Learning Hidden Markov Models from Pairwise Co-occurrences with Application to Topic Modeling

Kejun Huang [1] **Xiao Fu** [2] **Nicholas D. Sidiropoulos** [3]

Abstract

We present a new algorithm for identifying the transition and emission probabilities of a hidden Markov model (HMM) from the emitted data. Expectation-maximization becomes computationally prohibitive for long observation records, which are often required for identification. The new algorithm is particularly suitable for cases where the available sample size is large enough to accurately estimate second-order output probabilities, but not higher-order ones. We show that if one is only able to obtain a reliable estimate of the pairwise co-occurrence probabilities of the emissions, it is still possible to uniquely identify the HMM if the emission probability is *sufficiently scattered*. We apply our method to hidden topic Markov modeling, and demonstrate that we can learn topics with higher quality if documents are modeled as observations of HMMs sharing the same emission (topic) probability, compared to the simple but widely used bag-of-words model.

1. Introduction

Hidden Markov models (HMMs) are widely used in machine learning when the data samples are time *dependent*, for example in speech recognition, language processing, and video analysis. The graphical model of a HMM is shown in Figure 1. HMM models a (time-dependent) sequence of data $\{Y_t\}_{t=0}^T$ as indirect observations of an underlying Markov chain $\{X_t\}_{t=0}^T$ which is not available to us. Homogeneous HMMs are parsimonious models, in the sense that they are fully characterized by the transition probability $\Pr[X_{t+1}|X_t]$ and the emission probability $\Pr[Y_t|X_t]$ even though the size of the given data $\{Y_t\}_{t=0}^T$ can be very large.

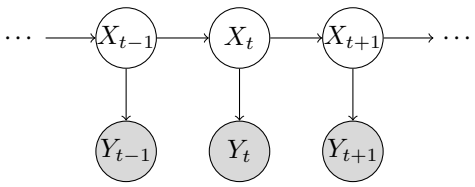


Figure 1: The graphical model of a HMM.

Consider a homogeneous HMM such that:

- a latent variable X_t can take K possible outcomes $x_1, ..., x_K$;
- an ambient variable Y_t can take N possible outcomes $y_1, ..., y_N$.

Recall that (Rabiner & Juang, 1986; Ghahramani, 2001):

- Given both $\{X_t\}_{t=0}^T$ and $\{Y_t\}_{t=0}^T$, the complete joint probability factors, and we can easily estimate the transition probability $\Pr[X_{t+1}|X_t]$ and the emission probability $\Pr[Y_t|X_t]$.
- Given only $\{Y_t\}_{t=0}^T$, but assuming we know the underlying transition and emission probabilities, we can calculate the observation likelihood using the forward algorithm, estimate the most likely hidden sequence using the Viterbi algorithm, and compute the posterior probability of the hidden states using the forward-backward algorithm.

The most natural problem setting, however, is when neither the hidden state sequence nor the underlying probabilities are known to us—we only have access to a sequence of observations, and our job is to reveal the HMM structure, characterized by the transition matrix $\Pr[X_{t+1}|X_t]$ and the emission probability $\Pr[Y_t|X_t]$ from the set of observations $\{Y_t\}_{t=0}^T$.

1.1. Related work

The traditional way of learning a HMM from $\{Y_t\}_{t=0}^T$ is via expectation-maximization (EM) (Rabiner & Juang, 1986), in which the expectation step is performed by calling the forward-backward algorithm. This specific instance of EM is also called the Baum-Welch algorithm (Baum et al., 1970; Ghahramani, 2001). However, the complexity of Baum-

[1]University of Minnesota, Minneapolis, MN 55455 [2]Oregon State University, Corvallis, OR 97331 [3]University of Virginia, Charlottesville, VA 22904. Correspondence to: Kejun Huang <huang663@umn.edu>, Xiao Fu <xiao.fu@oregonstate.edu>, Nicholas D. Sidiropoulos <nikos@virginia.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

Welch is prohibitive when T is relatively large—the complexity of the forward-backward algorithm is $\mathcal{O}(K^2T)$, but EM converges slowly, so the forward-backward algorithm must be called many times. This is a critical issue, because a HMM can only be learned with high accuracy when the number of observation samples T is large enough.

One way of designing scalable algorithms for learning HMMs is to work with sufficient statistics—a summary of the given observation sequence, whose size does not grow with T. Throughout this paper we assume that the HMM process is stationary (time-invariant), which is true almost surely if the underlying Markov process is ergodic and the process has been going on for a reasonable amount of time. With T large enough, we can accurately estimate the co-occurrence probability between two consecutive emissions $\Pr[Y_t, Y_{t+1}]$. According to the graphical model shown in Figure 1, it is easy to see that given the value of X_t, Y_t is conditionally independent of all the other variables, leading to the factorization

$$\Pr[Y_t, Y_{t+1}] = \sum_{k,j=1}^{K} \Pr[Y_t|X_t = x_k] \Pr[Y_{t+1}|X_{t+1} = x_j] \times \Pr[X_t = x_k, X_{t+1} = x_j] \quad (1)$$

Let $\boldsymbol{\Omega} \in \mathbb{R}^{N\times N}$, $\boldsymbol{M} \in \mathbb{R}^{N\times K}$, and $\boldsymbol{\Theta} \in \mathbb{R}^{K\times K}$, with their elements defined as

$$\begin{aligned} \Omega_{n\ell} &= \Pr[Y_t = y_n, Y_{t+1} = y_\ell], \\ M_{nk} &= \Pr[Y_t = y_n|X_t = x_k], \\ \Theta_{kj} &= \Pr[X_t = x_k, X_{t+1} = x_j]. \end{aligned}$$

Then, equations (1) can be written compactly as

$$\boldsymbol{\Omega} = \boldsymbol{M}\boldsymbol{\Theta}\boldsymbol{M}^\top. \quad (2)$$

Noticing that (2) is a nonnegative matrix tri-factorization with a number of inconsequential constraints for $\boldsymbol{M}$ and $\boldsymbol{\Theta}$ to properly represent probabilities, Vanluyten et al. (2008); Lakshminarayanan & Raich (2010); Cybenko & Crespi (2011) proposed using nonnegative matrix factorization (NMF) to estimate the HMM probabilities. However, NMF-based methods have a serious shortcoming in this context: the tri-factorization (2) is in general not unique, because it is fairly easy to find a nonsingular matrix $\boldsymbol{Q}$ such that both $\boldsymbol{M}\boldsymbol{Q} \geq 0$ and $\boldsymbol{Q}^{-1}\boldsymbol{\Theta}\boldsymbol{Q}^{-\top} \geq 0$, and then $\widetilde{\boldsymbol{M}} = \boldsymbol{M}\boldsymbol{Q}$ and $\widetilde{\boldsymbol{\Theta}} = \boldsymbol{Q}^{-1}\boldsymbol{\Theta}\boldsymbol{Q}^{-\top}$ are equally good solutions in terms of reconstructing the co-occurrence matrix $\boldsymbol{\Omega}$. When we use $(\boldsymbol{M}, \boldsymbol{\Theta})$ and $(\widetilde{\boldsymbol{M}}, \widetilde{\boldsymbol{\Theta}})$ to perform HMM inference, such as estimating hidden states or predicting new emissions, the two models often yield completely different results, unless $\boldsymbol{Q}$ is a permutation matrix.

