# A PERCEPTUAL HASHING ALGORITHM USING LATENT DIRICHLET ALLOCATION

*Nicholas Vretos, Nikos Nikolaidis and Ioannis Pitas*

Department of Informatics, Aristotle University of Thessaloniki
Thessaloniki 54124, Greece Tel,Fax: +30-2310996304
email: {nikolaid, pitas}@aiia.csd.auth.gr

**ABSTRACT**

This paper investigates the possibility of extracting latent aspects of a video, using visual information about humans (e.g. actors' faces), in order to develop a fingerprinting (replica detection) framework. We employ a generative probabilistic model, namely Latent Dirichlet Allocation (LDA), so as to capture latent aspects of a video, using facial semantic information derived from the video. We use the bag-of-words concept, (bag-of-faces in our case) in order to ensure exchangeability of the latent variables (e.g. topics). The video topics are modeled as a mixture of distributions of faces in each video. This generative probabilistic model has already been used in the case of text modeling with good results. Experimental results provide evidence that the proposed method performs very efficiently for video fingerprinting.

## 1. INTRODUCTION

Video perceptual hashing also known as video fingerprinting or near replica detection [1], refers to methods that try to identify whether a given video is a replica or a near replica of one of the videos existing in a video database. In this context, near replication means that two videos are either identical or that one video was generated from the other (the original) video through one or more typical video manipulation procedures. When queried with a video, the system should return a single video that is the original of the query video or report that there are no matching videos in the database. Each video is described by a feature vector, called fingerprint, that should be robust to content manipulation (attacks).

A limited number of video fingerprinting or replica detection techniques have been proposed in the literature so far. In [2], Indyk et al. use temporal fingerprints based on the shot boundaries of a video sequence in order to find pirated videos on the Internet. Oostveen et al. in [3] have proposed a spatio-temporal fingerprint based on the differential of luminance in spatiotemporal blocks.

Latent Dirchlet Allocation (LDA) is a generative probabilistic model introduced in [4]. It consists of a three-level hierarchical Bayesian network where each database item (in our case a video) is modeled as a finite mixture of an underlying set of topics. Each topic is, in turn, modeled as an infinity mixture of a latent set of distributions of discrete data. In subsequent sections we will further explain the LDA framework in more detail. LDA framework has been recently used in the context of image [5], [6] and video analysis and description [7], [8].

In this paper, we utilize the fact that, in a video, there are latent aspects of its content which are hopefully robust under attacks and we try to use LDA in order to extract them. The semantic units that are used as input data to the LDA are related to pictorial actor face instances and will be called from now on "facewords". In this context, actor denotes any person appearing in a video clip or movie. This choice has been adopted for two reasons: first, the actors appearing in a video are an important and distinctive high level video characteristic. Second, actors usually make appearances throughout the entire video and, thus, even if we have only a small video excerpt, we can still extract useful information in order to match it against the original video. In our case, information regarding the actor appearances in a video is derived through face detection/tracking. In addition, since we are not only interested in face detection but rather in the appearances of a specific actor, a face clustering approach utilizing SIFT features is used, in order to cluster detected facial images and, thus, discover appearances of the same actor in different time instances within a video and across the database videos. This results in the creation of a bag-of-faces for each video, without necessarily knowing the actor identity, i.e. without performing face recognition which, overall, is a much more difficult problem to solve.

The novelty of this paper lies mainly in the use of latent aspects of the video content. We aim to extract the underlying video topics and to use them in video fingerprinting. In more detail, this paper includes the following novelties:

- The use of face occurrences in a video as facewords that describe this video.
- The use of latent semantic analysis for video fingerprinting. Although many papers are using probabilistic Latent Semantic Analysis (pLSA) for a number of image and video processing tasks, only very recent publications like [5] and [9] have utilized the LDA algorithm. However, none did use this framework for video fingerprinting, to the best of our knowledge.

