

Mixture Models for Co-occurrence and Histogram Data*

Thomas Hofmann¹ and Jan Puzicha²

¹ Artificial Intelligence Laboratory, M.I.T., Cambridge, MA 02139, hofmann@ai.mit.edu

² Institut für Informatik III, University of Bonn, D-53177 Bonn, Germany, jan@cs.uni-bonn.de

Abstract

Modeling and predicting co-occurrences of events is a fundamental problem of unsupervised learning. In this contribution, we develop a general statistical framework for analyzing co-occurrence data based on probabilistic clustering by mixture models. More specifically, we discuss three models which pursue different modeling goals and which differ in the way they define the probabilistic partitioning of the observations. Adopting the maximum likelihood principle, annealed EM algorithms are derived for parameter estimation. From the class of potential applications in pattern recognition and data analysis, we have chosen document retrieval, language modeling, and unsupervised texture segmentation to test and evaluate the proposed algorithms.

1. Introduction

The type of data investigated in this paper is best described by the term *co-occurrence data* (COD). The general setting is as follows: Suppose two finite sets $\mathcal{X} = \{x_1, \dots, x_N\}$ and $\mathcal{Y} = \{y_1, \dots, y_M\}$ of abstract objects with arbitrary labeling are given. The elementary observations are pairs $(x_i, y_j) \in \mathcal{X} \times \mathcal{Y}$, i.e., joint occurrences of objects from \mathcal{X} and \mathcal{Y} . All observations are collected in a sample set $\mathcal{S} = \{(x_{i(r)}, y_{j(r)}, r) : 1 \leq r \leq L\}$ with arbitrary ordering and sufficient statistics $n_{ij} = |\{(x_i, y_j, r) \in \mathcal{S}\}|$ (co-occurrence frequencies). If \mathcal{S} is partitioned into subsets \mathcal{S}_i according to the \mathcal{X} -component, the sets \mathcal{S}_i correspond to *histograms* over a finite feature space \mathcal{Y} .

Co-occurrence and histogram data is found in many distinctive applications. For example, in information retrieval we may identify \mathcal{X} with a collection of documents and \mathcal{Y} with a set of keywords (context–event relation). In computational linguistics \mathcal{X} and \mathcal{Y} may correspond to words being part of a binary syntactic structure such as nouns with corresponding adjectives (structural relation). In computer vision \mathcal{X} may correspond to image locations and \mathcal{Y} to discretized or categorical feature values (histogram data). Other potential application domains are data mining, molecular biology, and preference analysis.

*This research has been supported by a M.I.T. Faculty Sponser's Discretionary Fund and by the German Research Foundation (DFG) under grant # BU 914/3–1. A detailed version of this paper can be found in [4]. It is a pleasure to thank Hans du Buf for providing aerial image data.

The intrinsic problem of COD is *data sparseness*. For large object sets a majority of pairs (x_i, y_j) only has a small probability of being observed even for large sample sets. To overcome the sparseness problem we discuss a family of *finite mixture models*. Mixture models provide a sound statistical foundation and can rely on the calculus of probability theory as a powerful inference mechanism. Moreover, they offer a natural framework for unifying statistical inference and clustering. This is particularly important, since one is often interested in *discovering structure*, typically represented by groups of similar objects as in *pairwise data clustering* [3]. As a major advantage clustering based on COD does not require an external similarity measure, but exclusively relies on the occurrence statistics. Several clustering and mixture models for COD have recently been investigated [1, 6, 8]. Our approach provides a unifying framework for all of these models.

2. The Separable Mixture Model

The first model we propose is the *Separable Mixture Model* (SMM). Introducing K abstract classes \mathcal{C}_α data is generated according to the following scheme: (i) choose a class \mathcal{C}_α with probability π_α , (ii) select x_i from a class-conditional distribution $p_{i|\alpha}$, (iii) select y_j with probability $q_{j|\alpha}$. This yields the joint probability $p_{ij} \equiv \sum_\alpha \pi_\alpha p_{i|\alpha} q_{j|\alpha}$. The component distributions are separable, i.e., x_i and y_j are conditionally independent given the class \mathcal{C}_α . We apply a standard maximum likelihood technique known as the Expectation–Maximization (EM) Algorithm [2] and introduce indicator variables $R_{r\alpha} \in \{0, 1\}$ to represent the unknown class \mathcal{C}_α of the r -th observation. The E-step requires the computation of posteriors $\langle R_{r\alpha} \rangle$ for the hidden variables which are given by

$$\langle R_{r\alpha} \rangle = \frac{\pi_\alpha p_{i(r)|\alpha} q_{j(r)|\alpha}}{\sum_{\nu=1}^K \pi_\nu p_{i(r)|\nu} q_{j(r)|\nu}}. \quad (1)$$