A number of works propose to use *tensor* methods to overcome the identifiability issue. Instead of working with the pairwise co-occurrence probabilities, they start by estimating the joint probabilities of three consecutive observations $\Pr[Y_{t-1}, Y_t, Y_{t+1}]$. Noticing that these three random variables are conditionally independent given X_t, the triple-occurrence probability factors into

$$\Pr[Y_{t-1}, Y_t, Y_{t+1}] = \sum_{k=1}^{K} \Pr[X_t = x_k] \Pr[Y_{t-1}|X_t = x_k] \times \Pr[Y_t|X_t = x_k] \Pr[Y_{t+1}|X_t = x_k],$$

which admits a tensor *canonical polyadic decomposition* (CPD) model (Hsu et al., 2009; Anandkumar et al., 2012; 2014). Assuming $K \leq N$, the CPD is essentially unique if two of the three factor matrices have full column rank, and the other one is not rank one (Harshman, 1970); in the context of HMMs, this is equivalent to assuming $\boldsymbol{M}$ and $\boldsymbol{\Theta}$ both have linearly independent columns, which is a relatively mild condition. The CPD is known to be unique under much more relaxed conditions (Sidiropoulos et al.), but in order to uniquely retrieve the transition probability using the relationship

$$\Pr[Y_{t+1}|X_t] = \sum_{j=1}^{K} \Pr[Y_{t+1}|X_{t+1}=x_j] \Pr[X_{t+1}=x_j|X_t],$$

$K \leq N$ is actually the best we can achieve using triple-occurrences without making further assumptions. [1] A salient feature in this case is that if the triple-occurrence probability $\Pr[Y_{t-1}, Y_t, Y_{t+1}]$ is exactly given (meaning the rank of the triple-occurrence tensor is indeed smaller than N), the CPD can be efficiently calculated using generalized eigendecomposition and related algebraic methods (Sanchez & Kowalski, 1990; Leurgans et al., 1993; De Lathauwer et al., 2004). These methods do not work well, however, when the low-rank tensor is perturbed; e.g., due to insufficient mixing / sample averaging of the triple occurrence probabilities.

It is also possible to handle cases where $K > N$. The key observation is that, given X_t, Y_t is conditionally independent of $Y_{t-1}, ..., Y_{t-\tau}$ and $Y_{t+1}, ..., Y_{t+\tau}$. Then, grouping $Y_{t-1}, ..., Y_{t-\tau}$ into a single categorical variable taking N^τ possible outcomes, and $Y_{t+1}, ..., Y_{t+\tau}$ into another one, we can construct a much bigger tensor of size $N^\tau \times N^\tau \times N$, and then uniquely identify the underlying HMM structure with $K \gg N$ as long as certain linear independence requirements are satisfied for the conditional distribution of the *grouped* variables (Allman et al., 2009; Bhaskara et al., 2014; Huang et al., 2016b; Sharan et al., 2017). It is intu-

[1] In the supplementary material, we prove that if the emission probability is *generic* and the transition probability is *sparse*, the HMM can be uniquely identified from triple-occurrence probability for $K < N^2/16$ using the latest tensor identifiability result (Chiantini & Ottaviani, 2012).

itively clear that for fixed N, we need a much larger realization length T in order to accurately estimate $(2\tau+1)$-occurrence probabilities as τ grows, which is the price we need to pay for learning a HMM with a larger number of hidden states.

1.2. This paper

The focus of this paper is on cases where $K \leq N$, and T is large enough to obtain accurate estimate of $\Pr[Y_t, Y_{t+1}]$, but not large enough to accurately estimate triple or higher-order occurrence probabilities. We *prove* that it is actually possible to recover the latent structure of an HMM only from pairwise co-occurrence probabilities $\Pr[Y_t, Y_{t+1}]$, provided that the underlying emission probability $\Pr[Y_t|X_t]$ is *sufficiently scattered*. Compared to the existing NMF-based HMM learning approaches, our formulation employs a different (determinant-based) criterion to ensure identifiability of the HMM parameters. Our matrix factorization approach resolves cases that cannot be handled by tensor methods, namely when T is insufficient to estimate third-order probabilities, under an additional condition that is quite mild: that the emission probability matrix $\boldsymbol{M}$ must be *sufficiently scattered*, rather than simply full column-rank.

We apply our method to hidden topic Markov modeling (HTMM) (Gruber et al., 2007), in which case the number of hidden states (topics) is indeed much smaller than the number of ambient states (words). HTMM goes beyond the simple and widely used bag-of-words model by assuming that (ordered) words in a document are emitted from a hidden topic sequence that evolves according to a Markov model. We show improved performance on real data when using this simple and intuitive model to take word ordering into account when learning topics, which also benefits from our identifiability guaranteed matrix factorization method.

As an illustrative example, we showcase the inferred topic of each word in a news article (removing stop words) in Figure 2, taken from the Reuters21578 data set obtained at (Mimaroglu, 2007). As we can see, HTMM gets much more consistent and smooth inferred topics compared to that obtained from a bag-of-words model (cf. supplementary material for details). This result agrees with human understanding.

2. Second-order vs. Third-order Learning

We start by arguing that for the same observation data $\{Y_t\}_{t=0}^T$, the estimate of the pairwise co-occurrence probability $\Pr[Y_t, Y_{t+1}]$ is always more accurate than that of the triple co-occurrence probability $\Pr[Y_{t-1}, Y_t, Y_{t+1}]$.

Let us first explicitly describe the estimator we use for these probabilities. For each observation Y_t, we define a coordinate vector $\boldsymbol{\psi}_t \in \mathbb{R}^K$, and $\boldsymbol{\psi}_t = \boldsymbol{e}_k$ if $Y_t = y_k$. The natural

china daily vermin eat pct grain stocks survey provinces and cities showed vermin consume and pct china grain stocks china daily each year mln tonnes pct china fruit output left rot and mln tonnes pct vegetables paper blamed waste inadequate storage and bad preservation methods government had launched national programme reduce waste calling for improved technology storage and preservation and greater production additives paper gave details

china daily vermin eat pct grain stocks survey provinces and cities showed vermin consume and pct china grain stocks china daily each year mln tonnes pct china fruit output left rot and mln tonnes pct vegetables paper blamed waste inadequate storage and bad preservation methods government had launched national programme reduce waste calling for improved technology storage and preservation and greater production additives paper gave details

Figure 2: Inferred topics of the words shown in different colors, obtained by probabilistic latent semantic analysis (top) and hidden topic Markov model (bottom).

estimator for the pairwise co-occurrence probability matrix $\boldsymbol{\Omega}$ is

$$\widehat{\boldsymbol{\Omega}} = \frac{1}{T}\sum_{t=0}^{T-1} \boldsymbol{\psi}_t \boldsymbol{\psi}_{t+1}^\top, \tag{3}$$

and similarly for the triple co-occurrence probability $\underline{\boldsymbol{\Omega}_3}$

$$\widehat{\underline{\boldsymbol{\Omega}_3}} = \frac{1}{T-1}\sum_{t=1}^{T-1} \boldsymbol{\psi}_{t-1} \circ \boldsymbol{\psi}_t \circ \boldsymbol{\psi}_{t+1}, \tag{4}$$

where $\circ$ denotes vector outer-product. [2]