4. Finally, conclusions are drawn in Section5.

## 2. FEATURE EXTRACTION AND DATA ORGANIZATION

In this section, we shall describe the facewords used in order characterize a video and outline the proposed framework for video fingerprinting. For each video, two steps are undertaken:

a) Face detection. The Viola and Jones face detector [10] is used in order to extract facial images from a video. We use the training set defined in OpenCV for frontal faces and, thus, the resulting facial images are frontal or nearly frontal.

b) Face clustering using SIFT features. Since the proposed approach is based on face appearances of specific actors, face detection is followed by a face clustering step. This step is accomplished by evaluating facial image similarity based on SIFT features [11]. The face clustering approach is inspired from the work of Antonopoulos et al. [12].

At the end of the face clustering procedure over the entire video database, the formed facial image clusters constitute the universal

vocabulary of this video database. The cardinality of the universal vocabulary is equal to the number of the formed facial image clusters. We use the term universal to distinguish between facewords in a video and facewords over the entire database, the term universal applying to the latter one. The appearance of a specific face in a particular time in the video is considered as an instance of a specific faceword from the universal vocabulary. For instance, a video is characterized by a sequence (A,A,B,A,B,B,C,...) where each of the facewords A, B, C corresponds to the appearance of three specific actors faces. The above mentioned process creates an actor appearance histogram for each video in the database. These histograms are used as an estimate of the probability distribution of the actors in each video in the database.

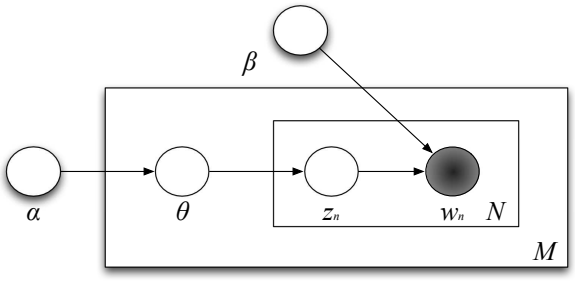


**Fig. 1**. The LDA graphical model.

### 2.1. Data Organization

Many latent semantic analysis approaches have been proposed so far for multimedia analysis[13]. Latent Dirichlet Allocation (LDA) [4], initially developed for text classification, is a recently proposed approach within this framework that produced good analysis and modeling results. LDA uses the following structures: 1) a finite universal vocabulary $\mathcal{W} = \{\underline{w}^1, \underline{w}^2, ..., \underline{w}^V\}$ of $V$ words (i.e. basic units of discrete data). Each $\underline{w}^i$ with $i \in [1..V]$ is a vector where the $i$-th element is 1 and all others 0 (i.e. $\underline{w}^i(i) = 1$ and $\underline{w}^i(j) = 0$ for $i \neq j$ ). For simplicity we will refer to $\underline{w}^i(i)$ as $w^i$. 2) Documents (videos in our case) where each document $\underline{v}$ is a sequence of $N$ words from the universal vocabulary $\mathcal{W}$, $\underline{v} = (\underline{w}_1^{g(1)}, \underline{w}_2^{g(2)}, ..., \underline{w}_N^{g(N)})$, where $g$ is a surjective map $g : [1..N] \rightarrow [1..V]$ and $\underline{w}_i^{g(i)}$ denotes that the $i$-th word in the sequence $\underline{v}$ is the $g(i)$-th word in the vocabulary $\mathcal{W}$. The fact that $g$ is surjective, is because in $\underline{v}$, we can have multiple instances of the word $\underline{w}^i$. 3) A corpus, namely a set of documents $\mathcal{C}_i = \{\underline{v}_1, \underline{v}_2, .., \underline{v}_m\}$, which are relevant to each other i.e., deal with the same topics. The term topic is used to denote the latent vector variables $\underline{z}^i$, which represent probability distributions on sets of facewords.The meaning and use of these variables will be explained in more detail in the following sections.

In the proposed video fingerprinting framework, a word $\underline{w}^i$ is a faceword (i.e. a certain facial image, ideally corresponding to a particular actor) and each video $\underline{v}$ is a document. The universal vocabulary is the set of all facewords $\mathcal{W}$, as discovered by the face clustering procedure (i.e. the face clusters centers). In our case, a corpus is a set of only one video (i.e. a singleton set) due to the fact that we need to retrieve the same and not simply similar videos. This assumption is not to be confused with the topics. In our case, we assume that a video may be generated from several topics but this distribution of topics is unique for each video.