For the M-step $\pi_\alpha = \frac{1}{L} \sum_{r=1}^L \langle R_{r\alpha} \rangle$, $p_{i|\alpha} \propto \sum_{r:i(r)=i} \langle R_{r\alpha} \rangle$, and $q_{j|\alpha} \propto \sum_{r:j(r)=j} \langle R_{r\alpha} \rangle$ is obtained. Iterating the E- and M-step, the parameters converge to a local maximum of the likelihood.¹

¹In the special case of $\mathcal{X} = \mathcal{Y}$ the SMM is equivalent to the word clustering model of Saul and Pereira [8] (proposed independently in [4]).

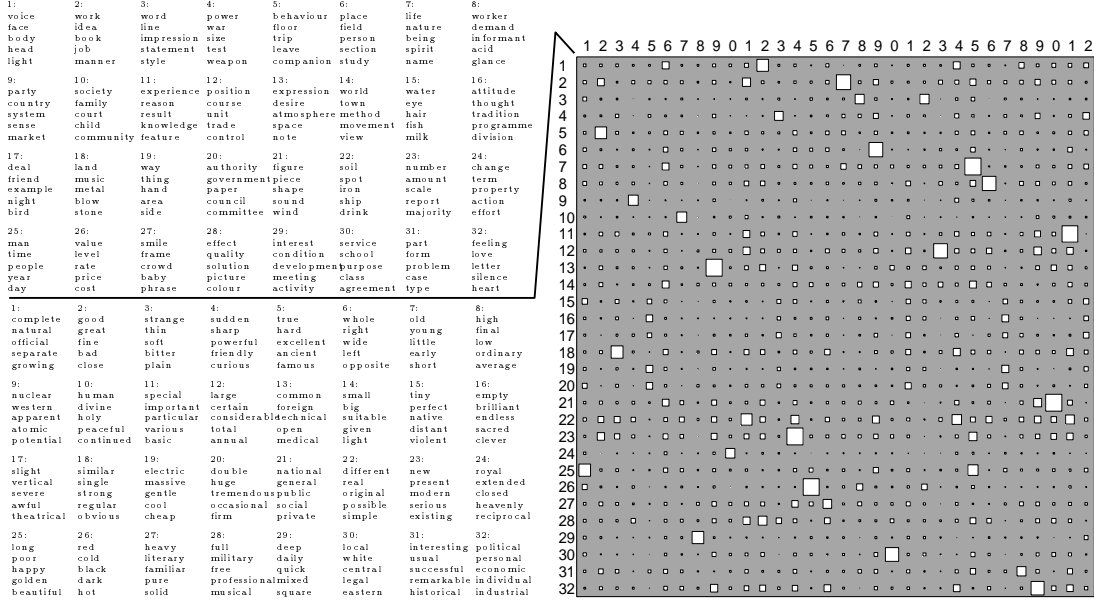


Figure 1. Clustering of LOB using the SCM ($K^{\mathcal{X}} = K^{\mathcal{Y}} = 32$) with a visualization of the $\pi_{\nu\mu}$ matrix and a characterization of clusters by their most probable words.

3. The Asymmetric Clustering Model

The grouping structure inferred by the SMM corresponds to a probabilistic partitioning of the observation space $\mathcal{X} \times \mathcal{Y}$. Depending on the specific application it might be more appropriate to assume a definitive assignment of objects (e.g. in \mathcal{X}) to clusters. The *Asymmetric Clustering Model (ACM)* introduces hidden variables $I_{i\alpha}$ to represent the class memberships of *objects* x_i , $p_{ij} \equiv p_i \sum_{\alpha=1}^K I_{i\alpha} q_{j|\alpha}$. Obviously, we can interchange the role of \mathcal{X} and \mathcal{Y} to obtain two distinct models.² The E-step for the ACM is given by