The first observation is that both $\widehat{\boldsymbol{\Omega}}$ and $\widehat{\underline{\boldsymbol{\Omega}_3}}$ are unbiased estimators: Obviously $\mathrm{E}(\boldsymbol{\psi}_t \boldsymbol{\psi}_{t+1}^\top) = \boldsymbol{\Omega}$ and likewise for the triple-occurrences, and taking their averages does not change the expectation. However, the individual terms in the summation are not independent of each other, making it hard to determine how fast estimates converge to their expectation. The state-of-the-art concentration result for HMMs (Kontorovich, 2006) states that for any 1-Lipschitz function f

$$\Pr[|f(\{Y_t\}) - \mathrm{E}\, f(\{Y_t\})| > \epsilon] \leq 2\exp\left(-T\epsilon^2/c\right),$$

where c is a constant that only depends on the specific HMM structure but not on the function f (cf. (Kontorovich, 2006) for details). Taking f as any entry in $\widehat{\boldsymbol{\Omega}}$ or $\widehat{\underline{\boldsymbol{\Omega}_3}}$, we can check that indeed it is 1-Lipschitz, meaning as T goes to infinity, both estimators converge to their expectation with negligible fluctuations.

We now prove that for a given set of observations $\{Y_t\}_{t=0}^T$, $\widehat{\boldsymbol{\Omega}}$ is always going to be more accurate than $\widehat{\underline{\boldsymbol{\Omega}_3}}$. Since both of them represent probabilities, we use two common metrics to measure the differences between the estimators and their expectations, the Kullback-Leibler divergence $D_{\mathrm{KL}}(\cdot)$ and the total-variation difference $D_{\mathrm{TV}}(\cdot)$.

[2] In some literature $\circ$ is written as the Kronecker product $\otimes$. Strictly speaking, the Kronecker product of three vectors is a very long vector, not a three-way array. For this reason, we chose to use $\circ$ instead of $\otimes$.

Proposition 1. *Let $\widehat{\boldsymbol{\Omega}}$ and $\widehat{\underline{\boldsymbol{\Omega}_3}}$ be obtained from the same set of observations $\{Y_t\}_{t=0}^T$, we have that*

$$D_{\mathrm{KL}}(\widehat{\boldsymbol{\Omega}}\|\boldsymbol{\Omega}) \leq D_{\mathrm{KL}}(\widehat{\underline{\boldsymbol{\Omega}_3}}\|\underline{\boldsymbol{\Omega}_3}) \qquad \textit{and}$$
$$D_{\mathrm{TV}}(\widehat{\boldsymbol{\Omega}}\|\boldsymbol{\Omega}) \leq D_{\mathrm{TV}}(\widehat{\underline{\boldsymbol{\Omega}_3}}\|\underline{\boldsymbol{\Omega}_3}).$$

The proof of Proposition 1 is relegated to the supplementary material.

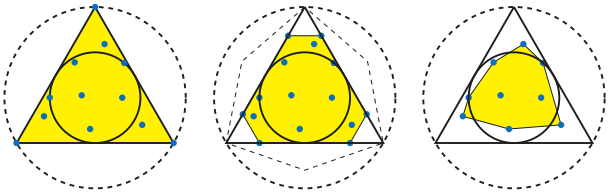
3. Identifiability of HMMs from Pairwise Co-occurrence Probabilities

The arguments made in the previous section motivate going back to matrix factorization methods for learning a HMM when the realization length T is not large enough to obtain accurate estimates of triple co-occurrence probabilities. As we have explained in §1.1, the co-occurrence probability matrix $\boldsymbol{\Omega}$ admits a nonnegative matrix tri-factorization model (2). There are a number of additional equality constraints. Columns of $\boldsymbol{M}$ represent conditional distributions, so $\boldsymbol{1}^\top\boldsymbol{M} = \boldsymbol{1}^\top$. Matrix $\boldsymbol{\Theta}$ represents the joint distribution between two consecutive Markovian variables, therefore $\boldsymbol{1}^\top\boldsymbol{\Theta}\boldsymbol{1} = 1$. Furthermore, we have that $\boldsymbol{\Theta}\boldsymbol{1}$ and $\boldsymbol{\Theta}^\top\boldsymbol{1}$ represent $\Pr[X_t]$ and $\Pr[X_{t+1}]$ respectively, and since we assume that the Markov chain is stationary, they are the same, i.e., $\boldsymbol{\Theta}\boldsymbol{1} = \boldsymbol{\Theta}^\top\boldsymbol{1}$. Notice that this does not imply that $\boldsymbol{\Theta}$ is symmetric, and in fact it is often not symmetric.

Huang et al. (2016a) considered a factorization model similar to (2) in a different context, and showed that identifiability can be achieved under a reasonable assumption called *sufficiently scattered*, defined as follows.

Definition 1 (sufficiently scattered)**.** Let $\mathrm{cone}(\boldsymbol{M}^\top)^*$ denote the polyhedral cone $\{\boldsymbol{x} : \boldsymbol{M}\boldsymbol{x} \geq 0\}$, and $\mathcal{C}$ denote the elliptical cone $\{\boldsymbol{x} : \|\boldsymbol{x}\| \leq \boldsymbol{1}^\top\boldsymbol{x}\}$. Matrix $\boldsymbol{M}$ is called **sufficiently scattered** if it satisfies that: (i) $\mathrm{cone}(\boldsymbol{M}^\top)^* \subseteq \mathcal{C}$, and (ii) $\mathrm{cone}(\boldsymbol{M}^\top)^* \cap \mathrm{bd}\mathcal{C} = \{\lambda\boldsymbol{e}_k : \lambda \geq 0, k = 1, ..., K\}$, where $\mathrm{bd}\mathcal{C}$ denotes the boundary of $\mathcal{C}$, $\{\boldsymbol{x} : \|\boldsymbol{x}\| = \boldsymbol{1}^\top\boldsymbol{x}\}$.

The sufficiently scattered condition was first proposed in (Huang et al., 2014) to establish uniqueness conditions for the widely used *nonnegative matrix factorization* (NMF). For the NMF model $\boldsymbol{\Omega} = \boldsymbol{W}\boldsymbol{H}^\top$, if both $\boldsymbol{W}$ and $\boldsymbol{H}$ are sufficiently scattered, then the nonnegative decomposition is unique up to column permutation and scaling. Follow up work strengthened and extended the identifiability results based on this geometry inspired condition (Fu et al., 2015; Huang et al., 2016a; Fu et al., 2018). A similar tri-factorization model was considered in (Huang et al., 2016a) in the context of bag-of-words topic modeling, and it was shown that among all feasible solutions of (2), if we find one that minimizes $|\det\boldsymbol{\Theta}|$, then it recovers the ground-truth latent factors $\boldsymbol{M}$ and $\boldsymbol{\Theta}$, assuming the ground-truth $\boldsymbol{M}$ is sufficiently scattered. In our present context, we therefore propose the following problem formulation:



(a) Separable (b) Sufficiently scattered (c) Not identifiable

Figure 3: A geometric illustration of the sufficiently scattered condition (middle), a special case that is separable (left), and a case that is not identifiable (right).

$$\underset{\boldsymbol{\Theta},\boldsymbol{M}}{\text{minimize}} \quad |\det\boldsymbol{\Theta}| \tag{5a}$$
$$\text{subject to} \quad \boldsymbol{\Omega} = \boldsymbol{M}\boldsymbol{\Theta}\boldsymbol{M}^\top, \tag{5b}$$
$$\boldsymbol{\Theta} \geq 0, \boldsymbol{\Theta}\boldsymbol{1} = \boldsymbol{\Theta}^\top\boldsymbol{1}, \boldsymbol{1}^\top\boldsymbol{\Theta}\boldsymbol{1} = 1, \tag{5c}$$
$$\boldsymbol{M} \geq 0, \boldsymbol{1}^\top\boldsymbol{M} = \boldsymbol{1}^\top. \tag{5d}$$

Regarding Problem (5), we have the following identifiability result.

