Hidden Markov Model Identifiability via Tensors

Paul Tune, Hung X. Nguyen and Matthew Roughan
School of Mathematical Sciences, University of Adelaide,
Adelaide, SA, Australia.
Email: {paul.tune, hung.nguyen, matthew.roughan}@adelaide.edu.au

Abstract—**The prevalence of hidden Markov models (HMMs) in various applications of statistical signal processing and communications is a testament to the power and flexibility of the model. In this paper, we link the identifiability problem with tensor decomposition, in particular, the Canonical Polyadic decomposition. Using recent results in deriving uniqueness conditions for tensor decomposition, we are able to provide a necessary and sufficient condition for the identification of the parameters of discrete time finite alphabet HMMs. This result resolves a long standing open problem regarding the derivation of a necessary and sufficient condition for uniquely identifying an HMM. We then further extend recent preliminary work on the identification of HMMs with multiple observers by deriving necessary and sufficient conditions for identifiability in this setting.**

I. Introduction

The hidden Markov model (HMM) was first introduced in the late 1950s by Blackwell and Koopmans [4] and generalised later by Baum and Petrie [2]. HMMs have been applied to a variety of domains, such as signal processing, machine learning, communications and many more, with particular emphasis on the inference of the parameters of the HMM, in particular, the hidden states of the system. Typically, an unbiased, or asymptotically unbiased, estimator such as a maximum likelihood estimator, is used to infer these states, using algorithms such as the famed Baum-Welch algorithm [3]. However, identifiability conditions, required to ensure the existence of an unbiased estimator, are generally not well-known, with only a select number of works proposing these conditions [1], [6], [7]. These conditions are probabilistic and difficult to verify in practice.

In this paper, we derive an identifiability condition for a stationary discrete time HMM, where the observations are the realisations of a probabilistic function. We show a strong connection between the identifiability of HMMs and the uniqueness of the Canonical Polyadic (CP) tensor decomposition (see [9]). Specifically through a tensor model called the restricted CP model, we derive a necessary and sufficient condition by using a result by Kruskal [10], called the permutation lemma. Our main result resolves an open problem regarding the derivation of a necessary and sufficient condition for uniquely identifying an HMM. A highlight of our results is that the condition is deterministic, compared to generic (probabilisitic) identifiability results. They are also easier to verify compared to previous conditions.

These results are particularly helpful in studying the recently proposed multi-observer HMMs [11], [12]. We consider two settings: the *homogeneous* setting where all observers possess the same observation matrix, and the *heterogeneous* setting, where at least two observers have distinct observation matrices. Surprisingly, the condition for identifiability in the homogeneous setting is equivalent to having just a single observer of the HMM. Thus, if the HMM cannot be identified with a single observer, no additional number of independent homogeneous observers can hope to identify the hidden states. The heterogeneous setting is shown to provide a significant advantage over the homogeneous setting, due to sufficient variability of the observations, contributed by different viewpoints of the independent observers.

The rest of this section introduces the notation used throughout the paper. Section II formulates our problem and defines the HMM. Section III provides an overview of the CP decomposition and some important results in tensor decomposition that we will invoke when proving our results. Section IV derives the identifiability condition of HMMs with only one observer. Section V further extends our framework to the multi-observer setting. Finally, we conclude and outline some future work in Section VI. We defer proofs and more details to our technical report [16].

Some notation and definitions are in order. All vectors and matrices are represented with lower and upper case boldface fonts respectively. Sets are represented with calligraphic font. Random variables are represented with italic fonts while their realisation is represented by lower case italic fonts. Tensors are represented by upper case, calligraphic boldface fonts. Let $\mathbf{I}_n$ be the identity matrix of size $n \times n$, while $\mathbf{1}_n$ denotes a column vector of n ones. $\mathrm{supp}(\mathbf{x})$ denotes the support of vector $\mathbf{x}$.

The Kronecker, or tensor, product between two matrices $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{p\times q}$ is denoted by $\mathbf{A} \otimes \mathbf{B} \in \mathbb{R}^{mp\times nq}$. We also define a row-wise tensor product, which we borrow from [1], where with matrices $\mathbf{A} \in \mathbb{R}^{m\times n_1}$ and $\mathbf{B} \in \mathbb{R}^{m\times n_2}$ with rows $\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_m$ and $\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_m$ respectively, the row-wise tensor product is equivalent to

$$\mathbf{A} \otimes^{\mathrm{row}} \mathbf{B} = \begin{bmatrix} \mathbf{a}_1 \otimes \mathbf{b}_1 \\ \mathbf{a}_2 \otimes \mathbf{b}_2 \\ \vdots \\ \mathbf{a}_m \otimes \mathbf{b}_m \end{bmatrix}.$$

This definition is related to the Khatri-Rao product, which is the column-wise tensor product. Note that for row vectors $\mathbf{a}$ and $\mathbf{b}$, $\mathbf{a} \otimes^{\mathrm{row}} \mathbf{b} = \mathbf{a} \otimes \mathbf{b}$. All other notation will be defined on an as needed basis.

