bd}\frac{P(y)}{P(b)} \tag{24}$$

which results in

$$P(x,y|C(x)=c,D(y)=d) = P(x)P(y)\sum_{a,b}\delta_{ac}\delta_{bd}\frac{P(a,b)}{P(a)P(b)}\quad . \tag{25}$$

Comparing (25) with (19) suggests that the cluster association parameters $\phi(c,d)$ correspond to the ratio of the aspect probabilities $P(a,b)$ and the product of marginal probabilities $P(a)P(b)$ for the corresponding aspect pairs. The equivalence of both formulations of the two–sided clustering model in a maximum likelihood approach – via cluster associations ϕ on one hand and as a constrained aspect model on the other hand – is proven in the following paragraph.[3]

[2] In order for Figure 3 to be a correct graphical model one has to demand the technical condition that no cluster c_k or d_l is empty. Otherwise it might happen, that $P(x|a,C(x)=c)=0$ for all $x \in \mathcal{X}$.

[3] Notice that $P(c)$ and $P(a)$ (and similarly $P(d)$ and $P(b)$) are conceptually very different. $P(c)$ is the prior probability of an *object* $x \in \mathcal{X}$ to have latent *class* c, while $P(a)$ is the probability for the latent $\mathcal{X}$–*aspect* of an observation to be a.

Approximate Expectation Maximization Algorithm As the main difficulty in the two-sided clustering model the latent mappings $\mathbf{c}$ and $\mathbf{d}$ are coupled via the cluster association strengths $\phi(c,d)$. This is made explicit by performing the maximization of the (augmented) complete data likelihood which constitutes the M-step of the EM procedure. Inserting the maximum likelihood estimates $\hat{P}(x) = n(x)/N$ and $\hat{P}(y) = n(y)/N$ one has to maximize

$$\begin{aligned}\mathcal{L}(\phi|\phi') &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} n(x,y) \sum_{c\in\mathcal{C}}\sum_{d\in\mathcal{D}} \mathrm{P}\{C(x)=c, D(y)=d|\mathcal{S},\phi'\}\log\phi(c,d) \qquad (26)\\ &+\lambda\left[\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\frac{n(x)n(y)}{N^2}\sum_{c\in\mathcal{C}}\sum_{d\in\mathcal{D}}\mathrm{P}\{C(x)=c, D(y)=d|\mathcal{S},\phi'\}\phi(c,d)-1\right].\end{aligned}$$

with respect to ϕ. Solving for the Lagrange multiplier reveals that $\lambda = -N$. Differentiation of (26) then yields

$$\phi(c,d) = \frac{\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\mathrm{P}\{C(x)=c, D(y)=d|\mathcal{S},\phi'\}\frac{n(x,y)}{N}}{\left[\sum_{x\in\mathcal{X}}\mathrm{P}\{C(x)=c|\mathcal{S},\phi'\}\frac{n(x)}{N}\right]\left[\sum_{y\in\mathcal{Y}}\mathrm{P}\{D(y)=d|\mathcal{S},\phi'\}\frac{n(y)}{N}\right]}\,. \qquad (27)$$

It is straightforward to verify that the expression in the numerator of (27) is equivalent to the M–step re-estimation equation one would obtain for the parameters $P(a,b)$ by taking the graphical model in Figure 3 and the corresponding formulation in (25) as a starting point for EM. The terms in the denominator of (27) correspond exactly to the marginals $P(a)$ and $P(b)$ of $P(a,b)$. The remaining update equations are simply given by

$$P(c) = \frac{1}{I}\sum_{x\in\mathcal{X}}\mathrm{P}\{C(x)=c|\mathcal{S},\phi'\}, \quad P(d) = \frac{1}{J}\sum_{y\in\mathcal{Y}}\mathrm{P}\{D(y)=d|\mathcal{S},\phi'\}\ . \qquad (28)$$

From (27) it follows that posterior joint probabilities $\mathrm{P}\{C(x)=c, D(y)=d|\mathcal{S},\phi'\}$ have to be computed in the E-step in order to perform an exact EM-based likelihood maximization. To preserve the tractability of the model fitting procedure, an approximate E–step seems to be adequate. We propose to perform a variational approximation (also known as *mean field approximation*) and replace the exact computation of posteriors in the E-step by a factorial approximation, i.e.,

$$\mathrm{P}\{C(x)=c, D(y)=d|\mathcal{S},\phi\} \approx Q\{C(x)=c|\mathcal{S},\phi\}Q\{D(y)=d|\mathcal{S},\phi\}\ , \qquad (29)$$

and to utilize this approximation for the parameter re-estimation in the M-step. Here, Q is an approximating probability distribution which is chosen in order to minimize the KL–divergence to the true posterior distribution (cf. [NH98, JGJS98, HPB98]). The approximate E-step equations are given by (cf. Appendix)

$$Q\{C(x)=c|\mathcal{S},\phi\} \propto P(c)\exp\left[\sum_y n(x,y)\sum_d Q\{D(y)=d|\mathcal{S},\phi\}\log\phi(c,d)\right], \qquad (30)$$

$$Q\{D(y)=d|\mathcal{S},\phi\} \propto P(d)\exp\left[\sum_x n(x,y)\sum_c Q\{C(x)=c|\mathcal{S},\phi\}\log\phi(c,d)\right]. \qquad (31)$$

Notice that the mean field conditions form a highly non-linear, coupled system of equations. A solution is found by a fixed–point iteration which alternates the computation of the latent variables in one space (or more precisely their approximate posterior probabilities) based on the intermediate solution in the other space, and vice versa. However, the alternating computation has to be interleaved with a re-computation of the ϕ–parameters in between, i.e., after updating one set of

$Q\{C(x)=c|\mathcal{S},\phi\}$ and $Q\{D(y)=d|\mathcal{S},\phi\}$, because certain term cancelations have been exploited in the derivation of (30,31) (cf. Appendix). The resulting alternation scheme optimizes a common objective function in each step and always maintains a valid probability distribution.[4] Alternatively, Markov chain Monte Carlo (MCMC) methods could be applied to approximate the required posterior probabilities. Yet, the mean field approximation has the advantage to be more efficient due to its deterministic nature.

Mutual Information Clustering It is elucidating to consider the hard clustering limit of the two-sided clustering model. If we substitute all continuous parameters by their M-step re-estimation formula, we obtain a goodness-of-fit measure defined over instantiations $\mathbf{c}$ and $\mathbf{d}$, namely,

$$\mathcal{I}(\mathbf{c},\mathbf{d})=\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}}\frac{n(x,y)}{N}\log\phi(c(x),d(y))=\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}}P(a,b)\log\frac{P(a,b)}{P(a)P(b)}. \tag{32}$$

$\mathcal{I}$ can be identified as the expected mutual information between the random variables A^n and B^n. Thus in the hard clustering case, the two-sided clustering model reduces to maximizing the mutual information between the $\mathcal{X}$– and the $\mathcal{Y}$–part of the aspects variables. In a more sloppy formulation one may state that the mutual information between the partitions in $\mathcal{X}$– and $\mathcal{Y}$–space is maximized.

2.4 Hierarchical Clustering Model

Model Specification: Combining Aspects and Clusters The aspect model as well as the clustering models define a non–hierarchical, 'flat' latent class structure. However, especially in the context of structure discovery, it is important to find a hierarchical data organization. There are well–known learning architectures like the *Hierarchical Mixtures of Experts* [JJ94] which can fit hierarchical models to data. Yet, in the case of dyadic data there is an alternative way to define a hierarchical model by combining aspects and clusters. In the hierarchical clustering proposed here, aspects are identified with the nodes (inner and terminal) of a hierarchy – e.g., a complete binary tree – while clusters are identified with the terminal nodes. As before, both, a latent aspect structure $(A^n)_{1\leq n\leq N}$ and a latent clustering structure $(C(x))_{x\in\mathcal{X}}$ are introduced. In order to enforce a hierarchical organization of aspects as intended by the given tree, these structures are related by the following compatibility constraints:

$$P(a|x,c)\equiv\mathrm{P}\{A^n=a|X^n=x,C(x)=c\}=0,\quad\text{if not } a\uparrow c\ . \tag{33}$$

The symbolic expression $a\uparrow c$ reads "a is on the path from the root to c". By comparing these constraints with the ones imposed in the one-sided clustering model in (12), it immediately follows that the latter reduces to a "degenerated" hierarchy without inner nodes. In the one-sided clustering model, all observations involving a particular x must have the same aspect, in the hierarchical model, this condition is weakened such that all observations must have aspects which are associated with a single path in the tree. The graphical model representation of the hierarchical clustering model thus is identical with the one-sided clustering model, the only difference is that the constraints are weakened such that the clustering structure does not completely determine the aspect variables. Hence, we have the freedom to introduce additional parameters $P(a|x,c)$ for $a\uparrow c$ (cf. Figure 2 (b)).[5] In addition to this fully parameterized model, the more parsimonious model $P(a|x,c)=P(a|c)$ and the entropy maximizing model without additional parameters $P(a|x,c)=1/depth(c)$ provide interesting alternatives.[6]

[4] The EM approach for the two-sided clustering model often converges to an inferior local minimum when started from a random initialization, because all posteriors typically approach uniform distributions. As a remedy, a solution of the one-sided clustering model is utilized to initialize one set of latent variables.

[5] In the sequel, we implicitly assume that $P(a|x,c)=0$ if not $a\uparrow c$.

[6] With the parameterization by $P(a|c)$ the model becomes non-identifiable and degenerates to the one-sided clustering model, however, we suggest to fit these mixing proportions that weight different abstraction levels in the hierarchy from held-out data.

By summing over the hidden variables one obtains the following nested mixture formulation of the hierarchical model: $P(\mathcal{S}) = \prod_{x \in \mathcal{X}} P(\mathcal{S}_x)$, where

$$P(\mathcal{S}_x) \quad = \quad \sum_{c \in \mathcal{C}} P(c) \prod_{y \in \mathcal{Y}} \left[\sum_{a \in \mathcal{A}} P(y|a) P(a|x,c) P(x) \right]^{n(x,y)} \quad . \tag{34}$$

In comparing (34) with the "flat" clustering model, one may notice that the class-conditional probability $P(y|a)$ has been replaced by an inner mixture which is a mixture along the "vertical" axis of the hierarchy.