Theorem 1. *(Huang et al., 2016a) Suppose $\boldsymbol{\Omega}$ is constructed as $\boldsymbol{\Omega} = \boldsymbol{M}_\natural\boldsymbol{\Theta}_\natural\boldsymbol{M}_\natural^\top$, where $\boldsymbol{M}_\natural$ and $\boldsymbol{\Theta}_\natural$ satisfy the constraints in (5), and in addition (i) $\mathrm{rank}(\boldsymbol{\Theta}_\natural) = K$ and (ii) $\boldsymbol{M}_\natural$ is* **sufficiently scattered**. *Let $(\boldsymbol{M}_\star, \boldsymbol{\Theta}_\star)$ be an optimal solution for (5), then there must exist a permutation matrix $\boldsymbol{\Pi} \in \mathbb{R}^{K\times K}$ such that*

$$\boldsymbol{M}_\natural = \boldsymbol{M}_\star\boldsymbol{\Pi}, \qquad \boldsymbol{\Theta}_\natural = \boldsymbol{\Pi}^\top\boldsymbol{\Theta}_\star\boldsymbol{\Pi}.$$

One may notice that in (Huang et al., 2016a), there are no constraints on the core matrix $\boldsymbol{\Theta}$ as we do in (5c). In terms of identifiability, it is easy to see that if the ground-truth $\boldsymbol{\Theta}_\natural$ satisfies (5c), solving (5) even without (5c) will produce a solution $\boldsymbol{\Theta}_\star$ that satisfies (5c), thanks to uniqueness. In practice when we are given a less accurate $\boldsymbol{\Omega}$, such "redundant" information will help us improve the estimation error, but that goes beyond identifiability consederations.

The proof of Theorem 1 is referred to (Huang et al., 2016a). Here we provide some insights on this geometry-inspired sufficiently scattered condition, and discuss why it is a reasonable (and thus practical) assumption. The notation $\mathrm{cone}(\boldsymbol{M}^\top)^* = \{\boldsymbol{x} : \boldsymbol{M}\boldsymbol{x} \geq 0\}$ comes from the convention in convex analysis that it is the *dual cone* of the conical hull of the row vectors of $\boldsymbol{M}$, i.e., $\mathrm{cone}(\boldsymbol{M}^\top) = \{\boldsymbol{M}^\top\boldsymbol{\alpha} : \boldsymbol{\alpha} \geq 0\}$. Similarly, we can derive that the dual cone of $\mathcal{C}$ is $\mathcal{C}^* = \{\boldsymbol{x} : \|\boldsymbol{x}\| \leq \boldsymbol{1}^\top\boldsymbol{x}/\sqrt{K-1}\}$. A useful property of the dual cone is that for two convex cones $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{A} \subseteq \mathcal{B}$ iff $\mathcal{B}^* \subseteq \mathcal{A}^*$. Therefore, the first requirement of sufficiently scattered in Definition 1 equivalently means

$$\mathcal{C}^* \subseteq \mathrm{cone}(\boldsymbol{M}^\top).$$

We give a geometric illustration of the sufficiently scattered condition in Figure 3b for $K = 3$, and we focus on the 2-dimensional plane $\mathbf{1}^\top \boldsymbol{x} = 1$. The intersection between this plane and the nonnegative orthant is the probability simplex, which is the triangle in Figure 3b. The outer circle represents $\mathcal{C}$, and the inner circle represents $\mathcal{C}^*$, again intersecting with the plane, respectively. The rows of $\boldsymbol{M}$ are scaled to sum up to one, and they are represented by black dots in Figure 3b. Their conical hull is represented by the shaded region. The polygon with dashed lines represents the dual of $\mathrm{cone}(\boldsymbol{M}^\top)$, which is indeed a subset of $\mathcal{C}$, and touches the boundary of $\mathcal{C}$ only at the coordinate vectors.

Figure 3a shows a special case of sufficiently scattered called *separability*, which first appeared in (Donoho & Stodden, 2004) also to establish uniqueness of NMF. In this case, all the coordinate vectors appear in rows of $\boldsymbol{M}$, therefore $\mathrm{cone}(\boldsymbol{M})$ equals the nonnegative orthant. It makes sense that this condition makes the identification problem easier, but it is also a very restrictive assumption. The sufficiently scattered condition, on the other hand, only requires that the shaded region contains the inner circle, as shown in Figure 3b. Intuitively this requires that the rows of $\boldsymbol{M}$ be "well scattered" in the probability simplex, but not to the extent of "separable". Separability-based HMM identification has been considered in (Barlier et al., 2015; Glaude et al., 2015). However, the way they construct second-order statistics is very different from ours. Figure 3c shows a case where $\boldsymbol{M}$ is not sufficiently scattered, and it also happens to be a case where $\boldsymbol{M}$ is not identifiable.

As we can see, the elliptical cone $\mathcal{C}^*$ is tangent to all the facets of the nonnegative orthant. As a result, for $\boldsymbol{M}$ to be sufficiently scattered, it is necessary that there are enough rows of $\boldsymbol{M}$ lie on the boundary of the nonnegative orthant, i.e., $\boldsymbol{M}$ is relatively sparse. Specifically, if $\boldsymbol{M}$ is sufficiently scattered, then each column of $\boldsymbol{M}$ contains at least $K - 1$ zeros (Huang et al., 2014). This is a very important insight, as exactly checking whether a matrix is sufficiently scattered may be computationally hard. In the present paper we further show the following result.

Proposition 2. *The ratio between the volume of the hyperball obtained by intersecting $\mathbf{1}^\top \boldsymbol{x} = 1$ and $\mathcal{C}^*$ and the probability simplex is*

$$\frac{1}{\sqrt{\pi K}} \left(\frac{4\pi}{K(K-1)} \right)^{\frac{K-1}{2}} \Gamma\left(\frac{K}{2}\right). \tag{6}$$

The proof is given in the supplementary material. As K grows larger, the volume ratio (6) goes to zero at a super-exponential decay rate. This implies that the volume of the inner sphere quickly becomes negligible compared to the volume of the probability simplex, as K becomes moderately large. The take home point is that, for a practical choice of K, say $K \geq 10$, as long as $\boldsymbol{M}$ satisfies that each column contains at least K zeros, and the positions of the zeros appear relatively random, it is very likely that it is sufficiently scattered, and thus can be uniquely recovered via solving (5).