II. Hidden Markov Models

Throughout the paper, we only consider the discrete time finite alphabet HMM. Time is organised into regularly spaced discrete intervals. Let $\{X_t\}_{t\geq 1}$ be the non-observable states of an irreducible, aperiodic Markov chain and $\{Y_t\}_{t\geq 1}$ be the observable states, both at time t, called an *observation process*. Thus, $\{X_t\}_{t\geq 1}$ constitutes the hidden Markov chain, as it cannot be directly measured. Without loss of generality, let the alphabets of X_t and Y_t be the sets $\mathcal{X} = \{1, 2, \cdots, q\}$ and $\mathcal{Y} = \{1, 2, \cdots, \kappa\}$ respectively. We further assume $q \geq 2$ and $\kappa \geq 2$.

We assume $\{X_t\}_{t\geq 1}$ is stationary for simplicity of exposition, although our results apply to non-stationary Markov chains as well, with appropriate modifications. The observations $\{Y_t\}_{t\geq 1}$ are assumed to be i.i.d. with Y_t only dependent on X_t. The HMM is described by the joint process $(X_t, Y_t) \in \mathcal{X} \times \mathcal{Y}$ for all $t = 1, 2, \cdots, N$, in terms of the state space model

$$X_{t+1} = f(X_t),\ Y_t = g(X_t),$$

with the initial state described by an initial random variable X_1, from an initial distribution $\Pr(X_1 = i) = \pi_i$, denoted by the q–length row vector $\boldsymbol{\pi}$. The function $f(\cdot)$ is probabilistic, and obeys the $q \times q$ *transition matrix* $\mathbf{A}$, where its (i, j)-th element is $a_{i,j} := \Pr(X_{t+1} = j \mid X_t = i)$. The function $g(\cdot)$ may be deterministic, but here, we consider it a probabilistic function, with the transition of observation states described by the $q \times \kappa$ *observation matrix* $\mathbf{B}$, where the (i, j)-th element is $b_{i,j} := \Pr(Y_t = j \mid X_t = i)$. The function $g(\cdot)$ is assumed to be surjective. Since the Markov chain is assumed to be stationary, the observation process $\{Y_t\}_{t\geq 1}$ is stationary as well. Also, the observation process $\{Y_t\}_{t\geq 1}$ may not be a Markov chain in general, even though the input process is. We are now in a position to formalise HMMs.

Definition 1: A discrete time finite alphabet HMM is parameterised by the set $\boldsymbol{\lambda} = \{\boldsymbol{\pi}; q, \kappa, \mathbf{A}, \mathbf{B}\}$:

- $\boldsymbol{\pi}$: initial state probabilities, which may or may not be sampled from a stationary distribution,
- q: number of hidden states of X_t, i.e. $|\mathcal{X}| = q$,
- κ: number of observation states of Y_t, i.e. $|\mathcal{Y}| = \kappa$,
- $\mathbf{A}$: $q \times q$ transition matrix of hidden states X_t, and
- $\mathbf{B}$: $q \times \kappa$ observation matrix of observation process Y_t.

One way of measuring the complexity of the HMM is by the number of states required to describe the Markov chain. The *order* of an HMM is minimum of $|\mathcal{X}|$ amongst all representations [6]. An HMM is *minimal* if it has a representation such that $|\mathcal{X}|$ is equal to its order.

An *observation letter* y_t is defined as a single realisation of the observation process at time t. A *sequence* from time t_1 to t_2 is defined as a series of consecutive observation letters from time t_1 to t_2. A sequence has *length* N if it consists of N observation letters.

The joint probability of a particular observed sequence $y_1, y_2, \cdots, y_N$ may be described by

$$\begin{aligned} &P_{\boldsymbol{\lambda}}(Y_1 = y_1, Y_2 = y_2, \cdots, Y_N = y_N) \\ &= \sum_{\mathbf{x}\in\mathcal{X}^N} \pi_{x_1} a_{x_1,x_2} b_{x_1,y_1} a_{x_2,x_3} b_{x_2,y_2} \cdots a_{x_{N-1},x_N} b_{x_N,y_N} \\ &= \boldsymbol{\pi}\mathbf{W}\mathbf{E}(y_1)\mathbf{W}\mathbf{E}(y_2)\cdots\mathbf{W}\mathbf{E}(y_N)\mathbf{1}_q, \qquad (1) \end{aligned}$$

where $\mathbf{E}(k)$ is a $\kappa q \times q$ matrix with the $q \times q$ identity matrix in the k-th row partition, and

$$\mathbf{W} = \mathbf{B} \otimes^{\text{row}} \mathbf{A} = \left[\mathbf{D}_1(\mathbf{B})\mathbf{A}\ \mathbf{D}_2(\mathbf{B})\mathbf{A}\ \cdots\ \mathbf{D}_\kappa(\mathbf{B})\mathbf{A}\right],$$

where $\mathbf{D}_k(\mathbf{B})$ denotes the diagonal matrix with the k-th column of $\mathbf{B}$ lying on its diagonal.