## 3. LATENT DIRCHLET ALLOCATION FOR VIDEO FINGERPRINTING

As briefly mentioned in Section 1, LDA consists of a generative probabilistic model. The graphical model of LDA is shown in Figure 1.

In our case, we are dealing with videos and thus we aim to use LDA to reveal the latent aspects of a video, based on actors appearances. As already explained, the latent aspects (topics) we refer to are essentially faceword distributions. The motivation behind the adopted approach stems from the fact that the distribution of actor face appearances throughout a movie can provide a description of a video clip or a movie, which will be robust enough to be used in video fingerprinitng.

The LDA probabilistic model consists of the following generative process that creates a video $\underline{v}$ made up of a sequence of $N$ facewords $(\underline{w}_1^{g(1)}, \underline{w}_2^{g(2)}, .., \underline{w}_N^{g(N)})$, where each $\underline{w}_i^{g(i)}$ is drawn from a topic distribution:

- Choose $N$ from a Poisson distribution $Poisson(\xi)$
- Choose a $K$-dimensional random vector variable $\underline{\theta} = [\theta_1, \theta_2, ..., \theta_K]$ from a Dirichlet distribution: $\underline{\theta} \sim Dir(\underline{\alpha})$, where $\underline{\alpha}$ is the vector hyperparameter of the prior Dirichlet distribution.
- For each of the $N$ facewords $\underline{w}_n^{g(n)}$:
  - Choose a topic $\underline{z}_n^{h(n)}$ from a multinomial distribution [14] parametrized with $\underline{\theta}$. $\underline{z}_n^{h(n)} \sim Multinomial(\underline{\theta})$, where $\underline{\theta}$ is a Dirichlet distributed vector variable and $h$ a surjective map $h : [1..N] \rightarrow [1..K]$ which provides that the $n$-th word is conditioned from the $h(n)$-th topic in the latent topics set of cardinality $K$. The fact that $h$ is surjective has the same explanation as for $g$.
  - Choose a word $\underline{w}_n^{g(n)}$ from $p(\underline{w}_n^{g(n)}|\underline{z}_n^{h(n)}, \underline{\beta})$, which will also be a multinomial distribution.

The above generative process, suggests that each faceword is generated with a probability conditioned on a topic (the latent variable). The topics, in turn, are generated from a multinomial distribution with a Dirichlet prior (i.e. $\underline{\theta}$), which is an assumption based on the fact that the Dirichlet distribution is a conjugate prior to the multinomial distribution and thus the most natural choice for a prior [5]. The dimensionality $K$ of the multinomial latent variable $\underline{z}_n^{h(n)}$ can not be known a-priori, and furthermore, no methods for its estimetion exist.In general, defining the dimensionality of the latent variable in the LDA model is still an open issue and certainly beyond the scope of this paper.

Let us suppose that we fix the dimensionality of the topic variable to $K$, and thus, the latent set of topics $\mathcal{Z}$ contains $K$ distinct topics $\mathcal{Z} = \{\underline{z}^1, \underline{z}^2, ..., \underline{z}^K\}$ where $\underline{z}^i$ is a vector where the $i$-th element is 1 and all others 0 (i.e. $\underline{z}^i(i) = 1$ and $\underline{z}^i(j) = 0$ for $i \neq j$ ). For simplicity $\underline{z}^i(i)$ will be denoted as $z^i$. A $K-$dimensional Dirichlet random vector variable $\underline{\theta}$ is chosen from a distribution with probability density function:

$$p(\underline{\theta}|\underline{\alpha}) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_K^{\alpha_K - 1}, \quad (1)$$

where $\underline{\theta}$ lies in a $(K-1)$-simplex (due to the fact that $\theta_i \geq 0$ and $\sum_{i=1}^{K} \theta_i = 1$), $\underline{\alpha}$ is the $K$-dimensional Dirichlet vector hyperparameter with $\alpha_i > 0$ and $\Gamma(x)$ is the Gamma function.