4. Algorithm

Our identifiability analysis based on the sufficiently scattered condition poses an interesting non-convex optimization problem (5). First of all, the given co-occurrence probability $\boldsymbol{\Omega}$ may not be exact, therefore it may not be a good idea to put (5b) as a hard constraint. For algorithm design, we propose the following modification to problem (5).

$$\begin{aligned} \underset{\boldsymbol{\Theta}, \boldsymbol{M}}{\text{minimize}} \quad & \sum_{n,\ell=1}^{N} -\Omega_{n\ell} \log \sum_{k,j=1}^{K} M_{nk} \Theta_{kj} M_{\ell j} + \lambda |\det \boldsymbol{\Theta}| \\ \text{subject to} \quad & \boldsymbol{M} \geq 0, \mathbf{1}^\top \boldsymbol{M} = \mathbf{1}^\top, \\ & \boldsymbol{\Theta} \geq 0, \boldsymbol{\Theta}\mathbf{1} = \boldsymbol{\Theta}^\top \mathbf{1}, \mathbf{1}^\top \boldsymbol{\Theta} \mathbf{1} = 1. \end{aligned} \tag{7}$$

In the loss function of (7), the first term is the Kullback-Leibler distance between the empirical probability $\boldsymbol{\Omega}$ and the parameterized version $\boldsymbol{M}\boldsymbol{\Theta}\boldsymbol{M}^\top$ (ignoring a constant), and the second term is our identifiability-driven regularization. We need to tune the parameter λ to yield good estimation results. However, intuitively we should use a λ with a relatively small value. Suppose $\boldsymbol{\Omega}$ is sufficiently accurate, then the priority is to minimize the difference between $\boldsymbol{\Omega}$ and $\boldsymbol{M}\boldsymbol{\Theta}\boldsymbol{M}^\top$; when there exist equally good fits, then the second term comes into play and helps us pick out a solution that is *sufficiently scattered*.

Noticing that the constraints of (7) are all convex, but not the loss function, we propose to design an iterative algorithm to solve (7) using successive convex approximation. At iteration r when the updates are $\boldsymbol{\Theta}^r$ and $\boldsymbol{M}^r$, we define

$$\Pi^r_{n\ell kj} = M^r_{nk} \Theta^r_{kj} M^r_{\ell j} \Big/ \sum_{\kappa,\iota=1}^{K} M^r_{n\kappa} \Theta^r_{\kappa\iota} M^r_{\ell\iota}. \tag{8}$$

Obviously, $\Pi^r_{n\ell kj} \geq 0$ and $\sum_{k,j=1}^{K} \Pi^r_{n\ell kj} = 1$, which defines a probability distribution for fixed n and ℓ. Using Jensen's inequality (Jensen, 1906), we have that

$$\begin{aligned} & -\Omega_{n\ell} \log \sum_{k,j=1}^{K} M_{nk} \Theta_{kj} M_{\ell j} \\ & \leq \sum_{k,j=1}^{K} -\Omega_{n\ell} \Pi^r_{n\ell kj} \left(\log M_{nk} + \log \Theta_{kj} + \log M_{\ell j} \right. \\ & \qquad \left. - \log \Pi^r_{n\ell kj} \right) \end{aligned} \tag{9}$$

which defines a convex and locally tight upperbound for the first term in the loss function of (7). Regarding the second

term in the loss of (7), we propose to simply take the linear approximation

$$|\det\boldsymbol{\Theta}| \approx |\det\boldsymbol{\Theta}^r| + |\det\boldsymbol{\Theta}^r| \operatorname{Tr}\big((\boldsymbol{\Theta}^r)^{-1}(\boldsymbol{\Theta}-\boldsymbol{\Theta}^r)\big) \quad (10)$$

Combining (9) and (10), our successive convex approximation algorithm tries to solve the following convex problem at iteration r:

$$\begin{aligned}\underset{\boldsymbol{\Theta},\boldsymbol{M}}{\text{minimize}} \quad & \sum_{n,\ell=1}^{N}\sum_{k,j=1}^{K} -\Omega_{n\ell}\Pi^r_{n\ell kj}\left(\log M_{nk} + \log M_{\ell j}\right. \\ & \left. + \log\Theta_{kj}\right) + \lambda\sum_{k,j=1}^{K}\Xi^r_{kj}\Theta_{kj} \quad (11)\\ \text{subject to} \quad & \boldsymbol{M}\geq 0, \boldsymbol{1}^\top\boldsymbol{M}=\boldsymbol{1}^\top,\\ & \boldsymbol{\Theta}\geq 0, \boldsymbol{\Theta}\boldsymbol{1}=\boldsymbol{\Theta}^\top\boldsymbol{1}, \boldsymbol{1}^\top\boldsymbol{\Theta}\boldsymbol{1}=1,\end{aligned}$$

where we define $\boldsymbol{\Xi}^r = |\det\boldsymbol{\Theta}^r|(\boldsymbol{\Theta}^r)^{-\top}$. Problem (11) decouples with respect to $\boldsymbol{M}$ and $\boldsymbol{\Theta}$, so we can work out their updates individually.

The update of $\boldsymbol{M}$ admits a simple closed form solution, which can be derived via checking the KKT conditions. We denote the dual variable corresponding to $\boldsymbol{1}^\top\boldsymbol{M}=\boldsymbol{1}^\top$ as $\boldsymbol{\mu}\in\mathbb{R}^K$. Setting the gradient of the Lagrangian with respect to M_{nk} equal to zero, we have

$$M_{nk} = \sum_{\ell=1}^{N}\sum_{j=1}^{K}\left(\Omega_{n\ell}\Pi^r_{n\ell kj} + \Omega_{\ell n}\Pi^r_{\ell njk}\right)\Big/\mu_k$$

and $\boldsymbol{\mu}$ should be chosen so that the constraint $\boldsymbol{1}^\top\boldsymbol{M}=\boldsymbol{1}^\top$ is satisfied, which amounts to a simple re-scaling.

The update of $\boldsymbol{\Theta}$ is not as simple as a closed form expression, but it can still be obtained very efficiently. Noticing that the nonnegativity constraint is implicitly implied by the individual $\log$ functions in the loss function, we propose to solve it using Newton's method with equality constraints (Boyd & Vandenberghe, 2004, §10.2). Although Newton's method requires solving a linear system of equations with K^2 number of variables in each iteration, there is special structure we can exploit to reduce the per-iteration complexity down to $\mathcal{O}(K^3)$: The Hessian of the loss function of (11) is diagonal, and the linear equality constraints are highly structured; using block elimination (Boyd & Vandenberghe, 2004, §10.4.2), we ultimately only need to solve a positive definite linear system with K variables. Together with the quadratic convergence rate of Newton's method, the complexity of updating $\boldsymbol{\Theta}$ is $\mathcal{O}(K^3\log\log\frac{1}{\varepsilon})$, where ε is the desired accuracy for the $\boldsymbol{\Theta}$ update. Noticing that the complexity of a naive implementation of Newton's method would be $\mathcal{O}(K^6\log\log\frac{1}{\varepsilon})$, the difference is big for moderately large K. The in-line implementation of this tailored Newton's method ThetaUpdate and the detailed derivation can be found in the supplementary material.