A. Equivalence and identifiability of HMMs

The observation process $\{Y_t\}_{t\geq 1}$ is assumed to admit a representation of a Markov chain with q states. It is possible to construct an HMM with more states that generates the same observation process. Let the process be alternatively parameterised by the set $\tilde{\boldsymbol{\lambda}} = \{\tilde{\boldsymbol{\pi}}; \tilde{q}, \tilde{\kappa}, \tilde{\mathbf{A}}, \tilde{\mathbf{B}}\}$. Equivalence of HMMs is defined as follows:

Definition 2: Two HMMs with parameterisations $\boldsymbol{\lambda}$ and $\tilde{\boldsymbol{\lambda}}$ respectively are *equivalent* if and only if for all sequences $y_1, y_2, \cdots, y_N$,

$$\begin{aligned} &P_{\boldsymbol{\lambda}}(Y_1 = y_1, Y_2 = y_2, \cdots, Y_N = y_N) \\ &\qquad = P_{\tilde{\boldsymbol{\lambda}}}(\tilde{Y}_1 = y_1, \tilde{Y}_2 = y_2, \cdots, \tilde{Y}_N = y_N), \end{aligned}$$

for any integer $N \geq 1$.

There are two types of identifiability: *deterministic* and *generic identifiability*. Deterministic identifiability implies that HMMs satisfying the condition can always be identified. Generic identifiability means that the HMM is identified with probability 1, i.e. identifiability holds everywhere except for some model parameters that lie in a set of Lebesgue measure zero. Allman et al. [1] defines it as all nonidentifiable parameters of the model lying in a proper subvariety.

Our main result is a condition when an HMM can or cannot be deterministically identified.

III. A Summary on Tensors

As shown in (1), the joint probability of a sequence can be expressed as matrix multiplication of row tensor products. We show that identifiability of HMMs simply boils down to decomposing the product into factors via tensor decomposition.

The tensor is essentially a multidimensional array of numbers, with a general overview found in [9]. The tensor order is the number of indices required to unambiguously label a component of the tensor, called a *way*.

One particular important decomposition of a tensor is the Canonical Polyadic (CP) decomposition. Let the components of a tensor be $\mathbf{A} \in \mathbb{R}^{q\times m}$, $\mathbf{B} \in \mathbb{R}^{q\times n}$ and $\mathbf{C} \in \mathbb{R}^{q\times p}$, written succinctly as $[\mathbf{A}, \mathbf{B}, \mathbf{C}]$. Then a tensor $\boldsymbol{\mathcal{X}}$ constructed from these components is expressed as

$$[\boldsymbol{\mathcal{X}}; \mathbf{A}, \mathbf{B}, \mathbf{C}] = \sum_{i=1}^{q} \mathbf{a}_i \otimes \mathbf{b}_i \otimes \mathbf{c}_i, \qquad (2)$$

where $\mathbf{a}_i, \mathbf{b}_i, \mathbf{c}_i,\ i = 1, 2 \cdots, q$ are the rows of $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ respectively, essentially a decomposition to q single rank tensors. A tensor is *irreducible* with q components if and only if it cannot be decomposed to fewer than q components. A tensor $\boldsymbol{\mathcal{X}}$ is *permutation and scaling indeterminate* if its components are unique up to a scaling and permutation of rows. Hence, for any alternative decomposition of $\boldsymbol{\mathcal{X}}$ with components $[\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \tilde{\mathbf{C}}]$, there exists a permutation matrix $\mathbf{\Pi}$ and nonsingular scaling matrices $\mathbf{\Lambda}_{\mathbf{A}}$, $\mathbf{\Lambda}_{\mathbf{B}}$ and $\mathbf{\Lambda}_{\mathbf{C}}$ where $\mathbf{\Lambda}_{\mathbf{A}}\mathbf{\Lambda}_{\mathbf{B}}\mathbf{\Lambda}_{\mathbf{C}} = \mathbf{I}_q$, such that $\tilde{\mathbf{A}} = \mathbf{\Pi}\mathbf{\Lambda}_{\mathbf{A}}\mathbf{A}$, $\tilde{\mathbf{B}} = \mathbf{\Pi}\mathbf{\Lambda}_{\mathbf{B}}\mathbf{B}$ and $\tilde{\mathbf{C}} = \mathbf{\Pi}\mathbf{\Lambda}_{\mathbf{C}}\mathbf{C}$. Finally, an equivalent representation of a tensor is given by its mode *matricisation*. For example, the first mode matricisation of $\boldsymbol{\mathcal{X}}$ is $\mathbf{A}^{\mathrm{T}}(\mathbf{B} \otimes^{\mathrm{row}} \mathbf{C})$, while its second mode matricisation is $\mathbf{B}^{\mathrm{T}}(\mathbf{C} \otimes^{\mathrm{row}} \mathbf{A})$.

Surprisingly, under certain mild conditions, a tensor of order 3 and above possess a unique CP decomposition, up to a scaling and permutation of rows of the components, unlike matrices. A general sufficient condition was proposed by Kruskal [10], with its generalisation in [14].