Expectation Maximization Algorithm The E-step for the hierarchical model can best be performed in a two stage manner. From the mixture formulation it is straightforward to generalize the posterior probability of the one-sided clustering model,

$$\mathrm{P}\{C(x)=c|\mathcal{S},\theta\} \quad \propto \quad P(c) \prod_{y \in \mathcal{Y}} \left[\sum_{a \in \mathcal{A}} P(y|a) P(a|x,c) \right]^{n(x,y)} \quad . \tag{35}$$

Because of the constraints imposed by the clustering variables, posteriors of aspect variables have to be conditioned on values of the former, i.e., one has to compute

$$\mathrm{P}\{A^n=a|x,y,C(x)=c;\theta\} = \frac{P(a|x,c)P(y|a)}{\sum_{a' \in \mathcal{A}} P(a'|x,c)P(y|a')} \quad . \tag{36}$$

The derivation of the M-step equations follows the standard procedure which yields

$$P(y|a) \quad \propto \quad \sum_{n:y^n=y} \mathrm{P}\{A^n=a|x,y;\theta\} \; , \tag{37}$$

$$P(a|x,c) \quad \propto \quad \sum_{n:x^n=x} \mathrm{P}\{A^n=a|x,y,C(x)=c;\theta\} \quad , \tag{38}$$

$$P(c) \quad \propto \quad \sum_{x \in \mathcal{X}} \mathrm{P}\{C(x)=c|\mathcal{S},\theta\} \quad , \tag{39}$$

completing the derivation of the EM algorithm.

2.5 Discussion

At this point it might be helpful to give a systematic overview of the presented models. It has been stressed how different models along with their parameterization can be derived by imposing constraints on the basic aspect model. As a consequence, the different models can be systematically arranged in a simple Venn diagram (cf. Figure 4), where the partial order distinguishes more and less constrained models. Obviously, the aspect model is the least constrained model, while the two-sided clustering model possesses the most strongly restricted latent aspect space.

Another way to compare the different models is by their predictive probabilities on new data, e.g., $P(y|x,\mathcal{S})$, which is summarized in Table 1. In regard of Table 1, we would like to make the following comments:

- There is an essential difference between parameters $P(\cdot)$ and posterior probabilities $\mathrm{P}\{\cdot\}$ for latent variables. For example, in the one-sided clustering model, the x-specific distribution $P(a|x)$ over aspects is replaced by the posterior probability $\mathrm{P}\{C(x)=c|\mathcal{S}\}$ of x having latent class c; the latter will asymptotically ($n(x) \to \infty$) approach Boolean values which is not necessarily true for the parameters in the aspect model. While the mixing weights $P(a|x)$ are free

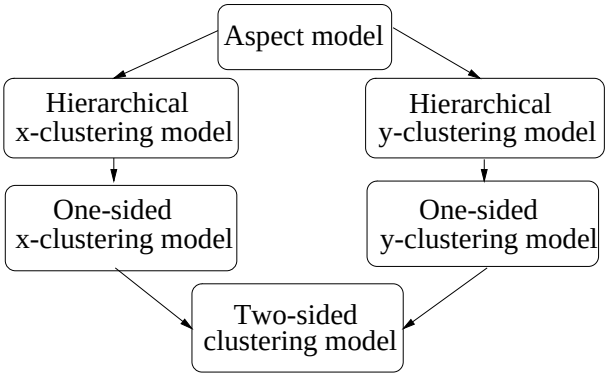

Figure 4: Venn diagram for the family of co-occurrence latent variable models.

parameters in the aspect model, the corresponding weights in the clustering models are only induced by the posterior uncertainty and would vanish, if the values of the latent variables were known.

- This has also consequences in interpreting the class-conditional distributions $P(y|a)$ and $P(y|c)$. In the aspect model, these do not have to be prototypical for any particular group of objects from $\mathcal{X}$, since $P(y|x)$ is obtained by convex combinations of the class-conditionals. In contrast, the one-sided clustering model attempts to identify class-conditionals which are characteristic for a group of objects in $\mathcal{X}$.
- In the hierarchical generalization of the "flat" clustering model the class conditional $P(y|c)$ is replaced by a more flexible weighted combination. On the other hand, the class-conditionals $P(y|c)$ of the one-sided clustering model are further restricted in the two-sided clustering model to be proportional to the marginal probability $P(y)$ times a cluster-specific modulation factor which depends on y only through its associated class. The latter reflects the additional "coarsening" induced by grouping objects in $\mathcal{Y}$.

2.6 Related Models and Previous Work

Brown et al. [BdM$^+$92] have proposed an information criterion for class–based n–gram models in language modeling. Word classes are formed such that the mutual information between classes of adjacent words is maximized. Their model is closely related to the two-sided clustering model with the main difference that in [BdM$^+$92] the word classes for the predicting and predicted word are identified (hence $C(x)$=$D(x)$).[7] Moreover, Brown et al. have proposed a non–probabilistic, hard clustering variant for which they have introduced a greedy cluster merging algorithm.

Many of the ideas presented in his paper have been solicitated by the pioneering work of Pereira, Tishby, and Lee [PTL93] on *distributional clustering*. In [PTL93], the aspect model mixture distribution is the starting point, however, neither are latent aspect variables introduced, nor is the corresponding EM algorithm derived. Instead, Pereira et al. propose a clustering method similar to the one-sided clustering model, which is based upon two different principles: The posterior class distributions $P(c|x)$ are obtained by minimizing a cross–entropy criterion (or, equivalently, Kullback–Leibler divergence) together with an entropic regularization with parameter β, which results in maximum entropy cluster memberships $P(c|x) \propto \exp\left[\beta \sum_{y \in \mathcal{Y}} \hat{P}(y|x) \log P(y|c)\right]$. For $\beta = 1$

[7] Whether this identification is advantageous in terms of the predictive log–likelihood is, at least, doubtful. Algorithmically, this identification poses additional problems, since the alternation scheme proposed for the two-sided clustering model no longer applies.

Model	$P(y\|x,\mathcal{S})$
Aspect	$\sum_a P(a\|x)P(y\|a)$
One-sided Clustering	$\sum_c \mathrm{P}\{C(x)=c\|\mathcal{S}\}P(y\|c)$
Hierarchical Clustering	$\sum_c \mathrm{P}\{C(x)=c\|\mathcal{S}\}\sum_a P(a\|x,c)P(y\|a)$
Two-sided Clustering	$\sum_c \mathrm{P}\{C(x)=c\|\mathcal{S}\}P(y)\sum_d \mathrm{P}\{D(y)=d\|\mathcal{S}\}\,\phi(c,d)$

Table 1: Systematic overview of the presented dyadic mixture models.

this differ from the E–step equations for the one–sided clustering model in (18) by a (missing) factor of $n(x)$ in the exponent and the class prior probabilities $P(c)$. The reason is that all x objects are treated equally, in particular irrespective of the number of observations in the sample set $\mathcal{S}_x$. Yet in [PTL93], the centroid condition is derived from the maximum likelihood principle, after inserting the maximum entropy cluster membership probabilities. Although both principles are considered to be "complementary" and re-estimation equations are iterated according to a pseudo–EM scheme. In [PTL93] it remains unresolved whether an underlying common objective function exists. In a new formulation [TP98], both types of equations have been derived more rigorously in a rate–distrotion theoretic ansatz from a mutual information principle. The latter might offer perspectives for a foundation of unsupervised learning beyond the maximum likelihood principle, which has been the basis of our work.

Saul and Pereira [SP97a] have – independently from our work – proposed a model called *aggregate Markov model* which they have utilized as a back-off model [Kat87] in the context of language modeling. Their approach is equivalent to the aspect model in its asymmetric parameterization, but was restricted to a class-based bigram model ($\mathcal{X} = \mathcal{Y}$). They have also derived an EM algorithm, but without employing annealing methods (cf. Section 3.1).

3 Generalizations and Extensions

3.1 Annealed EM

Annealed EM is an important generalization of EM-based model fitting which pursues three main goals:

- Avoiding overfitting by controlling the effective model complexity,
- reducing the sensitivity of EM to local maxima, and
- generating tree topologies for the hierarchical clustering model.

Annealed EM is closely related to deterministic annealing, a technique that has been applied to many clustering problems, including vectorial clustering [RGF90, RGF92, BK93], pairwise clustering [HB97, PHB99], and in the context of co-occurrence data for distributional clustering [PTL93]. The key idea is to introduce a temperature parameter T and to replace the minimization of a combinatorial objective function by a substitute known as the *free energy*. Here, we present annealing methods without reference to statistical physics. Consider therefore the general case of maximum likelihood estimation by the EM algorithm. The E-step by definition computes a posterior average of the complete data log-likelihood which is maximized in the M-step. The *annealed* E-step at temperature T performs this average w.r.t. a distribution which is obtained by generalizing Bayes' formula such that the likelihood contribution is taken to the power of $1/T$, i.e., in mnemonic notation: $annealed\text{–}posterior \sim prior \times likelihood^{1/T}$. For $T > 1$ this amounts to increasing the influence of the prior which in turn results in a larger entropy of the (annealed) posteriors.

For example, in the case of the aspect model and the one-sided clustering model, the annealed E-steps generalizing (5,15) are given by

$$P(a|x,y,\theta\} \propto P(a)\left[P(x|a)P(y|a)\right]^{1/T} \quad \text{and} \tag{40}$$

$$P(c|\mathcal{S}_x,\theta\} \propto P(c)\prod_{y\in\mathcal{Y}} P(y|c)^{n(x,y)/T} \ , \tag{41}$$

respectively. For fixed $T > 1$ the *annealed* E-step performs a regularization based on entropy. This is the reason why annealed EM not only reduces the sensitivity to local minima but also controls the effective model complexity. Annealing, thereby, has the potential to improve the generalization for otherwise overfitting models (for supervised learning problems cf. [PKM96, RMRG97]). Recent theoretical investigations emphasize the benefits of annealing to avoid overfitting phenomena [Buh98]. In this paper, the advantages of deterministic annealing are investigated experimentally (cf. Section 4).

In statistical learning, deterministic annealing is used in the $T \rightarrow T_{fin} \geq 1$ limit where the stopping temperature $T_{fin} \rightarrow 1$ in the inifinte data aspymptotics. In hard clustering applications one usually perform annealing in the limit $T \rightarrow 0$. Although there is no theoretical guarantee that deterministic annealing finds the global minimum in general, many independent empirical studies indicate that the typical solutions obtained are often significantly better than the corresponding 'unannealed' optimization. This is explained by the fact that annealing is a *homotopy method*, where the original objective function (such as the log-likelihood) is smoothed for large T.

For hierarchical models, the annealed EM algorithm offers a natural way to generate tree topologies. As is known from adaptive vector quantization [RGF90], starting at a high value of T and successively lowering T typically leads through a sequence of phase transitions. At each phase transition, the effective number of distinguishable clusters grows until some maximal number is reached or the annealing is stopped. This suggests a heuristic procedure where we start with a single cluster and recursively split clusters. In the course of the annealing, the sequence of splits is tracked. The respective 'phase diagram' is used as a tree topology. Note that merely the tree topology is successively grown, while the parameters and the posterior probabilities of the latent classes may drastically change during the annealing process.