Algorithm 1 Proposed Algorithm

Require: $\lambda > 0$
initialize $\boldsymbol{M}$ using (Huang et al., 2016a)
initialize $\boldsymbol{\Theta} \leftarrow \frac{1}{K(K+1)}(\boldsymbol{I}+\boldsymbol{1}\boldsymbol{1}^\top)$
repeat
  $\widetilde{\boldsymbol{\Omega}} \leftarrow \boldsymbol{\Omega}/\boldsymbol{M}\boldsymbol{\Theta}\boldsymbol{M}^\top$ ▷ element-wise division
  $\widetilde{\boldsymbol{M}} \leftarrow \boldsymbol{M} * \left(\widetilde{\boldsymbol{\Omega}}\boldsymbol{M}\boldsymbol{\Theta}^\top + \widetilde{\boldsymbol{\Omega}}^\top\boldsymbol{M}\boldsymbol{\Theta}\right)$
  $\widetilde{\boldsymbol{\Theta}} \leftarrow \boldsymbol{M}^\top\widetilde{\boldsymbol{\Omega}}\boldsymbol{M}$
  $\widetilde{\boldsymbol{M}} \leftarrow \widetilde{\boldsymbol{M}}\operatorname{Diag}(\boldsymbol{1}^\top\widetilde{\boldsymbol{M}})^{-1}$
  $\widetilde{\boldsymbol{\Theta}} \leftarrow$ ThetaUpdate ▷ cf. supplementary
  $(\boldsymbol{M},\boldsymbol{\Theta}) \leftarrow$ Amijo line search between $(\boldsymbol{M},\boldsymbol{\Theta})$ and $(\widetilde{\boldsymbol{M}},\widetilde{\boldsymbol{\Theta}})$
until convergence
return $\boldsymbol{M}$ and $\boldsymbol{\Theta}$

The entire proposed algorithm to solve Problem (7) is summarized in Algorithm 1. Notice that there is an additional line-search step to ensure decrease of the loss function. The constraint set of (7) is convex, so the line-search step will not incur infeasibility. Computationally, we find that any operation that involves $\Pi^r_{n\ell kj}$ can be carried out succinctly by defining the intermediate matrix $\widetilde{\boldsymbol{\Omega}} = \boldsymbol{\Omega}/\boldsymbol{M}\boldsymbol{\Theta}\boldsymbol{M}^\top$, where "/" denotes element-wise division between two matrices of the same size. The per-iteration complexity of Algorithm 1 is completely dominated by the operations that involve computing with $\widetilde{\boldsymbol{\Omega}}$, notably comparing with that of ThetaUpdate. In terms of initialization, which is important since we are optimizing a non-convex problem, we propose to use the method by Huang et al. (2016a) to obtain an initialization for $\boldsymbol{M}$; for $\boldsymbol{\Theta}$, it is best if we start with a feasible point (so that the Newton's iterates will remain feasible), and a simple choice is scaling the matrix $\boldsymbol{I}+\boldsymbol{1}\boldsymbol{1}^\top$ to sum up to one. Finally, we show that this algorithm converges to a stationary point of Problem (7), with proof relegated to the supplementary material based on (Razaviyayn et al., 2013).

Proposition 3. *Assume* ThetaUpdate *solves Problem* (11) *with respect to* $\boldsymbol{\Theta}$ *exactly, then Algorithm 1 converges to a stationary point of Problem* (7).

5. Validation on Synthetic Data

In this section we validate the identifiability performance on synthetic data. In this case, the underlying transition and emission probabilities are generated synthetically, and we compare them with the estimated ones to evaluate performance. The simulations are conducted in MATLAB using the HMM toolbox, which includes functions to generate observation sequences given transition and emission probabilities, as well as an implementation of the Baum-Welch algorithm (Baum et al., 1970), i.e., the EM algorithm, to estimate the transition and emission probabilities using the ob-

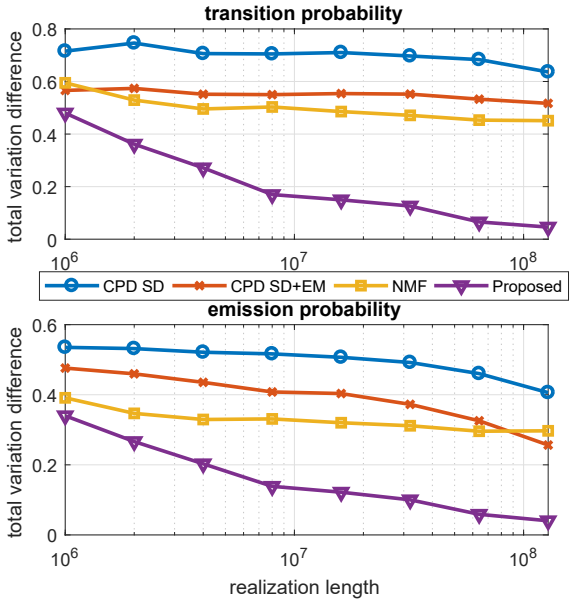


Figure 4: The total variation difference between the ground truth and estimated transition probability (top) and emission probability (bottom). The total variation difference of the emission probabilities is calculated as $\frac{1}{2K}\|\boldsymbol{M}_\natural - \boldsymbol{M}_\star\|_1$, since each column of the matrices indicates a (conditional) probability, and the total variation difference is equal to one half of the L_1-norm; and similarly for that of the transition probabilities after rescaling the rows of $\boldsymbol{\Omega}_\natural$ and $\boldsymbol{\Omega}_\star$ to sum up to one. The result is averaged over 10 random problem instances.

servations. Unfortunately, even for some moderate problem sizes we considered, the streamlined MATLAB implementation of the Baum-Welch algorithm was not able to execute within reasonable amount of time, so its performance is not included here. For the baselines, we compare with the plain NMF approach using multiplicative update (Vanluyten et al., 2008) and the tensor CPD approach (Sharan et al., 2017) using simultaneous diagonalization with Tensorlab (Vervliet et al., 2016). Since we work with empirical distributions instead of exact probabilities, the result of the simultaneous diagonalization is not going to be optimal. We therefore use it to initialize the EM algorithm for fitting a nonnegative tensor factorization with KL divergence loss (Shashanka et al., 2008) for refinement.

We focus on the cases when the number of hidden states K is smaller than the number observed states N. As we explained in the introduction, even for this seemingly easier case, it is not known that we can guarantee unique recovery of the HMM parameters *just from the pair-wise co-occurrence probability*. What is known is that the tensor CPD approach is able to guarantee identifiability given exact triple-occurrence probability. We will demonstrate in this section that it is much harder to obtain accurate triple-occurrence probability comparing with the co-occurrence probability. As a result, if the sufficiently scattered assumption holds for the emission probability, the estimated parameters obtained from our method are always more accurate than those obtained from tensor CPD.