$$\langle I_{i\alpha} \rangle = \frac{\rho_{\alpha} \prod_{j=1}^M (q_{j|\alpha})^{n_{ij}}}{\sum_{\nu=1}^K \rho_{\nu} \prod_{j=1}^M (q_{j|\nu})^{n_{ij}}} \quad (2)$$

where ρ_{α} are prior probabilities. The M-step equations are $q_{j|\alpha} \propto \sum_{i=1}^N n_{ij} \langle I_{i\alpha} \rangle$, $\rho_{\alpha} = \frac{1}{N} \sum_{i=1}^N \langle I_{i\alpha} \rangle$, $p_i = \sum_j n_{ij} / L$.

4. Symmetric Clustering Model

The last model we introduce infers clustering structure in both the \mathcal{X} and the \mathcal{Y} space simultaneously. We utilize hidden variables $I_{i\nu}$ and $J_{j\mu}$ to denote the assignment of objects x_i and y_j to clusters in \mathcal{X} and \mathcal{Y} , respectively. Introducing cluster association parameters $c_{\nu\mu}$, the joint probability of the *symmetric clustering model (SCM)* is defined by $p_{ij} \propto p_i q_j \sum_{\nu=1}^K \sum_{\mu=1}^K I_{i\nu} J_{j\mu} c_{\nu\mu}$. In the SCM, the

coupling of I and J makes the exact computation of posteriors in the E-step intractable and we have to recourse to a factorial approximation, $\langle I_{i\nu} J_{j\mu} \rangle \approx \langle I_{i\nu} \rangle \langle J_{j\mu} \rangle$ (mean-field approximation), which results in the following approximate E-step (analogously for $\langle J_{i\nu} \rangle$)

$$\langle I_{i\nu} \rangle \propto \rho_{\nu} \prod_{j,\mu} e^{-n_{ij} \langle J_{j\mu} \rangle \log c_{\nu\mu}}, \quad (3)$$

where ρ_{ν} are priors. Taking into account the normalization constraint on p_{ij} (cf. [4]), $c_{\nu\mu} = \pi_{\nu\mu} / (\pi_{\nu}^x \pi_{\mu}^y)$, where $\pi_{\nu\mu} = \sum_{i,j} \langle I_{i\nu} \rangle \langle J_{j\mu} \rangle n_{ij} / L$ and $\pi_{\nu}^x = \sum_{\mu} \pi_{\nu\mu}$, $\pi_{\mu}^y = \sum_{\nu} \pi_{\nu\mu}$ is obtained. The resulting approximate EM alternates the update of posterior marginals with an update of the continuous parameters. Considering the maximum likelihood estimator for $c_{\nu\mu}$ as a function of I and J , the optimization of \mathcal{L}^c amounts to finding partitions in \mathcal{X} and \mathcal{Y} which maximize a *mutual information* criterion.³

In summary, the proposed models can be systematically distinguished by the restriction imposed on the class-conditional distributions. In the SMM, both $p_{i|\alpha}$ and $q_{j|\alpha}$ are arbitrary multinomial distributions. The ACM introduces an asymmetry by restricting only one of the distributions $p_{i|\alpha} \propto p_i \langle I_{i\alpha} \rangle$. In the SCM both distributions are restricted. Additional model variants as well as more details can be found in [4].

5. Annealed EM

Annealed EM is a method to deal with two problems of the standard EM approach: the sensitivity to local maxima

²The ACM is similar to the distributional clustering model proposed in [6], cf. the discussion in [4].

³A similar (hard-clustering) objective function has been proposed by Brown et al. [1] in their (non-probabilistic) class-based n -gram model.

Dataset	K	SMM		ACM		SCM	
		β	\mathcal{P}	β	\mathcal{P}	β	\mathcal{P}
CRAN	1	-	685	-	-	-	-
	32	0.83	386	0.07	452	0.53	506
	64	0.79	360	0.06	527	0.48	477
	128	0.78	353	0.04	663	0.45	462
PENN	1	-	639	-	-	-	-
	32	0.71	205	0.07	254	0.46	286
	64	0.69	182	0.07	223	0.44	272
	128	0.68	166	0.06	231	0.40	241

Table 1. Perplexity for SMM, ACM, and SCM on two data sets (CRAN: predicting words conditioned on documents, PENN: predicting nouns conditioned on adjectives) based on ten-fold cross validation ($K^x = K^y = K$ for SCM).