Central to Kruskal's and our results is the concept of the *Kruskal rank* defined below:

Definition 3: The Kruskal rank of a matrix $\mathbf{X}$, $\mathrm{krank}(\mathbf{X})$ is defined as the largest integer K such that any subset of K rows is linearly independent.

Unlike the rank of a matrix, the Kruskal rank changes when one defines it for columns instead. Here, we stick to the above definition for rows, since this is directly relevant to our discussions.

The cornerstone of Kruskal's result, as pointed out by [8], [15] is Kruskal's permutation lemma, here modified for rows. The lemma is key to our proposed identifiability condition.

Lemma 4 (Permutation lemma): Given two matrices $\mathbf{H}$ and $\bar{\mathbf{H}}$, both with size $q \times r$, suppose that $\mathbf{H}$ has no identically zero rows, and assume the following implication holds for all column vectors $\mathbf{x}$:

$$|\mathrm{supp}(\bar{\mathbf{H}}\mathbf{x})| \leq q - \mathrm{rank}(\bar{\mathbf{H}}) + 1$$
$$\text{implies that } |\mathrm{supp}(\mathbf{H}\mathbf{x})| \leq |\mathrm{supp}(\bar{\mathbf{H}}\mathbf{x})|.$$

Then, $\bar{\mathbf{H}} = \mathbf{\Pi}\mathbf{\Lambda}\mathbf{H}$, where $\mathbf{\Pi}$ is a permutation matrix and $\mathbf{\Lambda}$ is a nonsingular diagonal scaling matrix.

IV. Single observer HMM Setting

Intuitively, the identifiability of an HMM rests on the number of states q and the number of observation states κ, both having a direct relationship with $\mathbf{A}$ and $\mathbf{B}$ respectively. Our reformulation using the properties of tensors allows us to explore this relationship. Since the underlying Markov chain is assumed to be irreducible and aperiodic and assuming all alphabets in $\mathcal{Y}$ are not redundant, $\mathrm{krank}(\mathbf{A}) \geq 1$ and $\mathrm{krank}(\mathbf{B}) \geq 1$ respectively.

A. Main result

Our main result is the following:

Theorem 5: For an HMM parameterised by $\boldsymbol{\lambda}$ to be unique up to a scaling and permutation of states, it is necessary and sufficient that $\mathrm{krank}(\mathbf{B} \otimes^{\mathrm{row}} \mathbf{A}) = q$.

Previous work [6], [7] studied the class of *regular* HMMs. An HMM is regular if there exists a set of $2q$ sequences whose joint probabilities can be described by a product of two linear subspaces of dimension q. A regular HMM is permutation and scaling indeterminate. Finesso [6] provided a simple sufficient condition for equivalence between two HMMs. Additionally, the author proved a necessary condition for a regular HMM to be equivalent to another HMM. The proof is probabilistic as he showed the set of parameters $\boldsymbol{\lambda}$ of HMMs almost surely leads to regularity in the Lebesgue measure.

Our result differs from Finesso's in the sense that we show an interaction between the hidden and the observation states, dispensing with assumption of regularity of the HMM and replacing it with a deterministic condition. Furthermore, regular HMMs are also minimal, implying $\mathrm{krank}(\mathbf{A}) = q$ (see [16]). Our result requires no restriction to regular HMMs, or as coined by Finesso [6], a *Petrie point* after Petrie's work [13] on regular HMMs, since the deterministic condition covers all possible cases. In this sense, the result is the strongest to date on the identifiability, whether deterministic or generic, of HMMs.

Theorem 5 is a consequence of the properties of a specific restricted CP tensor, where one mode of the tensor is full rank [8], which we call the *per letter tensor*. Let us consider $\boldsymbol{\mathcal{L}}$, a three way tensor of dimensions $q \times q \times \kappa$, with component matrices $[\mathbf{A}, \mathbf{I}_q, \mathbf{B}]$. For each element of $\boldsymbol{\mathcal{L}}$,

$$\begin{aligned}\boldsymbol{\mathcal{L}}_{i,j,k} &:= P_{\boldsymbol{\lambda}}(X_{t+1} = i \,|\, X_t = j) \cdot P_{\boldsymbol{\lambda}}(Y_t = k \,|\, X_t = j) \\ &= P_{\boldsymbol{\lambda}}(Y_t = k, X_{t+1} = i \,|\, X_t = j).\end{aligned}$$

Then, each slice of the third mode $\boldsymbol{\mathcal{L}}_k := \boldsymbol{\mathcal{L}}_{\cdot,\cdot,k} = \mathbf{I}_q\mathbf{D}_k(\mathbf{B})\mathbf{A}$, $k = 1, 2, \cdots, \kappa$ is the per observation letter and state probability of the set $\mathcal{Y}$ arranged in ascending order[1]. A key observation is the equivalence of $\boldsymbol{\mathcal{L}}$ and $\mathbf{I}_q(\mathbf{B} \otimes^{\mathrm{row}} \mathbf{A})$, its second mode matricisation.