Fixing $N = 100$ and $K = 20$, the transition probabilities are synthetically generated from a random exponential matrix of size $K \times K$ followed by row-normalization; for the emission probabilities, approximately 50% of the entries in the $N \times K$ random exponential matrices are set to zero before normalizing the columns, which is shown to satisfy the sufficiently scattered condition with very high probability (Huang et al., 2015). We let the number of HMM realizations go from 10^6 to 10^8, and compare the estimation error for the transition matrix and emission matrix by the aforementioned methods. We show the total variation distance between the ground truth probabilities $\Pr[X_{t+1}|X_t]$ and $\Pr[Y_t|X_t]$ and their estimations $\widehat{\Pr}[X_{t+1}|X_t]$ and $\widehat{\Pr}[Y_t|X_t]$ using various methods. The result is shown in Figure 4. As we can see, the proposed method indeed works best, obtaining almost perfect recovery when sample size is above 10^8. The CPD based method does not work as well since it cannot obtain accurate estimates of the third-order statistics that it needs. Initialized by CPD, EM improves from CPD but the performance is still far away from the proposed method. NMF is not working well since it does not have identifiability in this case.

6. Application: Hidden Topic Markov Model

Analyzing text data is one of the core application domains of machine learning. There are two prevailing approaches to model text data. The classical bag-of-words model assumes that each word is *independently* drawn from certain multinomial distributions. These distributions are different across documents, but can be efficiently summarized by a small number of *topics*, again mathematically modeled as distributions over words; this task is widely known as *topic modeling* (Hofmann, 2001; Blei et al., 2003). However, it is obvious that the bag-of-words representation is oversimplified. The n-gram model, on the other hand, assumes that words are conditionally dependent up to a window-length of n. This seems to be a much more realistic model, although the choice of n is totally unclear, and is often dictated by memory and computational limitations in practice—since the size of the joint distribution grows exponentially with n. What is more, it is somewhat difficult to extract "topics" from this model, despite some preliminary attempts (Wallach, 2006; Wang et al., 2007).

We propose to model a document as the realization of a HMM, in which the topics are hidden states emitting words, and the states are evolving according to a Markov chain, hence the name *hidden topic Markov model* (HTMM). For a

set of documents, this means we are working with a *collection* of HMMs. Similar to other topic modeling works, we assume that the topic matrix is shared among all documents, meaning all the given HMMs share the same emission probability. For the bag-of-words model, each document has a specific topic distribution $\boldsymbol{p}_d$, whereas for our model, each document has its own *topic transition probability* $\boldsymbol{\Theta}_d$; as per our previous discussion, the row-sum and column-sum of $\boldsymbol{\Theta}_d$ are the same, which are also the topic probability for the specific document. The difference is the Markovian assumption on the topics rather than the over-simplifying independence assumption.

We can see some immediate advantages for the HTMM. Since the Markovian assumption is only imposed on the topics, which are not exposed to us, the observations (words) are not independent from each other, which agrees with our intuition. On the other hand, we now understand that although word dependencies exist for a wide neighborhood, we only need to work with pair-wise co-occurrence probabilities, or 2-grams. This releases us from picking a window length n in the n-gram model, while maintaining dependencies between words well beyond a neighborhood of n words. It also includes the bag-of-words assumption as a special case: If the topics of the words are indeed independent, this just means that the transition probability has the special form $\boldsymbol{1}\boldsymbol{p}_d^\top$. The closest work to ours is by Gruber et al. (2007), which is also termed hidden topic Markov model. However, they make a simplifying assumption that the transition probability takes the form $(1-\epsilon)\boldsymbol{I} + \epsilon \boldsymbol{1}\boldsymbol{p}_d^\top$, meaning the topic of the word is either the same as the previous one, or independently drawn from $\boldsymbol{p}_d$. Both models are special cases of our general HTMM.

In order to learn the shared topic matrix $\boldsymbol{M}$, we can use the co-occurrence statistics for the entire corpus: Denote the co-occurrence statistics for the d-th document as $\boldsymbol{\Omega}_d$, then $\mathrm{E}\,\boldsymbol{\Omega}_d = \boldsymbol{M}\boldsymbol{\Theta}_d\boldsymbol{M}^\top$; consequently

$$\boldsymbol{\Omega} = \frac{1}{\sum_{d=1}^D L_d} \sum_{d=1}^D L_d \boldsymbol{\Omega}_d,$$

which is an unbiased estimator for

$$\boldsymbol{M}\boldsymbol{\Theta}\boldsymbol{M}^\top = \frac{1}{\sum_{d=1}^D L_d} \sum_{d=1}^D L_d \boldsymbol{M}\boldsymbol{\Theta}_d\boldsymbol{M}^\top,$$

where L_d is the length of the d-th document and $\boldsymbol{\Theta}$ is conceptually a weighted average of all the topic-transition matrices. Then we may apply Algorithm 1 to learn the topic matrix.

We illustrate the performance of our HTMM by comparing it to three popular bag-of-words topic modeling approaches: pLSA (Hofmann, 2001), LDA (Blei et al., 2003), and FastAnchor (Arora et al., 2013), which guarantees identifiability if every topic has a characteristic *anchor word*. Our HTMM model guarantees identifiability if the topic matrix is *sufficiently scattered*, which is a more relaxed condition than the anchor word one. On the Reuters21578 data set obtained at (Mimaroglu, 2007), we use the raw document to construct the word co-occurrence statistics, as well as bag-of-words representations for each document for the baseline algorithms. We use the version in which the stop-words have been removed, which makes the HTMM model more plausible since any syntactic dependencies have been removed, leaving only semantic dependencies. The vocabulary size of Reuters21578 is around $200,000$, making any method relying on triple-occurrences impossible to implement, and that is why tensor-based methods are not compared here.

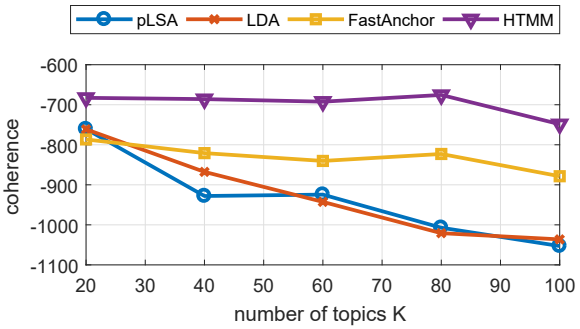


Figure 5: Coherence of the topics.

Because of page limitations, we only show the quality of the topics learned by various methods in terms of coherence. Simply put, a higher coherence means more meaningful topics, and the concrete definition can be found in (Arora et al., 2013) and in the supplementary material. In Figure 5, we can see that for different number of topics we tried on the entire dataset, HTMM consistently produces topics with the highest coherence. Additional evaluations can be found in the supplementary material.

7. Conclusion

We presented an algorithm for learning hidden Markov models in an unsupervised setting, i.e., using only a sequence of observations. Our approach is guaranteed to uniquely recover the ground-truth HMM structure using only pairwise co-occurrence probabilities of the observations, under the assumption that the emission probability is *sufficiently scattered*. Unlike EM, the complexity of the proposed algorithm does not grow with the length of the observation sequence. Compared to tensor-based methods for HMM learning, our approach only requires reliable estimates of pairwise co-occurrence probabilities, which are easier to obtain. We applied our method to topic modeling, assuming each document is a realization of a HMM rather than a simpler bag-of-words model, and obtained improved topic coherence results. We refer the reader to the supplementary material for detailed proofs of the propositions and additional experimental results.