The $K \times N$ parameter matrix $\underline{\beta}$ contains the probabilities $\beta(i,j)$ that the faceword $\underline{w}^j$ is generated from topic $\underline{z}^i$. The parameter matrix $\underline{\beta}$ is not known and has to be estimated, as we will demonstrate later on, from a variational EM algorithm. Given the hyperparameter $\underline{\alpha}$ and the matrix parameter $\underline{\beta}$ we can calculate the joint distribution of a topic mixture $\underline{\theta}$, a set of $N$ topics $\mathbf{Z} = (\underline{z}_1^{h(1)}, \underline{z}_2^{h(2)}, ..., \underline{z}_N^{h(N)})$ and a video $\underline{v}$ (sequence of $N$ words) by:

$$p(\underline{\theta}, \mathbf{Z}, \underline{v}|\underline{\alpha}, \underline{\beta}) = p(\underline{\theta}|\underline{\alpha}) \prod_{n=1}^{N} p(\underline{z}_n^{h(n)}|\underline{\theta}) p(\underline{w}_n^{g(n)}|\underline{z}_n^{h(n)}, \underline{\beta}) \quad (2)$$

By integrating (2) over $\underline{\theta}$ and summing over $\underline{z}_n^{h(n)}$, we obtain the marginal distribution for a video $\underline{v}$:

$$p(\underline{v}|\underline{\alpha}, \underline{\beta}) = \int p(\underline{\theta}|\underline{\alpha}) \left( \prod_{n=1}^{N} \sum_{n=1}^{N} p(\underline{z}_n^{h(n)}|\underline{\theta}) p(\underline{w}_n^{g(n)}|\underline{z}_n^{h(n)}, \underline{\beta}) \right) d\underline{\theta} \quad (3)$$

### 3.1. Training Through Variational Inference

Training the model involves in fact solving the inference problem of computing the posterior distribution of the hidden vectors $\underline{\theta}$ and $\underline{z}_n$ given a video $\underline{v}$ and the Dirichlet parameters $\underline{\alpha}$ and $\underline{\beta}$:

$$p(\underline{\theta}, \mathbf{Z}|\underline{v}, \underline{\alpha}, \underline{\beta}) = \frac{p(\underline{\theta}, \mathbf{Z}, \underline{v}|\underline{\alpha}, \underline{\beta})}{p(\underline{v}|\underline{\alpha}, \underline{\beta})} \quad (4)$$

Unfortunately, the computation of this distribution is in general intractable due to $p(\underline{v}|\underline{\alpha}, \underline{\beta})$. However, a wide variety of approximate inference algorithms can be used to this end, including Laplace approximation, variational approximation, and several Markov chain Monte Carlo methods [15]. In our case, we use a variational inference method as in [4]. Two variational parameters $\underline{\phi}$ and $\underline{\gamma}$ are inserted and thus we obtain a family of distributions of the latent variables.

$$q(\underline{\theta}, \mathbf{Z}|\underline{\gamma}, \underline{\phi}) = q(\underline{\theta}|\underline{\gamma}) \prod_{n=1}^{N} q(z_n, \underline{\phi}_n), \quad (5)$$

where $\underline{\gamma}$ is a $K$-dimensional Dirichlet distributed parameter vector and $(\underline{\phi}_1, \underline{\phi}_2, .., \underline{\phi}_N)$ are vectors of multinomial distributed parameters. In [4], it is proven that the values $\underline{\gamma}^*$ and $\underline{\phi}^*$ that lead to a tight lower bound on the log likelihood can be evaluated through the following optimization problem:

$$(\underline{\gamma}^*, \underline{\phi}^*) = \arg\min_{(\underline{\gamma}, \underline{\phi})} KL(q(\underline{\theta}, \mathbf{Z}|\underline{\gamma}, \underline{\phi}) \| p(\underline{\theta}, \mathbf{Z}, \underline{v}|\underline{\alpha}, \underline{\beta})), \quad (6)$$

where $KL$ is the Kulback-Leibler divergence [14]. The optimization procedure is described in [4].

We note that $\underline{\gamma}^*$ is a function of $\underline{v}$ due to the fact that (6) is evaluated for fixed $\underline{v}$, and thus, provides a unique representation of a video from the training set, in the simplex formed from the topics. In other words, each training video is uniquely characterized as a point in this $(K-1)$-simplex.