We now prove a necessary and sufficient condition on the uniqueness of the decomposition of $\boldsymbol{\mathcal{L}}$.

Lemma 6: The per letter tensor $\boldsymbol{\mathcal{L}}$ is unique up to a permutation and scaling of rows if and only if $\mathrm{krank}(\mathbf{B} \otimes^{\mathrm{row}} \mathbf{A}) = q$.

Proof: Crucial to our argument is the central claim that it is necessary and sufficient that none of the non-trivial linear combinations of rows of $\mathbf{B} \otimes^{\mathrm{row}} \mathbf{A}$ is expressible by a tensor product of two row vectors, that is, $\mathrm{krank}(\mathbf{B} \otimes^{\mathrm{row}} \mathbf{A}) = q$. Necessity is proven by contradiction. We borrow a counterexample from [8]. If the first two rows can be expressed as a vector $\mathbf{b}_1 \otimes \mathbf{a}_1 + \mathbf{b}_2 \otimes \mathbf{a}_2 = \tilde{\mathbf{b}}_1 \otimes \tilde{\mathbf{a}}_1$, an alternative decomposition of $\boldsymbol{\mathcal{L}}$ is as follows:

$$\begin{aligned}\mathbf{I}_q(\mathbf{B} \otimes^{\mathrm{row}} \mathbf{A}) &= \begin{bmatrix} 1 & 0 & \mathbf{0} \\ -1 & 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_{q-2} \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} \tilde{\mathbf{b}}_1 \otimes \tilde{\mathbf{a}}_1 \\ \mathbf{b}_2 \otimes \mathbf{a}_2 \\ \vdots \\ \mathbf{b}_q \otimes \mathbf{a}_q \end{bmatrix} \\ &= \mathbf{C}^{\mathrm{T}}(\tilde{\mathbf{B}} \otimes^{\mathrm{row}} \tilde{\mathbf{A}}). \qquad (3)\end{aligned}$$

[1] This is one way of labelling the letters, as labelling is non-unique.

As $\mathbf{C}^{\mathrm{T}}$ is not a permutation or scaling, there is no unique decomposition for $\boldsymbol{\mathcal{L}}$.

For sufficiency, we only need to verify $|\mathrm{supp}(\mathbf{x})| = |\mathrm{supp}(\mathbf{I}_q\mathbf{x})| \leq |\mathrm{supp}(\tilde{\mathbf{C}}\mathbf{x})|$ for all $|\mathrm{supp}(\tilde{\mathbf{C}}\mathbf{x})| = 1$ (since $\mathbf{x} = \mathbf{0}$ is the only zero support vector) for some $\tilde{\mathbf{C}}$, a component of an alternative decomposition of $\boldsymbol{\mathcal{L}}$, to satisfy Lemma 4. With an alternative decomposition $\mathbf{I}_q(\mathbf{B}\otimes^{\mathrm{row}}\mathbf{A}) = \tilde{\mathbf{C}}^{\mathrm{T}}(\tilde{\mathbf{B}}\otimes^{\mathrm{row}}\tilde{\mathbf{A}})$, then $\forall \mathbf{x}$,

$$\mathbf{x}^{\mathrm{T}}(\mathbf{B}\otimes^{\mathrm{row}}\mathbf{A}) = \mathbf{x}^{\mathrm{T}}\tilde{\mathbf{C}}^{\mathrm{T}}(\tilde{\mathbf{B}}\otimes^{\mathrm{row}}\tilde{\mathbf{A}}).$$

Consider $\mathbf{x}$ with $|\mathrm{supp}(\tilde{\mathbf{C}}\mathbf{x})| = 1$. Then, $\mathbf{x}^{\mathrm{T}}\tilde{\mathbf{C}}^{\mathrm{T}}(\tilde{\mathbf{B}}\otimes^{\mathrm{row}}\tilde{\mathbf{A}})$ is just a scaled tensor product of one row of $\tilde{\mathbf{A}}$ and the corresponding row of $\tilde{\mathbf{B}}$, by the above equation. If $|\mathrm{supp}(\mathbf{x})| > 1$, then more than one row of $\mathbf{B}\otimes^{\mathrm{row}}\mathbf{A}$ is needed to represent a row of $\tilde{\mathbf{B}}\otimes^{\mathrm{row}}\tilde{\mathbf{A}}$. This means a row of $\tilde{\mathbf{B}}\otimes^{\mathrm{row}}\tilde{\mathbf{A}}$ is not just a scaling and permutation, so $|\mathrm{supp}(\mathbf{x})| \leq 1$ must hold. Kruskal's permutation lemma (Lemma 4) then implies $\mathbf{I}_q$ and $\tilde{\mathbf{C}}$ are equivalent up to a permutation and scaling of rows. Putting these arguments together implies the components of $\boldsymbol{\mathcal{L}}$, i.e. $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{I}_q$ are all unique up to a permutation and scaling of rows, and the result follows. ■