Acknowledgments

This work is supported in part by the National Science Foundation (NSF) under grants CIF-1525194, ECCS-1608961, and IIS-1704074.

References

Allman, Elizabeth S., Matias, Catherine, and Rhodes, John A. Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics*, 37(6 A):3099–3132, 2009.

Anandkumar, Animashree, Hsu, Daniel, and Kakade, Sham M. A method of moments for mixture models and hidden Markov models. In *Conference on Learning Theory (COLT)*, pp. 33.1–33.34, 2012.

Anandkumar, Animashree, Ge, Rong, Hsu, Daniel, Kakade, Sham M., and Telgarsky, Matus. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014.

Arora, Sanjeev, Ge, Rong, Halpern, Yonatan, Mimno, David, Moitra, Ankur, Sontag, David, Wu, Yichen, and Zhu, Michael. A practical algorithm for topic modeling with provable guarantees. In *International Conference on Machine Learning*, pp. 280–288, 2013.

Barlier, Merwan, Laroche, Romain, and Pietquin, Olivier. Learning dialogue dynamics with the method of moments. In *IEEE Spoken Language Technology Workshop (SLT)*, pp. 98–105, 2015.

Baum, Leonard E, Petrie, Ted, Soules, George, and Weiss, Norman. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1): 164–171, 1970.

Bhaskara, Aditya, Charikar, Moses, and Vijayaraghavan, Aravindan. Uniqueness of tensor decompositions with applications to polynomial identifiability. In *Conference on Learning Theory (COLT)*, pp. 742–778, 2014.

Blei, David M, Ng, Andrew Y, and Jordan, Michael I. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

Boyd, Stephen and Vandenberghe, Lieven. *Convex Optimization*. Cambridge University Press, 2004.

Chiantini, Luca and Ottaviani, Giorgio. On generic identifiability of 3-tensors of small rank. *SIAM Journal on Matrix Analysis and Applications*, 33(3):1018–1037, 2012.

Cybenko, George and Crespi, Valentino. Learning hidden Markov models using nonnegative matrix factorization. *IEEE Transactions on Information Theory*, 57(6):3963–3970, 2011.

De Lathauwer, Lieven, De Moor, Bart, and Vandewalle, Joos. Computation of the canonical decomposition by means of a simultaneous generalized Schur decomposition. *SIAM Journal on Matrix Analysis and Applications*, 26(2):295–327, Jan 2004.

Donoho, David and Stodden, Victoria. When does nonnegative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems*, pp. 1141–1148, 2004.

Fu, Xiao, Ma, Wing-Kin, Huang, Kejun, and Sidiropoulos, Nicholas D. Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain. *IEEE Transactions on Signal Processing*, 63(9): 2306–2320, May 2015.

Fu, Xiao, Huang, Kejun, and Sidiropoulos, Nicholas D. On identifiability of nonnegative matrix factorization. *IEEE Signal Processing Letters*, 25(3):328–332, 2018.

Ghahramani, Zoubin. An introduction to hidden Markov models and Baysian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 15(1): 9–42, 2001.

Glaude, Hadrien, Enderli, Cyrille, and Pietquin, Olivier. Spectral learning with non negative probabilities for finite state automaton. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 71–77, 2015.

Gruber, Amit, Weiss, Yair, and Rosen-Zvi, Michal. Hidden topic Markov models. In *Artificial Intelligence and Statistics*, pp. 163–170, 2007.

Harshman, Richard A. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multimodal factor analysis. Technical report, University of California at Los Angeles, 1970.

Hofmann, Thomas. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1-2): 177–196, 2001.

Hsu, Daniel J, Kakade, Sham M, and Zhang, Tong. A spectral algorithm for learning hidden Markov models. In *Conference on Learning Theory (COLT)*, 2009.

Huang, Kejun, Sidiropoulos, Nicholas D., and Swami, Ananthram. Non-negative matrix factorization revisited: Uniqueness and algorithm for symmetric decomposition. *IEEE Transactions on Signal Processing*, 62(1):211–224, Jan. 2014.

Huang, Kejun, Sidiropoulos, Nicholas D., Papalexakis, Evangelos, Christos, Faloutsos, Talukdar, Partha P., and Mitchell, Tom. Principled neuro-functional connectivity discovery. In *SIAM International Conference on Data Mining (SDM)*, 2015.

Huang, Kejun, Fu, Xiao, and Sidiropoulos, Nicholas D. Anchor-free correlated topic modeling: Identifiability and algorithm. In *Advances in Neural Information Processing Systems (NIPS)*, 2016a.

Huang, Qingqing, Ge, Rong, Kakade, Sham, and Dahleh, Munther. Minimal realization problems for hidden Markov models. *IEEE Transactions on Signal Processing*, 64(7):1896–1904, 2016b.

Jensen, Johan. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30(1): 175–193, 1906.

Kontorovich, Leonid. Measure concentration of hidden Markov processes. *arXiv preprint math/0608064*, 2006.

Lakshminarayanan, Balaji and Raich, Raviv. Non-negative matrix factorization for parameter estimation in hidden Markov models. In *IEEE International Workshop on Machine Learning for Signal Processing*, pp. 89–94, 2010.

Leurgans, SE, Ross, RT, and Abel, RB. A decomposition for three-way arrays. *SIAM Journal on Matrix Analysis and Applications*, 14(4):1064–1083, 1993.

Mimaroglu, Selim. Some text datasets, 2007. https://www.cs.umb.edu/~smimarog/textmining/datasets/.

Rabiner, Lawrence and Juang, B. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1):4–16, 1986.

Razaviyayn, Meisam, Hong, Mingyi, and Luo, Zhi-Quan. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153, 2013.

Sanchez, Eugenio and Kowalski, Bruce R. Tensorial resolution: A direct trilinear decomposition. *Journal of Chemometrics*, 4(1):29–45, 1990.

Sharan, Vatsal, Kakade, Sham, Liang, Percy, and Valiant, Gregory. Learning overcomplete HMMs. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Shashanka, Madhusudana, Raj, Bhiksha, and Smaragdis, Paris. Probabilistic latent variable models as nonnegative factorizations. *Computational Intelligence and Neuroscience*, 2008, 2008.

Sidiropoulos, Nicholas D, De Lathauwer, Lieven, Fu, Xiao, Huang, Kejun, Papalexakis, Evangelos E, and Faloutsos, Christos. Tensor decomposition for signal processing and machine learning. *IEEE Transactions on Signal Processing*, 65(13):3551–3582.

Vanluyten, Bart, Willems, Jan C., and De Moor, Bart. Structured nonnegative matrix factorization with applications to hidden Markov realization and clustering. *Linear Algebra and Its Applications*, 429(7):1409–1424, 2008.

Vervliet, N., Debals, O., Sorber, L., Van Barel, M., and De Lathauwer, L. Tensorlab 3.0, Mar. 2016. URL https://www.tensorlab.net. Available online.

Wallach, Hanna M. Topic modeling: Beyond bag-of-words. In *International Conference on Machine Learning*, pp. 977–984, 2006.

Wang, Xuerui, McCallum, Andrew, and Wei, Xing. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *IEEE International Conference on Data Mining*, pp. 697–702, 2007.