3.2 Expectation Maximization with Cross–Validation

Data sparseness is pervasive in the analysis of dyadic data. It is an an often neglected problem of maximum likelihood estimation – which it shares with all empirical risk minimization methods – the maximization of the likelihood based on the training data does not automatically guarantee an equivalent performance on unseen new data. This effect is known as *overfitting* and is often severe for sparse data.

Cross validation is the conceptually simplest criterion for model assessment: the model is trained on one part of the available data and tested on the remaining observations. Then, this procedure is repeated with a different data split. The average predictive log-likelihood may then serve as a model score. Although cross-validation is of great use in model selection, it does not directly indicate how to modify the model fitting procedure in order to obtain a better generalization performance. *Early stopping* is a common way to utilize cross–validation in model fitting: an iterative fitting method is aborted, before the test error performance starts to degrade. This, of course, makes specific assumptions about the nature of the fitting method. Moreover, there is no indication that early stopping is optimal in any sense to circumvent overfitting phenomena.

In the context of the discussed EM procedures, a simple modification leads to an efficient (one-step) approximation of leave-one-out cross-validation. We propose to modify the E-step such, that the parameter estimates θ used in computing posterior probabilities are corrected by subtracting the influence of the observations related to the latent variable in question. To be more precise,

consider the E-step of the basic aspect model. Neglecting the n–th observation associated with a latent variable A^n leads to the modified equations

$$P(a|x,y,\theta^{-n}) \quad \propto \quad P^{-n}(a)P^{-n}(x^n|a)P^{-n}(y^n|a) \ . \tag{42}$$

Here P^{-n} denote modified M-step estimates obtained by

$$P^{-n}(a) \quad \propto \quad N \cdot P(a) - P(a|x,y,\theta^{-n}) \ , \tag{43}$$

$$P^{-n}(y|a) \quad \propto \quad N \cdot P(a) \cdot P(y|a) - P(a|x,y,\theta^{-n}) \ , \tag{44}$$

$$P^{-n}(a) \quad \propto \quad N \cdot P(a) \cdot P(x|a) - P(a|x,y,\theta^{-n}) \ , \tag{45}$$

where the occurring posterior probabilities are the estimates computed in the preceding E-step. Essentially, this amounts to removing the (direct) influence of the n–th observation on the last M-step parameter estimation. Since EM is an iterative procedure, there is of course a remaining indirect influence which may still lead to over-fitting problems.

In the one-sided clustering model all observations $\mathcal{S}_x$ are removed in estimating the posterior for $C(x)$, i.e.,

$$P^{-x}(c) \quad = \quad \frac{1}{I-1} \sum_{x' \neq x} \mathrm{P}\{C(x')=c|\mathcal{S},\theta^{-x'}\} \ , \tag{46}$$

$$P^{-x}(y|c) \quad \propto \quad \sum_{x' \neq x} \mathrm{P}\{C(x')=c|\mathcal{S},\theta^{-x'}\} n(x',y) \ . \tag{47}$$

An efficient computation of these corrections based on additional bookkeeping quantities is straightforward. The interpretation of the leave-one-out E-step is very intuitive: for example, in the one-sided clustering model, posterior probabilities $\mathrm{P}\{C(x)=c|\mathcal{S},\theta^{-x}\}$ are calculated based on how well the averaged statistics of the other objects x' assigned to a particular cluster c predict the observations which involve x.

In order to preserve strict optimization principles, the modification of the E-step has to be carried out slightly more careful than by just eliminating certain observations (cf. Appendix). However, we have found empirically that the naive corrections discussed above do typically not lead to convergence or stability problems. We have, therefore, used the simpler equations in our experiments.

3.3 Accelerated EM

EM algorithms have important advantages over gradient-based methods. However, for many problems the convergence speed of EM restricts its applicability to large data sets. A simple way to accelerate EM algorithms is by *over-relaxation* in the M-step. This has been discussed in the context of mixture models [PW78] and was recently 'rediscovered' under the title of $EM(\eta)$ in [BKS97]. We found this method useful in accelerating the fitting procedure for all discussed models. Essentially the estimator for a generic parameter θ in the M-step is modified by

$$\hat{\theta}^{(t+1)} \quad = \quad (1-\eta)\hat{\theta}^{(t)} + \eta \overline{\theta}^{(t+1)} \ , \tag{48}$$

where $\overline{\theta}^{(t+1)}$ is the M-step estimate, i.e., $\eta = 1$ is the usual M-step. Choosing $1 < \eta < 2$ still guarantees convergence, and typically $\eta \approx 1.8$ has been found to be a good choice to speed up convergence. In case that a positivity or normalization constraint is violated after performing an over-relaxed M-step, the parameter vector is projected back on the admissible parameter space (replacing negative probabilities by a small positive constant). For an overview on more elaborated acceleration methods for EM we refer to [MK97].

	Aspect		$\mathcal{X}$-cluster		Hierarchical		$\mathcal{X}/\mathcal{Y}$-cluster	
K	$1/T$	$\mathcal{P}$	$1/T$	$\mathcal{P}$	$1/T$	$\mathcal{P}$	$1/T$	$\mathcal{P}$
1	-	685	-	-	-	-	-	-
16	0.85	431	0.07	482	0.14	471	0.60	543
32	0.83	386	0.07	**452**	0.12	438	0.53	506
64	0.79	360	0.06	527	0.11	422	0.48	477
128	0.78	**353**	0.04	663	0.10	**410**	0.45	**462**

Table 2: Comparative results for context–dependent unigram modelling for all discussed models on the Cranfield IR test collection (Cranfield, I=1400, J=1664, N=111803). All results are based on ten–fold cross validation.

3.4 Multiscale EM

Multiscale optimization [HPB94, PB98] is an approach for accelerating clustering algorithms whenever a neighborhood structure exists on the object space(s). In image segmentation, for example, it is a natural assumption that adjacent image sites belong with high probability to the same cluster or image segment. This fact can be exploited to significantly accelerate the estimation process by maximizing over a suitable nested sequence of variable subspaces in a coarse–to–fine manner. This is achieved by temporarily tying adjacent sites in a joint assignment variable. For notational convenience we again restrict the presentation to the one-sided clustering model, while extensions to the two-sided clustering are straightforward.[8]

More formally a coarsening hierarchy for $\mathcal{X}$ is given by a nested sequence of equivalence relations $\mathcal{M}^{(l)}$ over $\mathcal{X}$, where $\mathcal{M}^{(l)} \subset \mathcal{M}^{(l+1)}$ and $\mathcal{M}^{(0)} = \{(x,x) : x \in \mathcal{X}\}$. In the context of image analysis these equivalence relations typically correspond to multi–resolution pixel grids obtained by subsampling. The log–likelihood is minimized at coarsening level l by imposing constraints of the form $C(x) = C(x')$ whenever $(x,x') \in \mathcal{M}^{(l)}$. The coarse scale optimization at level l thus simply yields a further constrained aspect model. The computational advantage is a reduced number of posterior computations in the E-step which has only to be performed once for each equivalence class. After the maximization procedure at a resolution level l is converged, the optimization proceeds at the next level $l-1$ by prolongating the found solution in $\mathcal{M}^{(l)}$ to the subset defined by $\mathcal{M}^{(l-1)}$, thus initializing the optimization at level $l-1$ with the solution at level l.

We like to emphasize that in contrast to most multiresolution optimization schemes, multiscale optimization has the advantage to maximize the *original* log–likelihood at all resolution levels. It is only the set of hidden variables which is effectively reduced by imposing the constraints on the set of hidden variables $\mathcal{M}^{(l)}$. We applied multiscale optimization in all image analysis experiments resulting in typical accelerations by factors 10–100 compared to single–level optimization.

4 Applications and Experimental Results

This section presents four different application domains and discusses the applicability of statistical models for dyadic data: (i) information retrieval, (ii) data mining in text databases, (iii) computational linguistics, and (iv) image segmentation in computer vision. Given the conception of this paper, namely to present a largely domain-independent theory of unsupervised learning from dyadic data, we will not attempt to perform a thorough and detailed performance comparison with state of the art techniques on all of these problems; accompanying and forthcoming publications will deal with the different applications separately. We will, however, show in sufficient detail, how the different models can be applied to realistic problems and report some representative results on real-world

[8] Multiscale optimization in its current form is not applicable to the aspect model.

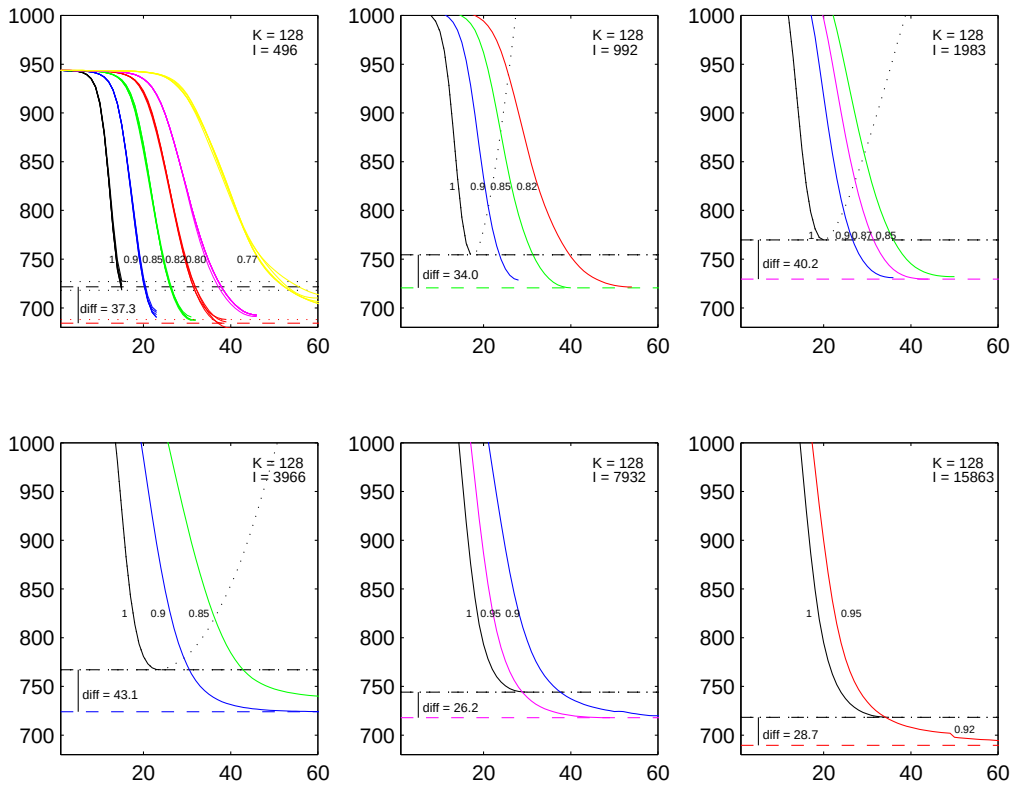

Figure 5: Annealing experiments on the TDT1 dataset for a model with $K = 128$ components at different subsampling factors (from upper left to lower right: 32x,16x,8x,4x,2x,1x). I denotes the number of documents in the training/held out data set, samll numbers indicate the inverse temperatur $1/T$ utilized for the respective training curves.

data. In addition, the different models are evaluated on data from different domains and the variants and extensions of EM discussed in Section 3 are systematically compared. Our main goal is thus to demonstrate the relevance of dyadic data models for many important applications and to investigate questions of modeling and model fitting across domains.