The above immediately implies a sufficient condition for HMM identifiability. For necessity, suppose $\mathrm{krank}(\mathbf{B}\otimes^{\mathrm{row}}\mathbf{A}) < q$, but the HMM is identifiable. As the condition does not hold, one can construct $\boldsymbol{\pi} = \tilde{\boldsymbol{\pi}}\tilde{\mathbf{C}}^{\mathrm{T}}$, $\mathbf{1}_q = (\tilde{\mathbf{C}}^{\mathrm{T}})^{-1}\tilde{\mathbf{1}}_q, (\mathbf{B}\otimes^{\mathrm{row}}\mathbf{A})\mathbf{E}(k) = (\tilde{\mathbf{B}}\otimes^{\mathrm{row}}\tilde{\mathbf{A}})(\mathbf{E}(k)\tilde{\mathbf{C}}^{\mathrm{T}}), \forall k$, with $\tilde{\mathbf{C}}$ as above in the proof of Lemma 6, resulting in a contradiction. Thus, $\mathrm{krank}(\mathbf{B}\otimes^{\mathrm{row}}\mathbf{A}) = q$ is a necessary and sufficient condition for identifiability.

Intuitively, the condition ensures all information to uniquely identify $\boldsymbol{\pi}$ from $\mathbf{B}\otimes^{\mathrm{row}}\mathbf{A}$ is preserved since $\mathbf{B}\otimes^{\mathrm{row}}\mathbf{A}$ has full row rank. Furthermore, the condition implies that $\mathbf{B}\otimes^{\mathrm{row}}\mathbf{A}$ each row is uniquely encoded, so information about $\mathbf{A}$ and $\mathbf{B}$ is preserved as well.

Evaluating the Kruskal rank of $\mathbf{B}\otimes^{\mathrm{row}}\mathbf{A}$ is computationally difficult as it requires checking over all possible combinations of rows of a matrix. The computational complexity worsens as q and κ become large. The conditions above may be weakened using the concept of coherence [5], found in compressed sensing and dictionary learning literature to derive polynomial time algorithms for verifying HMM identifiability. Further details are found in [16].

V. Multi-observer HMM Setting

The above results prove useful in the study of multi-observer HMMs, which have applications in machine learning [11], and detecting attacks on Internet Service providers [12]. We shall study identifiability in this setting, but first, we need to lay the foundations for the multi-observer case.

In the multi-observer setting, there are $m \geq 2$ observers of an underlying Markov chain. The observers are assumed independent to each other, since dependence would weaken the information content of their observations. The underlying irreducible, aperiodic Markov chain $\{X_t\}_{t\geq 1}$ is being observed by all m observers, with each $X_t \in \mathcal{X}$, starting from initial state probabilities $\boldsymbol{\pi}$, drawn from a stationary distribution. Each observer j may have a different perspective of the chain, denoted by the processes $\{Y_t^{(j)}\}_{t\geq 1}$, with each $Y^{(j)} \in \mathcal{Y}^{(j)}$ $\forall j = 1, 2, \cdots, m$. Without loss of generality, let $\mathcal{X} = \{1, 2, \cdots, q\}$ and for each j, $\mathcal{Y}^{(j)} = \{1, 2, \cdots, \kappa_j\}$.

While the transition matrix of the hidden states $\mathbf{A}$ remains the same for all observers, their associated observation matrices may differ. If at least two observation matrices are distinct, i.e. there exists indices ℓ and ℓ' such that $\mathbf{B}^{(\ell)} \neq \mathbf{B}^{(\ell')}$, the set of independent observers are called *heterogeneous* observers, otherwise they are called *homogeneous* observers. We model the separate observation matrices $\mathbf{B}^{(j)}$ for each observer $j = 1, 2, \cdots, m$, each of size $q \times \kappa_j$ in the heterogeneous case, and $\mathbf{B}^{(j)} := \mathbf{B}$ for all j in the homogeneous case. In either setting, the HMM is described by the parameter set $\boldsymbol{\lambda}_{\mathrm{multi}} := \{\boldsymbol{\pi}; m, q, \{\kappa_j\}_{j=1}^m, \mathbf{A}, \{\mathbf{B}^{(j)}\}_{j=1}^m\}$, with appropriate modifications to the observation matrix depending on the setting. For the homogeneous setting, we let $\kappa_j := \kappa$ and $\mathbf{B}^{(j)} := \mathbf{B}$, $\forall j$.

As we shall see, whether the observers are heterogenous or homogenous makes a significant difference to the identifiability of the HMM.

A. Multi–letter tensor

Unlike the single observer scenario, there are multiple observations in a single time step. We first consider the heterogeneous case, where, without loss of generality, we define the multi-letter tensor $\boldsymbol{\mathcal{M}}_*$, an order $m+2$ tensor of dimensions $q \times q \times \kappa_1 \times \cdots \times \kappa_m$, with component matrices $[\mathbf{A}, \mathbf{I}_q, \{\mathbf{B}^{(j)}\}_{j=1}^m]$.