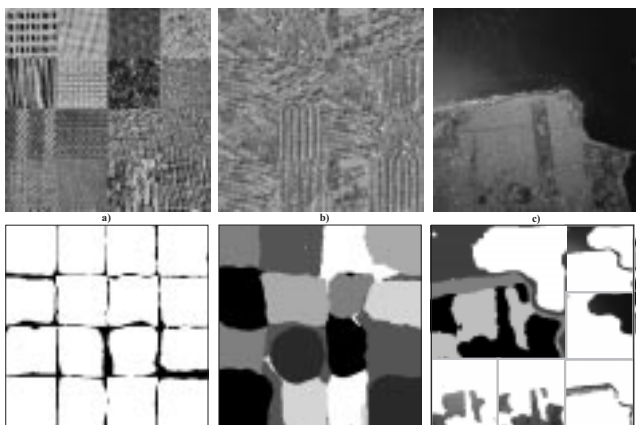


Figure 2. Typical segmentation results: (a) mondrian of 16 Brodatz textures (errors are depicted in black), (b) mondrian of 7 textures from aerial images, (c) aerial image of San Francisco.

and the problem of overfitting. Annealed EM is based on the idea of *deterministic annealing* which has been applied to many clustering problems, including vectorial clustering [7], pairwise clustering [3], and for distributional clustering [6]. The key idea is to introduce an (inverse temperature) parameter β , and to replace the negative (averaged) log-likelihood by a substitute known as the *free energy*. This can be understood as a homotopy method, where the likelihood is smoothed for small β and is recovered in the limit of $\beta \rightarrow 1$. Since annealing effectively performs a regularization based on entropy, it is utilized to improve the generalization by stopping at $\beta < 1$. In annealed EM the E-step is modified by taking the likelihood contribution in Bayes' rule to the power of β . For example, in the case of the ACM (2) is modified by replacing n_{ij} in the exponential with βn_{ij} , thus reducing the effective sample size. In order to determine the optimal value for β one may use an additional validation set.

6. Results

For the reported experiments we have utilized data from three different domains: (i) Word occurrence data from the Cranfield document collection (CRAN), (ii) adjective–noun co-occurrences from tagged versions of the Penn Treebank (PENN) and the LOB corpus (LOB), (iii) localized histogram data based on a Gabor multi-scale image representation of aerial images (cf. [5]).

In a series of experiments we have investigated how well different models perform in predicting occurrences, evaluating the perplexity \mathcal{P} .⁴ The results are summarized in Table 1. The main conclusions are: (i) the SMM obtains the lowest perplexity, (ii) annealed EM consistently improves over standard EM, (iii) temperature-based complexity control is superior to restricting the number of components.

A result of a simultaneous clustering of adjectives and nouns with the SCM is reported in Fig. 1. 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. Figure 2 depicts some exemplary segmentations of textured images obtained by clustering local histograms with the ACM. A detailed benchmark study of this novel segmentation algorithm including comparisons with state-of-the-art techniques will appear in a forthcoming paper.

References

- [1] P. Brown, P. deSouza, R. Mercer, V. Della Pietra, and J. Lai. Class-based n -gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.
- [2] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.
- [3] T. Hofmann and J. Buhmann. Pairwise data clustering by deterministic annealing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(1), 1997.
- [4] T. Hofmann and J. Puzicha. Statistical models for cooccurrence data. Technical report, Artificial Intelligence Laboratory Memo 1625, M.I.T., 1998.
- [5] T. Hofmann, J. Puzicha, and J. Buhmann. Deterministic annealing for unsupervised texture segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998 (to appear).
- [6] F. Pereira, N. Tishby, and L. Lee. Distributional clustering of English words. In *30th Annual Meeting of the ACL*, pages 183–190, 1993.
- [7] K. Rose, E. Gurewitz, and G. Fox. Statistical mechanics and phase transitions in clustering. *Physical Review Letters*, 65(8):945–948, 1990.
- [8] 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.

⁴ \mathcal{P} is related to the average test set log-likelihood l by $\mathcal{P} = e^{-l}$.