4.1 Information Retrieval

Intelligent information retrieval in databases is one of the key topics in *data mining*. The problem is most severe in cases where the query cannot be formulated precisely, e.g., in natural language interfaces for document collections and digital libraries or in image and multi-media databases. Typically, one would like to obtain those entries (documents, images, etc.) from a database which best match a given query according to some similarity measure. Yet, it is often difficult to reliably estimate similarities, because the query may not contain enough information, e.g., not all possibly relevant keywords might occur in a query for documents.

Context–Dependent Unigram Models In a first series of experiments, we have investigated the general question of how well the different mixture models perform in predicting the occurrences of words $y \in \mathcal{Y}$ in the context of a particular document $x \in \mathcal{X}$. The latter can be thought of as a

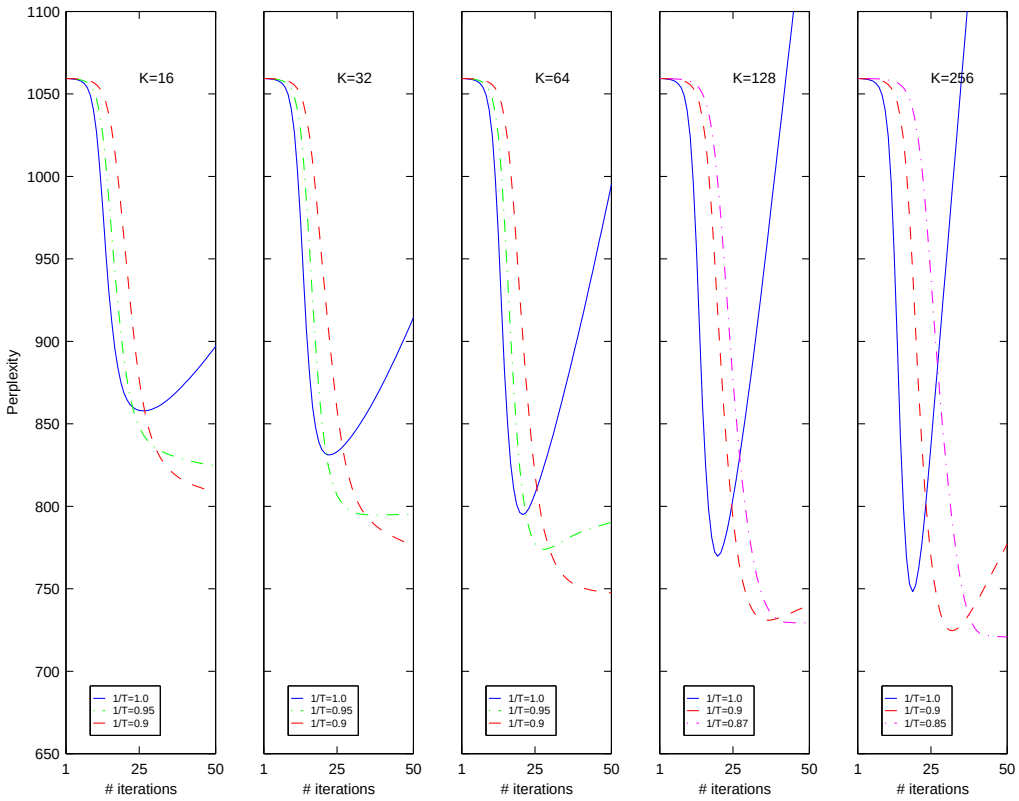

Figure 6: Annealing experiments on the TDT1 documents (subsampling 8x) with different number of components K and for different inverse temperatures $1/T$.

context–dependent unigram language model. Following the standard procedure in statistical learning, the set of all word occurrences has been divided into a training, a validation and a test set. From a statistical point of view the canonical goodness-of-fit measure is the average log–likelihood on the test set. In the context of natural language processing it is, however, more customary to report the perplexity $\mathcal{P}$ which is related to the average (per pattern) test set log–likelihood l by $\mathcal{P} = \exp(-l)$. The validation set was used to determine the optimal choice of the computational temperature in the annealed EM algorithm.

We have used a standard IR collection with 1400 documents from mechanical engineering (Cranfield) to evaluate perplexity results. As a preprocessing step very infrequent and very common words (stop word list) have been eliminated, moreover the Porter stemmer was used the generate word stem. The results for different (maximal) number of components K are summarized in Table 1.

The main conclusions are:

- The lowest perplexity – with a perplexity reduction of almost 50% compared to the unigram model – is obtained with the aspect model. The hierarchical clustering model performs better than the more constrained flat clustering models. Hence, in terms of perplexity the least constraints mixture models should be preferred over the clustering models.

- The optimal temperature for the aspect model is consistently above $T = 1$ which is the standard EM algorithm. For the clustering models the optimal generalization performance even requires a much higher temperature.

Method	Recall 10%	30%	50%	70%	90%
Medline, 30 queries					
TF	72.7	56.7	44.6	32.4	15.3
TFIDF	71.4	64.4	49.7	37.4	17.8
ML ($K = 32$, $\lambda = 0.5$)	76.3	62.0	49.6	37.1	21.1
ML ($K = 64$, $\lambda = 0.5$)	70.6	60.3	48.1	34.9	20.6
AML ($K = 256$, $\lambda = 0.7$)	78.4	69.2	58.4	45.9	32.0
AML (...) + TDIDF	79.3	69.9	63.4	52.2	33.0
abs. impr. vs. baseline	+6.6	+13.2	+18.2	+19.8	+17.7
rel. impr. vs. baseline	+9.1	+23.3	+36.1	+61.1	+115.7
Cranfield, 225 queries					
TF	64.4	39.4	29.0	13.2	7.7
TFIDF	69.9	48.0	34.9	18.1	9.2
AML ($K = 256$, $\lambda = 0.5$)	66.0	42.2	35.0	16.1	10.6
AML (...) + TFIDF	70.5	48.8	36.9	22.2	14.1
abs. impr. vs. baseline	+9.4	+13.9	+12.3	+13.0	+9.2
rel. impr. vs. baseline	+14.6	+35.3	+42.4	+98.5	+119.5

Table 3: Retrieval precision results in percent evaluated on the Medline and Cranfield collection by macro averaging. TDF and TDIDF denote the different term weighting schemes, ML and AML refer to an aspect model trained with EM (maximum likelihood) and annealed EM (annealed maximum likelihood).

- Temperature–based complexity control clearly does much better than restricting the number K of components. Even the aspect model with $K = 8$ components suffers from overfitting, if trained with non–annealed EM.

To stress the advantages of annealed EM, we have investigated the effect of a temperature–based regularization in more detail. In order to take the sample set size into account, we have utilized the Topic Detection and Tracking (TDT1) corpus [LDC97] which consists of approximately 7 million words in 15863 documents and which is large enough for subsampling experiments. Since we want to focus on the control of overfitting and not on the problem of local maxima, all models have been trained at a fixed temperature. Figure 5 shows perplexity curves for different inverse temperatures $1/T$ as a function of the number of annealed EM iterations. At all temperatures we have performed early stopping once the perplexity on held-out data increased. For the 32x subsampling experiments ($I = 496$ documents), we have repeated runs 5 times in order to evaluate the solution variability for different (randomized) initial conditions. To further investiagte the effect of the model size, we have also varied the number of components K at the subsampling level 8x. The resulting curves are depicted in Figure 6.

The following observations can be made:

- The use of a temperature $T > 1$ yields a significant and consistent improvement for all sample sizes (cf. Figure 5) and for models with a largely varying number of parameters (cf. 6).
- Although it has to be expected that the benefits of annealing are diminishing for larger data sets and may even vanish in the infinite data limit, the effect of a $T > 1$ temperature is still considerable, even for the full 7 million word TDT1 dataset. Thus overfitting is also an important problem in large-scale data sets.
- Early stopping is quite successful to prevent overfitting. The experiments indicate that it may

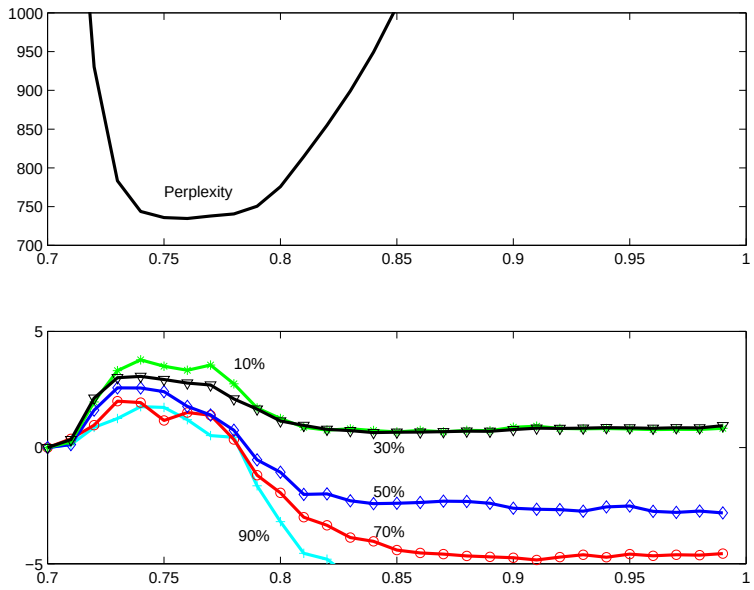

Figure 7: Model performance $K = 256$ on the Cranfield collection in terms of perplexity and precision (absolute gain vs. baseline) at different inverse temperatures $1/T$.

in fact take advantage of models with a huge number of parameters. In Figure 6 one observes that although overfitting gets more and more severe with increasing K (the training curves get steeper), the perplexity at the respective minima is nevertheless decreasing. However, compared to an annealed training one needs approximately 4 times more parameters to achieve the same performance. For example the annealed solution at $K = 32/64$ is of the same quality as the one obtained at $K = 128/256$ with early stopping EM.