Just as in the case of a single observer, the multi-letter tensor is connected to the HMM via its matricisation. Let $\kappa' := \prod_{j=1}^m \kappa_j$. Thus, the joint probability of a particular observed sequence $\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N$, noting each observation is now a vector, may be described by

$$\begin{aligned} &P_{\boldsymbol{\lambda}}(Y_1 = \mathbf{y}_1, Y_2 = \mathbf{y}_2, \cdots, Y_N = \mathbf{y}_N) \\ &\qquad = \boldsymbol{\pi}\mathbf{W}_*\mathbf{E}(\mathbf{y}_1)\mathbf{W}_*\mathbf{E}(\mathbf{y}_2)\cdots\mathbf{W}_*\mathbf{E}(\mathbf{y}_N)\mathbf{1}_q, \end{aligned} \tag{4}$$

where $\mathbf{E}(k)$ is a $\kappa' q \times q$ matrix, divided to κ' row partitions, with the $q \times q$ identity matrix in the k-th row partition, $\mathbf{1}_q$ is the $q \times 1$ vector of ones, and

$$\begin{aligned} \mathbf{W}_* &:= \bigotimes_{j=1}^{m}{}^{\mathrm{row}}\,\mathbf{B}^{(j)}\otimes^{\mathrm{row}}\mathbf{A} \\ &= \mathbf{B}^{(1)}\otimes^{\mathrm{row}}\mathbf{B}^{(2)}\otimes^{\mathrm{row}}\cdots\otimes^{\mathrm{row}}\mathbf{B}^{(m)}\otimes^{\mathrm{row}}\mathbf{A}. \end{aligned} \tag{5}$$

In this formulation, all possible output sequences from space $\mathcal{Y}^{(1)} \times \mathcal{Y}^{(2)} \times \cdots \times \mathcal{Y}^{(m)}$ are ordered lexicographically, with $\mathbf{E}(\mathbf{y})$ selecting the correct position of $\mathbf{y}$ out of this ordering. Note that $\mathbf{W}_*$ is equivalent to $\boldsymbol{\mathcal{M}}_*$ since it is the second mode matricisation of the tensor. Similarly, in the homogeneous setting, let $\mathbf{W}_\circ$ have the same structure as (5), with $\mathbf{B}^{(j)} := \mathbf{B}$, $\forall j$, and $\kappa' = \kappa^m$. We have the following result for $\boldsymbol{\mathcal{M}}_*$.

Lemma 7: Assume $m \geq 2$. The heterogeneous multi-letter tensor $\boldsymbol{\mathcal{M}}_*$ is unique up to permutation and scaling of rows if and only if $\mathrm{krank}(\bigotimes_{j=1}^m{}^{\mathrm{row}}\,\mathbf{B}^{(j)}\otimes^{\mathrm{row}}\mathbf{A}) = q$.

Proof: The proof is similar to the proof of Lemma 6, extended to the multidimensional case, thus, we only need to sketch the proof here. We claim that it is necessary and sufficient that none of the non-trivial linear combinations of rows of $\bigotimes_{j=1}^{m}{}^{\text{row}} \mathbf{B}^{(j)} \otimes^{\text{row}} \mathbf{A}$ is expressed by a tensor product of two row vectors, i.e. $\text{krank}(\bigotimes_{j=1}^{m}{}^{\text{row}} \mathbf{B}^{(j)} \otimes^{\text{row}} \mathbf{A}) = q$. The chief ingredients are, (1) show a counterexample to proof necessity, where the same example from the proof of Lemma 6 can be used, appropriately modified to account for additional dimensions, to construct an alternative decomposition of $\boldsymbol{\mathcal{M}}_*$, and (2) for sufficiency, show that $\mathbf{I}_q$, the full rank component of $\boldsymbol{\mathcal{M}}_*$ satisfies Kruskal's permutation lemma. Then, the claim is established. ■

We must, however, be careful when the observers are independent and homogeneous. In this scenario, the above result no longer holds. Instead, the homogeneous setting is equivalent to the single observer setting, evidenced by the following result.

Lemma 8: It is necessary and sufficient that $\text{krank}(\mathbf{B} \otimes^{\text{row}} \mathbf{A}) = q$ for the homogeneous multi-letter tensor $\boldsymbol{\mathcal{M}}_\circ$ to be unique up to permutation and scaling of rows.

An intuitive explanation is that each additional component, $\mathbf{B}$, is exactly the same and do not provide sufficient variability for unique decomposition. Thus, even if the tensor $\boldsymbol{\mathcal{M}}_\circ$ is matricised in different ways, one can always find an alternative decomposition of $\boldsymbol{\mathcal{M}}_\circ$, similar to an example by Stegeman et al. [15] for the 3-way tensor. It is for this reason, decomposition-wise, $\boldsymbol{\mathcal{M}}_\circ$ is no different from the single letter tensor $\boldsymbol{\mathcal{L}}$.