The parameters $\underline{\alpha}$ and $\underline{\beta}$, involved in the model, are estimated by training our model. To do so, we follow the approach in [4] which is an empirical Bayes method and consists of the following EM process:

- E-step: For each video, find the optimal values of the variational parameters $\underline{\phi}^*$, $\underline{\gamma}^*$. This estimation step uses the aforementioned methodology for fixed values of $\underline{\alpha}$ and $\underline{\beta}$
- (M-step) Maximize the resulting lower bound on the likelihood with respect to parameters $\underline{\alpha}$ and $\underline{\beta}$.

### 3.2. Video Fingerprinting Using LDA

Assuming that the parameters $\underline{\alpha}$, $\underline{\beta}$ have been estimated from the training set, we need to develop a method for finding if a video, introduced as a query to the database, is a replica or not. The video is first subjected to face detection and, then, each of the detected facial images is assigned to one of the face clusters, established offline for the entire database, using the face clustering algorithm. Thus the query video is represented as a sequence of words $\underline{v}_{query} = (\underline{w}_1^{g(1)}, \underline{w}_2^{g(2)}, .., \underline{w}_N^{g(N)})$.

The query video is then characterized by the $K$-dimensional parameter $\underline{\gamma}^*$ which is an estimate of the mixture of topics distributions $p(\underline{\theta}, \mathbf{Z}|\underline{v}, \underline{\alpha}, \underline{\beta})$ in this video and is found via inference with the trained model using (6). Thus $\underline{\gamma}^*$ is used as the feature vector (i.e. the fingerprint) of the query video.

In order to decide whether a video $\underline{v}_{query}$ is a replica of one of the videos in the database, we use the KL divergence between its variational parameter $\underline{\gamma}^*$ and the ones of the videos stored in the database. By doing so, we find the index $F$ of the closest database video:

$$F = \arg\min_i KL(\underline{\gamma}^*(\underline{v}_i) \| \underline{\gamma}^*(\underline{v}_{query})) \quad (7)$$

where $\underline{v}_i$ is $i$-th video in the database.

Besides matching query videos to the ones in the database, we also need to handle videos that are not replicas of the ones in the database. This is done in a two-level process. First, query videos whose facial images do not match any (or match only few) facial images stored in the database (universal) vocabulary are characterized as having no match in the database. In case a video has enough face matches with the database vocabulary (typically more than 20 in all clusters) we decide that the query video has a match in the database only if the KL divergence in (7) is below a certain threshold $T_{KL}$. This threshold is experimentally found by introducing into the system query videos that are not in the database and have enough face matches with the database, and videos in the database. Through this experiment the threshold that minimizes the false acceptance and false rejection ratios was found to be equal to $T_{KL} = 0.95$.

## 4. EXPERIMENTAL RESULTS

The performance of the method has been evaluated on a video set that includes short, low quality videos, randomly selected over the Internet. It consists of 332 videos, each 2-5 minutes long (approximately 4000-7000 frames per video clip).

In this video set we have applied face detection every 10 frames. Even at this face detection rate (one every approximately 0.5 sec) we are almost sure that we will detect all actor faces involved in the videos. For this data set the length of the universal vocabulary was 1088 (1088 different facewords, i.e. face clusters, were created).

In order to evaluate the performance of the proposed method, two different types of experiments must be performed:

- Tests with query videos that are replicas of the database videos (test A). Two types of errors are expected in this case: a misclassification error (MC), measured by the percentage of query videos that were classified to a wrong original video in the database and the false rejection error (FR), which is the percentage of query videos that were erroneously tagged as not belonging to the database.
- Tests with query videos that do not belong to the database (test B). In this case, the performance is measured in terms of the false acceptance (FA) error, i.e., the percentage of query videos

that are erroneously tagged as being a replica of a database video.

The experiments aimed at showing evaluating the system performance when queried with videos that are identical (in the test A) with those in the database, i.e. they have not been manipulated. In this set of experiments, the test A involved 332 videos which were used to both populate the system database (and train the system) and as query videos. For the test B we trained the model and populated the system database with 247 videos out of the original 332 and used the rest 85 videos as query videos for testing. Results are depicted in Table 1.