- The computational complexity of annealed EM at a fixed K is slightly higher compared to standard EM, typically twice as many iterations are necessary to converge. Yet, in the above experiments it achieves a prespecified performance with roughly 50% of the computing time and 25% of the memory requirements of standard EM. This is because an iteration for a model with $2K$ components is roughly twice as expensive as an iteration for K components.

Compared to the effect of annealing, the predictive EM variant showed only a slight though consistent improvement for the aspect model. Given the additional computations necessary to calculate the corrections, the practical use of predictive EM is somewhat limited for the aspect model. Yet, predictive EM has shown to have a stronger effect in combination with clustering models, which has to be expected, since the correction terms are more significant in this case.

The overrelaxed EM variant has also proven to be a valuable tool in our simulations with a typical acceleration factor of $2-3$. Overrelaxation helps in particular to accelerate the convergence process of annealed EM and, of course, less important in combination with early stopped EM training.

Probabilistic Latent Semantic Indexing Beyond controversy, the most popular family of techniques utilized in information retrieval is the so-called *Vector Space Model* (VSM), introduced by Salton et al. [SM83, SB91, BAS95] in the SMART system. In the VSM, each document is represented by a term vector with (transformed) frequency counts for term occurrences as components. The two most important ingredients of the VSM are: (i) a function to measure similarity between documents, typically the cosine between the two vector representations, (ii) a term weighting scheme to re-weight

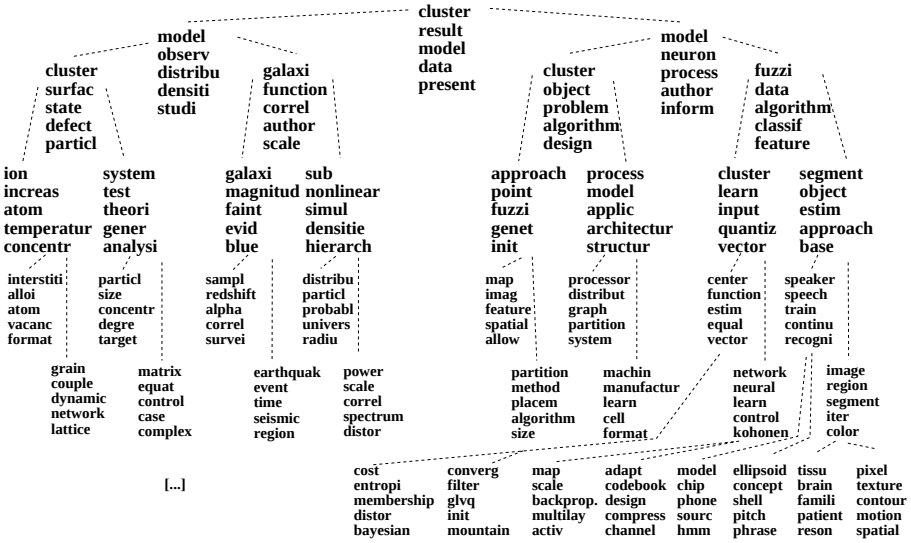

Figure 8: Upper levels of the CLUSTER hierarchy. Each node is described by the most probable terms of the conditional probabilities $P(y|a)$.

the influence of different terms. One very successfully applied weighting scheme is the TFIDF (term frequency inverse document frequency) which transforms counts by $\tilde{n}(x,y) = -\log(f_y) \cdot n(x,y)$, where f_y denotes the fraction of documents which contain the term y. The similarity between a document x and a query x' (or a second document) is then computed by

$$S(x,x') = \frac{\sum_y \tilde{n}(x,y)\tilde{n}(x',y)}{\sqrt{\left(\sum_y \tilde{n}(x,y)^2\right)\left(\sum_y \tilde{n}(x',y)^2\right)}} \,. \tag{49}$$

Since this angular similarity measure is scale invariant, we may as well replace the counts $n(x,y)$ by the empirical probabilities $\hat{P}(y|x) = n(x,y)/n(x)$.

The key idea in Probabilistic Latent Semantic Indexing (PLSI) is to replace empirical conditionals by multinomial word distributions derived from the aspect model. Namely, we consider linear interpolations between the empirical probabilities and the conditional distributions derived from the model

$$P_\lambda(y|x) = (1-\lambda)\hat{P}(y|x) + \lambda \sum_a P(y|a)P(a|x) \,. \tag{50}$$

Intuitively this aims at exploiting semantic relations extracted by the aspect model, e.g., about synonyms and words with similar meanings, to get improved matches. For example, although a word y in a query may not occur in a document, the (indirect) conditional probability based on the aspect model can nevertheless be high. This is similar to the dimension–reduction approach pursued in standard LSI [DTGL90].

We tested the PLSI method on a number of medium–sized standard document test collection with relevance assessments by computing *precision–recall curves*. The precision–recall curve reported in the sequel have been obtained by so–called macro averaging (cf. [VR75, Chapter7]). For each query q we determine n–best lists from which precision/recall pairs are computed for every n. A single precision/recall pair is obtained as follows: If R_q is the total number of relevant documents for a given query q and $r_q(n)$ is the number of relevant documents in the n–best list, then precision and

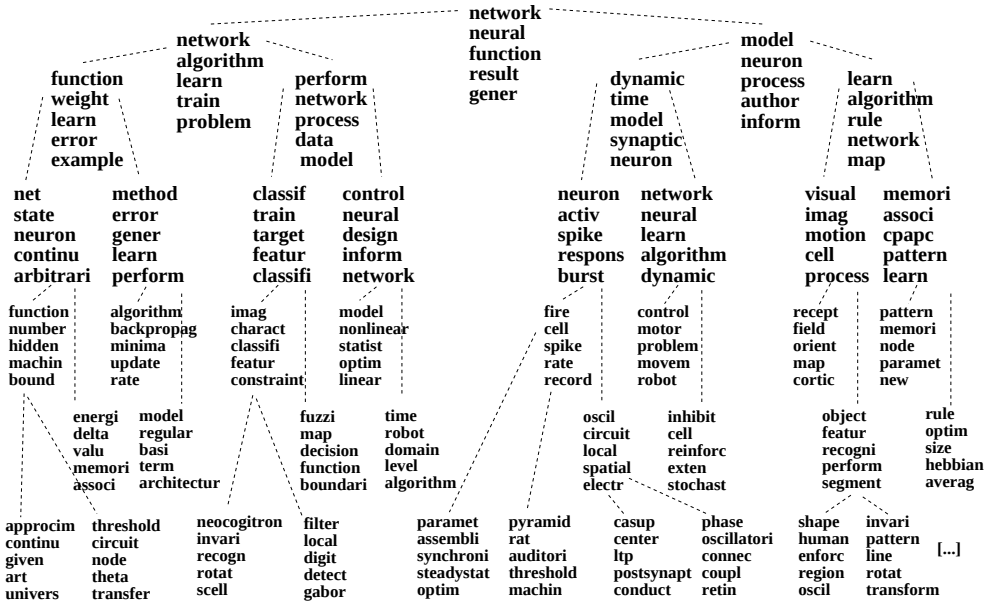

Figure 9: Upper levels of the NEURAL hierarchy. Each node is described by the most probable terms of the conditional probabilities $P(y|a)$.

recall at n are defined as $\mathrm{Prec}_q(n) = \frac{r_q(n)}{n}$ and $\mathrm{Rec}_q(n) = \frac{r_q(n)}{R_q}$. In Table 3, we have summarized the results for the Medline collection ($I = 1033$) and the Cranfield ($I = 1400$) collection, with and without the inverse document frequency term weighting and for models obtained by EM and annealed EM training. The main conclusions we can draw are: (i) The use of the aspect model to estimate document–specific word distribution yields substantial improvements in terms of retrieval precision. In particular, in the regime of high recall, the relative gain is more than 100%, e.g., the average number of irrelevant documents returned at the Rec = 90% level is more than halved. (ii) In many experiments, we have observed that improvements in word perplexity and retrieval precision are strongly correlated. This is stressed in Table 3 by comparing overfitted models trained according to the likelihood criterion with perplexity optimized models trained on the annealed likelihood. An exemplary run of annealed EM beyond the optimal stopping temperature which compares the simultaneous development of word perplexity and precision curves is shown in Figure 7. It can be seen that it is crucial to control the generalization performance of the model, since the precision is inversely correlated with the perplexity. In particular, notice that the model obtained by maximum likelihood estimation (at $T = 1$) actually deteriorates the retrieval performance.

4.2 Data Mining in Text Databases

Structuring and visualizing large databases is one of the key topics in *data mining*. Applications range from interactive and coarse-to-fine information access to efficient data management and representation. In this paragraph, we will focus on an application of the hierarchical clustering model for structuring large text repositories. Clustering of documents provides a popular way of pre-structuring a database that has been applied with mixed success in the context of query–based retrieval (cf. [Wil88] for an overview), but is of great importance in interactive retrieval. The most frequently used methods in this context are linkage algorithms (single linkage, complete linkage, Wards method, cf. [JD88]), or hybrid combinations of agglomerative and centroid–based methods

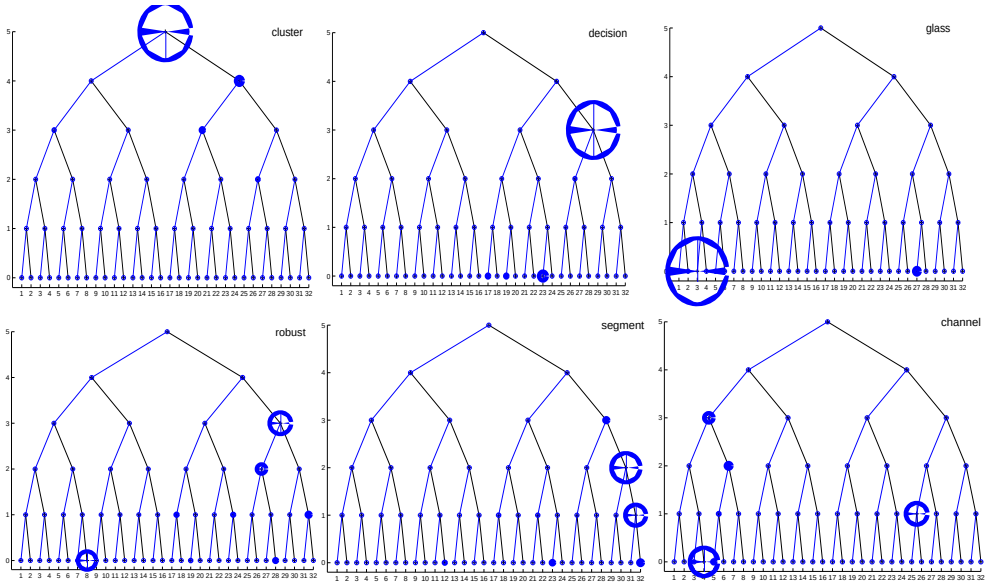

Figure 10: Exemplary relative word distributions over nodes for the *CLUSTER* dataset for the keywords 'cluster', 'decision', 'glass', 'robust', 'segment', and 'channel'.