B. Identifiability conditions

Theorem 9: Suppose the m observers are independent and homogeneous. For an HMM parameterised by $\boldsymbol{\lambda}_{\text{multi}}$ to be unique up to a scaling and permutation of states, it is necessary and sufficient that the per letter tensor of the HMM satisfies Lemma 8, i.e. $\text{krank}(\mathbf{B} \otimes^{\text{row}} \mathbf{A}) = q$.

Proof: The proof follows from the properties of $\boldsymbol{\mathcal{M}}_\circ$. Then, it is clear $\text{krank}(\mathbf{B} \otimes^{\text{row}} \mathbf{A}) = q$ if and only if $\text{krank}(\bigotimes^{m\ \text{row}} \mathbf{B} \otimes^{\text{row}} \mathbf{A}) = q$, from Lemma 8, otherwise an equivalent HMM can be constructed, such that the original HMM is no longer permutation and scaling indeterminate. ■

The result shows that the homogeneous setting is essentially equivalent to the single observer setting. As mentioned in [12], if an HMM is unidentifiable in the single observer case, it is also unidentifiable in the multiple independent homogenous observer case, as there is not enough variability in $\boldsymbol{\mathcal{M}}_\circ$. Thus, no matter how many independent homogeneous observers are present, if the model cannot be identified in the single observer setting, then the model remains unidentifiable.

We next turn our attention to the heterogeneous case, a consequence of Lemma 7.

Theorem 10: Suppose the m observers are independent and heterogenous. For an HMM parameterised by $\boldsymbol{\lambda}_{\text{multi}}$ to be unique up to a scaling and permutation of states, it is necessary and sufficient that the multi-letter tensor of the HMM satisfies Lemma 7, i.e. $\text{krank}(\bigotimes_{j=1}^{m}{}^{\text{row}} \mathbf{B}^{(j)} \otimes^{\text{row}} \mathbf{A}) = q$.

The extra dimensions of $\boldsymbol{\mathcal{M}}_*$ are related to the additional advantage of having multiple independent observers. Each single observer $j = 1, 2, \cdots, m$ is essentially restricted to a per letter tensor $\boldsymbol{\mathcal{L}}^{(j)}$ which may not satisfy Lemma 6 individually, but satisfies Lemma 7 when their observations are jointly considered. This proves a natural advantage multiple independent heterogeneous observers have over a single observer, used to great effect in [12].

VI. Conclusion

In this paper, we revisit the identifiability of hidden Markov models, via tensor decomposition, where there are well-known results regarding the permutation and scaling indeterminacy of tensors. We proved deterministic identifiability conditions of single and multi-observer HMMs using well–established results on the Kruskal rank. Our results are stronger than previous results, where only generic identifiability based on the regularity of the HMM is assumed. Future work includes using our framework to provide insights in the inference of the transition and observation matrices of HMMs and extend the work to dependent observers.

References

[1] E. S. Allman, C. Matias, and J. A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Ann. Stats.*, 37(6A):3099–3132, 2009.

[2] L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.*, 37:15541563, 1966.

[3] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, 41:164–171, 1970.

[4] D. Blackwell and L. Koopmans. On the identifiability problem for functions of finite Markov chains. *Ann. Math. Stat.*, 28(4):1011–1015, 1957.

[5] D. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proc. Nat. Acad. Sci.*, 100(5):2197–2202, March 2003.

[6] L. Finesso. *Consistent Estimation of the Order for Markov and Hidden Markov Chains*. PhD thesis, University of Maryland, 1990.

[7] E. J. Gilbert. On the identifiability problem for functions of finite Markov chains. *Ann. Math. Statist.*, 30:688697, 1959.

[8] T. Jiang and N. D. Sidiropoulos. Kruskal's permutation lemma and the identification of CANDECOMP/PARAFAC and bilinear models with constant modulus constraints. *IEEE Trans. Sig. Proc.*, 52(9):2625 – 2636, September 2004.

[9] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, September 2009.

[10] J. B. Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and Appl.*, 18(2):95–138, 1977.

[11] X. Li, M. Parizeau, and R. Plamondon. Training hidden Markov models with multiple observations: A combinatorial method. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(4):371–377, April 2000.

[12] H. X. Nguyen and M. Roughan. Improving hidden Markov model inferences with private data from multiple observers. *IEEE Sig. Proc. Letters*, 19(10):696–699, October 2012.

[13] T. Petrie. Probabilistic functions of finite state Markov chains. *Ann. Math. Statist.*, 40(1):97–115, 1969.

[14] N. D. Sidiropoulos and R. Bro. On the uniqueness of multilinear decomposition of N-way arrays. *J. Chemometrics*, 14(3):229–239, June 2000.

[15] A. Stegeman and N. Sidiropoulos. On Kruskal's uniqueness condition for Candecomp/Parafac decomposition. *Linear Algebra and Appl.*, 420(2–3):540–552, January 2007.

[16] P. Tune, H. X. Nguyen, and M. Roughan. Identifiability of Hidden Markov Models via tensor decomposition: Single and multiple observers. Technical report, University of Adelaide, 2012.