**Table 1**. Fingerprinting Performance Metrics

| | TEST A | | TEST B |
|---|---|---|---|
| | MC | FR | FA |
| $\mathcal{VC}$ | 2.11% | 1.2% | 0% |

As it can be seen, the MC, FR errors are sufficiently low, where FA is zero. The false acceptance (FA) rate can be further analyzed due to the fact that, as already mentioned in the previous section, declaring that a video is not a replica of a video in the database is a two step process. In our experiments (test B), out of the 85 non-replica videos that were used for querying the database, only 4 of them (that is 4.7%) were rejected in the first step, i.e. due to the small number of face mathces with the vocabulary.

Some preliminary experiments with query videos that have been modified by histogram equalization, temporal cropping, removal of random frames and spatial cropping have also been performed. Results were very promising showing robustness to such attacks.

## 5. CONCLUSIONS AND FUTURE WORK

In this work, a new framework for video fingerprinting has been presented. The intuition behind this work is that facewords can carry a significant amount of information and can be used to capture very distinctive video features, thus characterizing uniquely each video. By applying a generative probabilistic model, namely the Latent Dirichlet Allocation, in this context, we aim at discovering latent aspects of a video based on the semantic information related to the actors appearance. The distribution of these latent video aspects, for each video, can be used effectively to discriminate and match videos in a database for video fingerprinting applications, as shown in the experimental results.

In the future, more thorough experimental testing will be performed. In addition, we will further explore this approach by using a more complex vocabulary including e.g. human pose, human interactions etc. By doing so we believe to provide a better representation for movie topics, and thus a more robust and discriminative fingerprinting algorithm

## 6. ACKNOWLEDGMENT

This work has been partially supported by the European Commission through the IST Programme under Contract IST-2002-507932 ECRYPT.

## 7. REFERENCES

[1] D. Kundur and K. Karthik, "Video Fingerprinting and Encryption Principles for Digital Rights Management," *Proceedings of the IEEE*, vol. 92, no. 6, pp. 918–932, 2004.

[2] P. Indyk, G. Iyengar, and N. Shivakumar, "Finding pirated video sequences on the internet," *Tecnical Report, Stanford University, 1999*.

[3] Job Oostveen, Ton Kalker, and Jaap Haitsma, "Feature extraction and a database strategy for video fingerprinting," in *Proceedings of the 5th International Conference on Recent Advances in Visual Information Systems VISUAL '02*, London, UK, 2002, pp. 117–128, Springer-Verlag.

[4] D.M. Blei, A.Y. Ng, M.I. Jordan, and J. Lafferty, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, no. 4-5, pp. 993–1022, 2003.

[5] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proceedings of International Conference of Computer Vision and Pattern Recognition. CVPR '05*, 2005, vol. 5.

[6] J.C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words," *International Journal of Computer Vision*, pp. 1–20, 2008.

[7] J. Cao, J. Li, Y. Zhang, and S. Tang, "LDA-Based Retrieval Framework for Semantic News Video Retrieval," 2007, pp. 155–160, IEEE Computer Society Washington, DC, USA.

[8] M. Fleischman, "Unsupervised content-based indexing of sports video," 2007, pp. 87–94, ACM Press New York, NY, USA.

[9] M. Heritier, S. Foucher, and L. Gagnon, "Key-Places Detection and Clustering in Movies Using Latent Aspects," in *Proceedings of Internations Conference on Image Processing, ICIP '07*, 2007, vol. 2.

[10] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2002.

[11] D.G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of International Conference on Computer Vision*. 1999, vol. 2, pp. 1150–1157, Kerkyra, Greece.

[12] P. Antonopoulos, N. Nikolaidis, and I. Pitas, "Hierarchical Face Clustering using SIFT Image Features," in *Computational Intelligence in Image and Signal Processing, 2007. CIISP 2007. IEEE Symposium on*, 2007, pp. 325–329.

[13] C.G.M. Snoek and M. Worring, "Multimodal Video Indexing: A Review of the State-of-the-art," *Multimedia Tools and Applications*, vol. 25, no. 1, pp. 5–35, 2005.

[14] A. Papoulis and S.U. Pillai, *Probability, random variables, and stochastic processes*, McGraw-Hill New York, 1991.

[15] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul, "An Introduction to Variational Methods for Graphical Models," *Machine Learning*, vol. 37, no. 2, pp. 183–233, 1999.