[CKP92] which have no probabilistic interpretation and have a number of other disadvantages. In contrast, the hierarchical mixture model provide a sound statistical basis and also has many additional features which make it a suitable candidate in this context.

We have performed experiments for information retrieval on different collections of abstracts. To facilitate the assessment of the extracted *structure* we have investigated a dataset of $N = 1584$ documents containing abstracts of papers with *clustering* as a title word (CLUSTER) and a second dataset of 1278 documents with abstracts from the journals *Neural Computation* and *Neural Networks* (NEURAL). This data is presumably more amenable to an interpretation by the reader than are the standard text collections. The top–levels of a cluster hierarchy for CLUSTER and NEURAL generated by the hierarchical clustering model are visualized in Figure 8 and Figure 9, respectively.

The overall hierarchical organization of the documents is very satisfying, the topological relations between clusters seems to capture important aspects of the inter-document similarities. In contrast to most multi–resolution approaches the distributions at inner nodes of the hierarchy are not obtained by a coarsening procedure which typically performs some sort of averaging over the respective subtree of the hierarchy. The abstraction mechanism in fact leads to a specialization of the inner nodes. This specialization effect makes the probabilities $P(y|a)$ suitable for *cluster summarization*. Notice, how the low–level nodes capture the specific vocabulary of the documents associated with clusters in the subtree below. The specific terms become automatically the most probable words in the component distribution, because higher level nodes account for more general terms.

To further demonstrate the ability of the hierarchical clustering model to identify abstraction levels in the hierarchy, we have visualized the distribution of the responsibility for observations involving the same index word y for some particularly interesting examples in Figure 10. The first tree for the word 'cluster' shows that, as expected, the occurrences of 'cluster' in documents are explained to be a common feature of all documents, hence most of the occurrences are assigned to the root. The word 'decision' is found on a level 3 node, indicating that it is a typical word for all algorithmically oriented documents assigned to nodes in the subtree, but e.g. not for the left branch of papers from physics and astronomy. The index term 'robust' occurs in two different meanings:

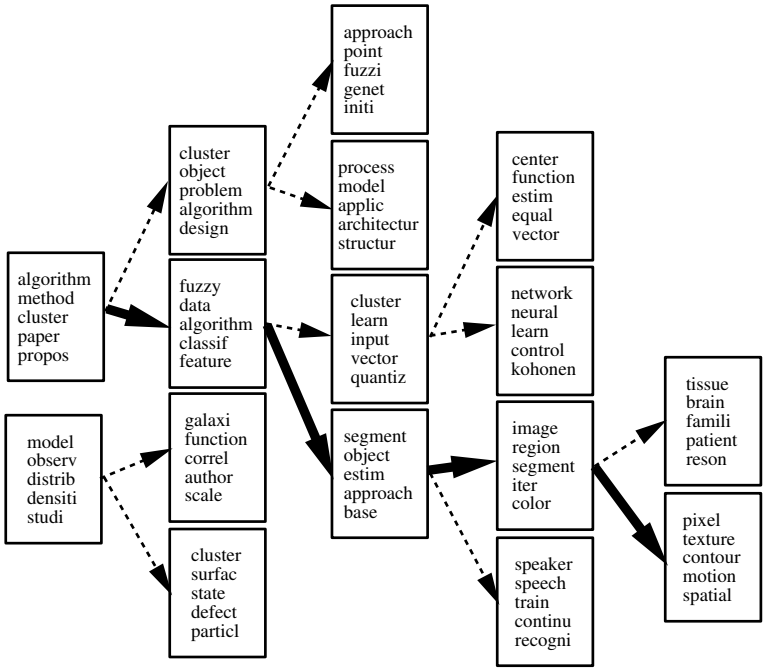

Figure 11: Example run of an interactive retrieval session for documents on 'texture–based image segmentation' with one level look-ahead in the cluster hierarchy.

first, it has a highly specific meaning in the context of stability analysis (an aspect characterized by the terms 'plane', 'perturb', 'eigenvalue', 'root', etc.) and a rather broad meaning in the sense of robust methods and algorithms. The word 'segment' occurs mainly in documents about computer vision and language processing, but it is used to a significant larger extend in the first field. 'glass' is a specific term in solid state physics, it thus is found on the lowest level of the hierarchy. 'channel' is again ambivalent, it is used in the context of physics as well as in communication theory. The bimodal distribution clearly captures this fact.

These examples are only spotlights, but they demonstrate that the extracted hierarchical organization reflects interesting structure inherent in the co-occurrence data. To demonstrate the usefulness at least for one example, we have depicted a virtual interactive expansion of the CLUSTER hierarchy for a retrieval session on 'texture segmentation' in Figure 11.

4.3 Computational Linguistics

In computational linguistics, the statistical analysis of word co-occurrences in lexical structures like adjective/noun or verb/direct object has recently received a considerable degree of attention [Hin90, PTL93, DLP93, DLP97]. Potential applications of these methods are in word–sense disambiguation, a problem which occurs in different linguistic tasks ranging from parsing and tagging to machine translation.

The data we have utilized to test the different models consists of adjective–noun pairs extracted from a tagged version of the Penn Treebank corpus ($I = 6931$, $J = 4995$, $N = 55214$) and the LOB corpus ($I = 5548$, $J = 6275$, $N = 36723$)[9]. Perplexity results on the Penn dataset are reported in Table 4. The results are qualitatively very similar to the ones obtained on the Cranfield document

[9] Singular and plural forms have been identified.

	Aspect		$\mathcal{X}$-cluster		Hierarchical		$\mathcal{X}/\mathcal{Y}$-cluster	
K	β	$\mathcal{P}$	β	$\mathcal{P}$	β	$\mathcal{P}$	β	$\mathcal{P}$
1	-	685	-	-	-	-	-	-
16	0.72	255	0.07	302	0.10	268	0.51	335
32	0.71	205	0.07	254	0.08	226	0.46	286
64	0.69	182	0.07	**223**	0.07	204	0.44	272
128	0.68	**166**	0.06	231	0.06	**179**	0.40	**241**

Table 4: Comparative perplexity results for adjective–noun pairs from the Treebank corpus.

collection, although this application is quite different from the one in information retrieval. This further supports the conclusions drawn above.

A result for a simultaneous hard clustering of the LOB data with the SCM is reported in Figure 12. The visualization of the $\pi_{\nu\mu}$ matrix reveals that many groups in either space are preferably combined with mainly one group in the complementary space. For example the adjective group 'holy', 'divine', 'human' has its occurrences almost exclusively with nouns from the cluster 'life', 'nature', 'being'. Some groups are very much indifferent with respect to the groups in the corresponding set, e.g., the adjective group headed by 'small', 'big', 'suitable'. '

4.4 Unsupervised Texture Segmentation

The unsupervised segmentation of textured images is still one of the most challenging. Numerous approaches to texture segmentation have been proposed over the past decades, most of which obey a two–stage scheme. In the *modeling stage*: characteristic features are extracted from the textured input image, e.g. spatial frequencies [JF91, HPB98], MRF–models [MJ92]. In the *clustering stage* features are grouped into homogeneous segments, where homogeneity of features is typically formalized by a clustering optimization criterion. Most widely, features are interpreted as vectors in a Euclidean space [JF91, MJ92, PH95] and a segmentation is obtained by minimizing the K–means criterion, which sums over the square distances between feature vectors and their assigned, group–specific *prototype feature vectors*. Occasionally, the grouping process has been based on *pairwise similarity* measurements between image sites, where similarity is measured by a non–parametric statistical test applied to the feature distribution of a surrounding neighborhood [GGGD90, HPB98]. Pairwise similarity clustering thus provides an indirect way to group (discrete) feature distributions without reducing information in a distribution to their mean. Mixture models for dyadic data, especially the one–sided clustering model, formalize the grouping of feature distribution in a more direct manner. In contrast to pairwise similarity clustering, they offer a sound generative model for texture class description which can be utilized in subsequent processing stages like edge localization [SB95]. Furthermore, there is no need to compute a large matrix of pairwise similarity scores between image sites, which greatly reduces the overall processing time and memory requirements. Compared to K–means, these techniques provides significantly more flexibility in distribution modeling. Especially in the texture segmentation application class features often exhibit a non–Gaussian, e.g., multi–modal distribution, which is the main reason for the success of pairwise similarity clustering approaches and the one–sided clustering model compared to standard K–means.

We applied the one-sided clustering model to the unsupervised segmentation of textured images, where objects x correspond to image locations. Since the number of observed features is identical for all sites, one can simply set $\mathrm{P}(x) = 1/N$. In the experiments, we have adopted the framework of [JF91, HPB98, PB98] and utilized an image representation based on the modulus of complex Gabor filters. For each site, the empirical distribution of coefficients in a surrounding (filter-specific) window is determined. All reported segmentation results are based on a filter bank of twelve Gabor filters with four orientations and three scales. Each filter output was discretized into 16 equally

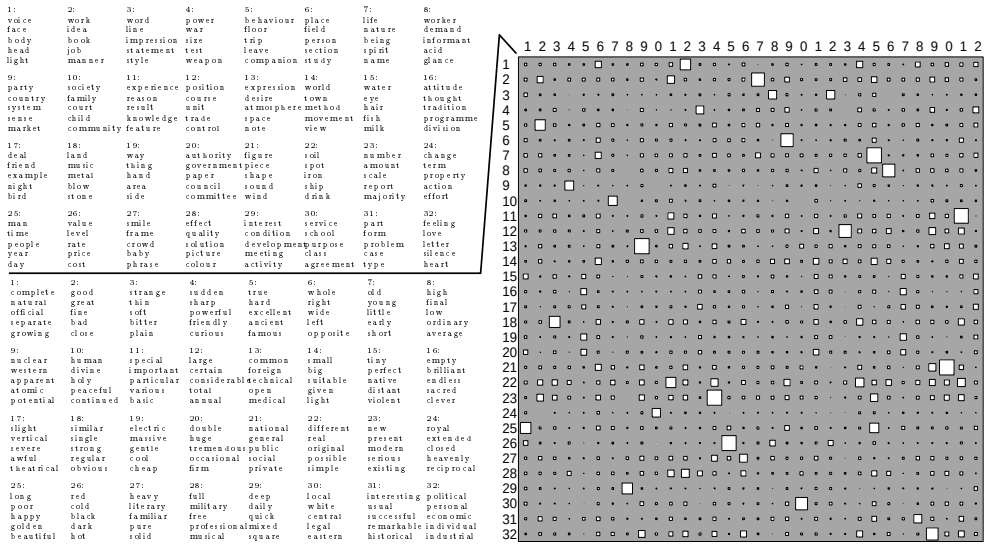

Figure 12: Clustering of LOB using the SCM ($K^{\mathcal{X}} = K^{\mathcal{Y}} = 32$) with a visualization of the $\phi(c,d)$ matrix and a characterization of clusters by their most probable words.

sized bins. As a consequence of the conditional independence assumptions, this results in a feature space $\mathcal{Y}$ of size $M = 192$. For each channel, Gabor coefficients were sampled in a local window of a size proportional to the scale of the filter. For the finest scale a rectangular 16×16 window was utilized. A simple spatial prior was utilized for the class distribution to suppress small regions. For such a prior distribution, multiscale EM is essential to compute high quality solutions.

The excellent segmentation quality obtained by the ACM histogram clustering algorithm is illustrated by the results in Fig. 13. The mixture of $K = 16$ different Brodatz textures has been partitioned accurately with an error rate of 4.7%. The errors basically correspond to boundary sites. The result obtained for the mondrian of aerial images is satisfactory but due to missing ground truth the quality could not be quantified. Disconnected texture regions of the same type have been identified correctly, while problems again occur at texture boundaries. The segmentation quality achieved on outdoor images in Fig. 14 are both visually and semantically satisfying. A detailed evaluation of the one-sided clustering model for unsupervised texture segmentation is out of the scope of this paper and will be published elsewhere [PH98].

5 Conclusion

As the main contribution of this paper a novel class of statistical models for the analysis of co-occurrence data has been proposed and evaluated. We have introduced and discussed several different models. These have been distinguished from a systematic point of view, namely by the way that hidden variables are introduced, which effectively imposes constraints on the component distributions of the mixture. Several recently proposed statistical models have turned out to be special cases. All models have a sound statistical foundation in that they define a generative distribution, and all of them can be fitted by an (at least approximate) EM algorithm.

The proper selection of the method of choice for a given problem crucially depends on the modeling goal. We have argued, that it is often required to detect groups structure or hierarchical representations. In these situations, one may sacrifice some precision in terms of statistical accuracy (i.e., perplexity reduction) in order to extract the structure of interest. Within the proposed

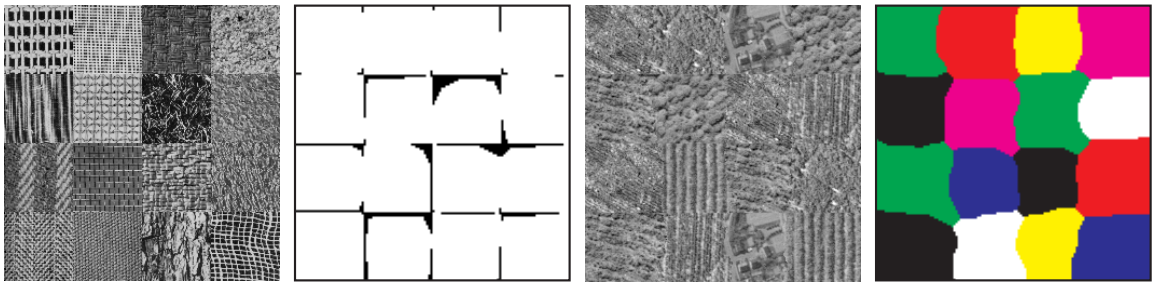

Figure 13: Left: Mixture image containing $K = 16$ different textures from the Brodatz album-Segmentation errors are depicted in black. Right: Mixture images containing 7 textures extracted from aerial images.

framework, models have been derived to extract group structure on either one or simultaneously both object spaces and to model hierarchical dependencies of clusters. We strongly believe that the proposed framework is flexible enough to be adapted to many different tasks. The generality of the developed methods has been stressed by revealing their benefits in the context of a broad range of potential applications.

In addition to the modeling problem, we have addressed computational issues, in particular focusing on improved variants of the basic EM algorithm. Most importantly, our experiments underline the possible advantages of the annealed version of EM, which is a fruitful combination of ideas and methods from statistics and statistical physics.

Acknowledgment

The authors wish to thank Michael Jordan, Peter Dayan, Tali Tishby, and Joachim Buhmann for helpful comments and suggestions. The authors are grateful to Carl de Marcken and Joshua Goodman for sharing their expertise and data in natural language processing as well as to J.M.H. du Buf for providing the image data depicted in Fig. 13.

Appendix

EM, Annealed EM and Free Energy Minimization

In [NH98], it has been shown that both, the E-step and M-step of the EM algorithm, are minimizing a (generalized) free energy criterion. This fact is of importance, in particular for deriving approximate E-steps. Let Z denote a generic latent variable with realizations z and consider the following family of objective functions:

$$\mathcal{F}_T(\theta, \theta') = \sum_z P(z|\mathcal{S}, \theta') \left[\log P(\mathcal{S}|z, \theta) + T \log \frac{P(z|\theta)}{P(z|\mathcal{S}, \theta')} \right] \quad . \tag{51}$$

Here $\mathcal{S}$ is a sample set, θ a parameter vector, and $T > 0$ corresponds to the *computational temperature*. $\mathcal{F}_1$ is the sum of the expected complete data log-likelihood $\mathcal{L}$ and the entropy of the posterior distribution. The derivative of $\mathcal{L}$ and $\mathcal{F}_1$ w.r.t. θ is thus identical. By maximizing $\mathcal{F}_T$ w.r.t. the probability distribution $P(z|\mathcal{S}, \theta')$ (treating the latter as variational parameters without their intended meaning) we get

$$P(z|\ldots) \propto P(\mathcal{S}|z, \theta)^{1/T} P(z|\theta) \quad . \tag{52}$$

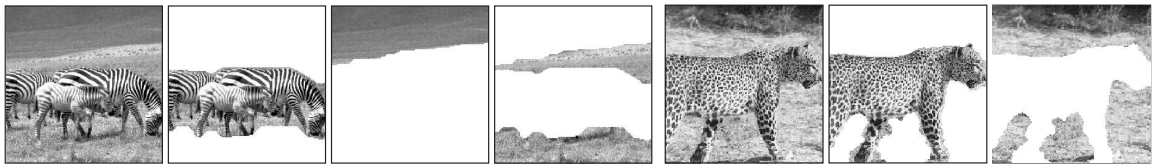

Figure 14: Typical segmentation result on a real-world image with $K = 3$ (left) and $K = 2$ (right) segments obtained by ACM.

For $T = 1$ this indeed recovers the posterior probabilities. This proves that $\mathcal{F}_1$ is maximized in every E-step w.r.t. $P(z|\ldots)$ and in every M-step w.r.t. θ; hence the free energy $-\mathcal{F}_1$ is a Lyapunov function of the EM algorithm. For arbitrary $T > 0$ (52) motivates annealed EM from an optimization principle.

Approximate EM and Mean Field Approximation

The above optimization principle offers a systematic approach for deriving approximate E-steps by restricting the optimization w.r.t. $P(z|\ldots)$ to a particular sub-family $\mathcal{Q}$ of distributions $Q(z)$. The approximating sub-family is chosen according to tractability considerations, i.e., it will typically have a simplified factorial form. In the general formulation of the approximate E-step thus the probability distribution $Q(z) \in \mathcal{Q}$ is chosen which maximizes $\mathcal{F}_1$. It can also be shown that this variational principle is equivalent to minimizing the KL–divergence between the approximation Q and the true posterior distribution $P(z|\mathcal{S}, \theta')$.

In the co-occurrence modeling framework, we have utilized this general principle for the two-sided clustering model, parameterizing $\mathcal{Q}$ by

$$Q(c,d) \quad = \quad \left[\prod_{x \in \mathcal{X}} Q(x, c(x))\right] \cdot \left[\prod_{y \in \mathcal{Y}} Q(y, d(y))\right] \quad , \tag{53}$$

with the set of constraints $\sum_c Q(x,c) = 1$ and $\sum_d Q(y,d) = 1$. Inserting into the expression for $\mathcal{F}_1$ and making use of factorial form of Q one obtains for the relevant (i.e., Q–dependent) part $\tilde{\mathcal{F}}_1$ of $\mathcal{F}_1$ (augmented by the Lagrange multiplier term to enforce the normalization)

$$\begin{aligned} \tilde{\mathcal{F}}_1(\theta, \theta') \quad = \quad & \sum_{x,y} n(x,y) \sum_{c,d} Q(x,c) Q(y,d) \log \phi(c,d) \\ & + \sum_{x,c} Q(x,c) \log \frac{\pi(c)}{Q(x,c)} + \sum_{y,d} Q(y,d) \log \frac{\pi(d)}{Q(y,d)} \\ & + \lambda \left[\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \frac{n(x) n(y)}{N^2} \sum_{c \in \mathcal{C}} \sum_{d \in \mathcal{D}} Q(x,c) Q(y,d) \phi(c,d) - 1 \right] \quad . \end{aligned} \tag{54}$$

Performing the maximization of $\tilde{\mathcal{F}}_1$ w.r.t. one set of variables, e.g., $Q(x,c)$, then yields

$$Q(x,c) \quad \propto \quad \pi(c) \exp \left[\sum_y n(x,y) \sum_d Q(y,d) \log \phi(c,d) \right] \quad , \tag{55}$$

provided that

$$\phi(c,d) = \frac{N \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} Q(x,c) Q(y,d) n(x,y)}{\left[\sum_{x \in \mathcal{X}} Q(x,c) n(x)\right] \cdot \left[\sum_{y \in \mathcal{Y}} Q(y,d) n(y)\right]} \quad . \tag{56}$$

Eq. (56) is simply the M-step equation with approximated posterior probabilities (cf. (27)), but one has to assure that the variables $Q(y,d)$ in (56) and (54) are identical in order to get the simple expression in (55).

Leave-One-Out E-step

The same optimization framework can be utilized to derive a leave-one-out modified E-step from a strict optimization principle. We demonstrate the idea for the aspect model and introduce variational parameters $Q(x,a)$ for (corrected) posterior probabilities. After substituting the (corrected) M-step equations of all parameters into $\mathcal{F}$ one obtains

$$\begin{aligned}\mathcal{F}(Q) &= \sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{A}} Q(x,a)\left[\log\frac{1}{I-1}\sum_{x'\neq x} Q(x',a) - \log Q(x,a)\right. \\ &\quad \left. + \sum_{y\in\mathcal{Y}} n(x,y)\log\frac{\sum_{x'\neq x} n(x,y)Q(x',a)}{\sum_{x'\neq x} n(x)Q(x',a)}\right] .\end{aligned} \tag{57}$$

Now, we can maximize $\mathcal{F}$ w.r.t. Q. The derivation of stationary conditions is straightforward, but involves some very technical algebraic manipulation.

References

[And97] E. Anderson. *Introduction to the Statistical Analysis of Categorial Data*. Springer, 1997.

[Bar87] D. J. Bartholomew. *Latent variable models and factor analysis*. Number 40 in Griffin's statistical monographs & courses. Oxford University Press, New York, 1987.

[BAS95] C. Buckley, J. Allan, and G. Salton. Automatic routing and retrieval using SMART – TREC2. *Information Processing and Managment*, 31(3):315–326, 1995.

[BdM+92] P.F. Brown, P.V. deSouza, R.L. Mercer, V.J. Della Pietra, and J.C. Lai. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.

[BK93] J.M. Buhmann and H. Kühnel. Vector quantisation with complexity costs. *IEEE Transactions on Information Theory*, 39:1133–1145, 1993.

[BKS97] E. Bauer, D. Koller, and Y. Singer. Update rules for parameter estimation in bayesian networks. In *Proceedings of the 13th Annual Conference on Uncertainty in AI (UAI)*, Providence, Rhode Island, August 1997.

[Boc74] H. H. Bock. *Automatische Klassifikation : theoretische und praktische Methoden zur Gruppierung und Strukturierung von Daten (Cluster-Analyse)*. Number 24 in Studia mathematica; mathematische Lehrbücher. Vandenhoeck und Ruprecht, Göttingen, 1974.

[Buh98] J. M. Buhmann. Empirical risk approximation: An induction principle for unsupervised learning. Technical Report IAI-TR-98-3, Institut für Informatik III, University of Bonn, 1998.

[BWD97] M. Blatt, S. Wiseman, and E. Domany. Data clustering using a model granular magnet. *Neural Computation*, 9(8):1805–1842, 1997.

[CA80] J. D. Carroll and P. Arabie. Multidimensional scaling. *Annual Review of Psychology*, 31:607–649, 1980.

[CKP92] D.R. Cutting, D.R. Karger, and J.O. Pedersen. Scatter/gather: a cluster-based approach to browsing large document collections. In *Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1992.

[Coo64] C. H. Coombs. *A Theory of Data.* John Wiley & Son, 1964.

[CSS98] W. W. Cohen, R. E. Shapire, and Y. Singer. Learning to order things. In *Advances in Neural Information Processing Systems 10*, 1998.

[CT91] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley & Sons, New York, 1991.

[DLP93] I. Dagan, L. Lee, and F.C.N. Pereira. Similarity-based estimation of word cooccurence probabilities. In *Proceedings of the Association for Computational Linguistics*, 1993.

[DLP97] I. Dagan, L. Lee, and F.C.N. Pereira. Similarity-based methods for word sense disambiguation. In *Proceedings of the Association for Computational Linguistics*, 1997.

[DLR77] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. B*, 39:1–38, 1977.

[DTGL90] S. Deerwester, Dumais S. T., Furnas G.W., and T.K. Landauer. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 1990.

[ES92] U. Essen and V. Steinbiss. Cooccurrence smoothing for stochastic language modeling. In *Proceedings of the IEEE nternational Conference on Acoustics, Speech, and Signal Processing*, pages 161–164, 1992.

[GGGD90] D. Geman, S. Geman, C. Graffigne, and P. Dong. Boundary detection by constrained optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(7):609–628, 1990.

[GNOD92] D. Goldberg, D. Nichols, B. M. Oki, and Terry D. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70, 1992.

[Goo53] I.J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3):237–264, 1953.

[Goo65] I.J. Good. *The Estimation of Probabilities.* Research Monograph 30. MIT Press, Cambridge, MA, 1965.

[HB97] T. Hofmann and J.M. Buhmann. Pairwise data clustering by deterministic annealing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(1), 1997.

[Hin90] D. Hindle. Noun classification from predicate–argument structures. In *Proceedings of the ACL*, pages 268–275, 1990.

[HPB94] F. Heitz, P. Perez, and P. Bouthemy. Multiscale minimization of global energy functions in some visual recovery problems. *CVGIP: Image Understanding*, 59(1):125–134, 1994.

[HPB98] T. Hofmann, J. Puzicha, and J.M. Buhmann. Unsupervised texture segmentation in a deterministic annealing framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998.

[HPJ99] T. Hofmann, J. Puzicha, and M. I. Jordan. Learning from dyadic data. In *Advances in Neural Information Processing Systems 11*, 1999.

[JD88] A.K. Jain and R.C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, NJ 07632, 1988.

[Jel85] F. Jelinek. The development of an experimental discrete dictation recogniser. *Proceedings of the IEEE*, 73(11), 1985.

[JF91] A. Jain and F. Farrokhnia. Unsupervised texture segmentation using Gabor filters. *Pattern Recognition*, 24(12):1167–1186, 1991.

[JGJS98] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul. An introduction to variational methods for graphical models. In M.I. Jordan, editor, *Learning in Graphical Models*, pages 105–161. Kluwer Academic Publishers, 1998.

[JJ94] M.I. Jordan and R.A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, 1994.

[JM80] F. Jelinek and R. Mercer. Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the Workshop of Pattern Recognition in Practice*, 1980.

[Kat87] S.M. Katz. Estimation of probabilities for sparse data for the language model component of a speech recogniser. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(3):400–401, 1987.

[KMMH97] J.A. Konstan, B. N. Miller, D. Maltz, and J. L. Herlocker. Grouplens: Applying collaborative filtering to Usenet news. *Communications of the ACM*, 40(3):77–87, 1997.

[Kru78] J. B. Kruskal. *Multidimensional Scaling*. Sage Publications, Beverly Hills, CA, 1978.

[LDC97] LDC. Linguistic Data Consortium: TDT pilot study corpus documentation. http://www.ldc.upenn.edu/TDT, 1997.

[LS89] K.E. Lochbaum and L.A. Streeter. Comparing and combining the effectiveness of latent semantic indexing and the ordinary vector space model for information retrieval. *Information Processing & Management*, 1989.

[MB88] G.J. McLachlan and K. E. Basford. *Mixture Models*. Marcel Dekker, INC, New York Basel, 1988.

[MJ92] J. Mao and A. Jain. Texture classification and segmentation using multiresolution simultaneous autoregressive models. *Pattern Recognition*, 25:173–188, 1992.

[MJ98] M. Meila and M. I. Jordan. Estimating dependency structure as a hidden variable. In *Advances in Neural Information Processing Systems 10*, 1998.

[MK97] G.J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, New York, 1997.

[NH98] R.M. Neal and G.E. Hinton. A view of the EM algorithm that justifies incremental and other variants. In M.I. Jordan, editor, *Learning in Graphical Models*, pages 355–368. Kluwer Academic Publishers, 1998.

[PB98] Jan Puzicha and Joachim Buhmann. Multi–scale annealing for real–time unsupervised texture segmentation. In *Proceedings of the International Conference on Computer Vision (ICCV'98)*, pages 267–273, 1998.

[PH95] D. Panjwani and G. Healey. Markov random field models for unsupervised segmentation of textured color images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(10):939–954, 1995.

[PH98] J. Puzicha and T. Hofmann. Histogram clustering for unsupervised segmentation and image retrieval. Technical report, Echtzeit-Optimierung Preprint 98-33, submitted to Pattern Recognition Letters, 1998.

[PHB99] J. Puzicha, T. Hofmann, and J. Buhmann. A theory of proximity based clustering: Structure detection by optimization. *Pattern Recognition*, 1999.

[PKM96] K. Pawelzik, J. Kohlmorgen, and K. R. Müller. Annealed competition of experts for a segmentation and classification of switching dynamics. *Neural Computation*, 1996.

[PTL93] F.C.N. Pereira, N.Z. Tishby, and L. Lee. Distributional clustering of english words. In *Proceedings of the ACL*, pages 183–190, 1993.

[PW78] B. C. Peters and H. F. Walker. An iterative procedure for obtaining maximum-likelihood estimates of the parameters for a mixture of normal distributions. *SIAM Journal of Applied Mathematics*, 35:362–378, 1978.

[RGF90] K. Rose, E. Gurewitz, and G. Fox. Statistical mechanics and phase transitions in clustering. *Physical Review Letters*, 65(8):945–948, 1990.

[RGF92] K. Rose, E. Gurewitz, and G. Fox. Vector quantization by deterministic annealing. *IEEE Transactions on Information Theory*, 38(4):1249–1257, 1992.

[RMRG97] A. Rao, D. Miller, K. Rose, and A.. Gersho. Deterministically annealed mixture of experts models for statistical regression. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 3201–3204. IEEE Comput. Soc. Press, 1997.

[SB91] G. Salton and C. Buckley. Global text matching for information retrieval. *Science*, 253(5023):1012–1015, August 1991.

[SB95] P. Schroeter and J. Bigun. Hierarchical image segmentation by multi-dimensional clustering and orientation-adaptive boundary refinement. *Pattern Recognition*, 28(5):695–709, 1995.

[Sch98] H. Schutze. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123, 1998.

[SM83] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw–Hill, 1983.

[SP97a] L. Saul and F. Pereira. Aggregate and mixed–order Markov models for statistical language processing. In *Proceedings of the 2nd International Conference on Empirical Methods in Natural Language Processing*, 1997.

[SP97b] H Schutze and J.O. Pedersen. A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing and Managment*, 33(3):307–318, 1997.

[TP98] N.Z Tishby and F.C.N. Pereira. personal communication and manuscript, in preparation, 1998.

[TSM85] D.M. Titterington, A.F.M. Smith, and U.E. Makov. *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons, 1985.

[VR75] C. J. Van Rijsbergen. *Information Retrieval.* Butterworths, London, Boston, 1975.

[WB91] I.H. Witten and T.C. Bell. The zero–frequency problem: estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, 37(4):1085–1094, 1991.

[Whi87] J. Whittaker. *Graphical Models in Applied Multivariate Statistics.* John Wiley & Son, Chichester, 1987.

[Wil88] P. Willett. Recent trends in hierarchical document clustering: a critical review. *Information Processing & Management*, 24(5):577–597, 1988.

[WL90] N. Wermuth and S.L. Lauritzen. On substantive research hypotheses, conditional independence graphs and graphical chain models. *Journal of the Royal Statistical Society Series B-Methodological*, 1990.