Fast and Consistent Learning of Hidden Markov Models by Incorporating Non-Consecutive Correlations

Robert Mattila, Cristian Rojas, Eric Moulines, Vikram Krishnamurthy, Bo Wahlberg

Fast and Consistent Learning of Hidden Markov Models by Incorporating Non-Consecutive Correlations

Robert Mattila [1] **Cristian R. Rojas** [1] **Eric Moulines** [2 3] **Vikram Krishnamurthy** [4] **Bo Wahlberg** [1]

Abstract

Can the parameters of a *hidden Markov model* (HMM) be estimated from a single sweep through the observations – and additionally, without being trapped at a local optimum in the likelihood surface? That is the premise of recent method of moments algorithms devised for HMMs. In these, correlations between consecutive pair- or triplet-wise observations are empirically estimated and used to compute estimates of the HMM parameters. Albeit computationally very attractive, the main drawback is that by restricting to only low-order correlations in the data, information is being neglected which results in a loss of accuracy (compared to standard maximum likelihood schemes). In this paper, we propose extending these methods (both pair- and triplet-based) by also including non-consecutive correlations in a way which does not significantly increase the computational cost (which scales linearly with the number of additional lags included). We prove strong consistency of the new methods, and demonstrate an improved performance in numerical experiments on both synthetic and real-world financial time-series datasets.

1. Introduction

The *hidden Markov model* (HMM) is a standard tool in statistical modeling of stochastic time-series (Cappé et al., 2005; Krishnamurthy, 2016). Despite its structural simplicity – a Markov chain observed via a noisy sensor –, the HMM has been successfully applied in a vast range of fields: from computational biology (Durbin, 1998; Vidyasagar, 2014) and speech recognition (Rabiner, 1989; Gales & Young, 2007) to finance (Mamon & Elliott, 2007; 2014) and human intent modeling (Yang et al., 1997; Xia et al., 2012), etc. Mathematically, an HMM is described by a latent finite-dimensional Markov chain $x_k \in \{1, \dots, X\}$ that evolves according to an $X \times X$ transition matrix P:

$$[P]_{ij} = \Pr[x_{k+1} = j | x_k = i], \tag{1}$$

where k denotes discrete time. The state x_k is observed, in the case of a finite observation alphabet $y_k \in \{1, \dots, Y\}$, via an $X \times Y$ observation matrix B:

$$[B]_{ij} = \Pr[y_k = j | x_k = i]. \tag{2}$$

In order to employ an HMM in any application that requires filtering or prediction, the parameters (1) and (2) have first to be determined – usually, via either domain-expertise or data-driven modeling. In this paper, we consider the latter.

Even though data-driven parameter estimation for HMMs has been studied for more than fifty years (Baum & Petrie, 1966), it remains a challenging problem in practice. The *de facto* standard methods employ iterative ("hill-climbing") local-search procedures that aim to maximize the likelihood of observed data (e.g., the *expectation-maximization* (EM), or Baum-Welch, algorithm) (Cappé et al., 2005; Krishnamurthy, 2016). In practice, these methods suffer from drawbacks related to convergence to bad local maxima (see Section 6.1 for an explicit numerical example) as well as slow convergence with an associated high computational cost.

In order to address such drawbacks, methods of moments have been introduced for HMMs (e.g., Chang, 1996; Mossel & Roch, 2005; Vanluyten et al., 2008; Lakshminarayanan & Raich, 2010; Hsu et al., 2012; Anandkumar et al., 2012; 2014; Kontorovich et al., 2013; Subakan et al., 2015; Tran et al., 2016; Mattila et al., 2017; Huang et al., 2018). These methods begin by estimating low-order correlations in the data, such as those of pairs $\Pr[y_k, y_{k+1}]$ or triplets $\Pr[y_k, y_{k+1}, y_{k+2}]$. The analytical relations between these correlations and the HMM parameters are then used "in reverse" to obtain empirical parameter estimates.

[1]Division of Decision and Control Systems, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden. [2]Centre de Mathématiques Appliquées de Polytechnique, Ecole Polytechnique, Paris, France. [3]HSE University, Moscow, Russia. [4]School of Electrical and Computer Engineering, Cornell University, Ithaca, New York, USA. Correspondence to: Robert Mattila <rmattila@kth.se>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

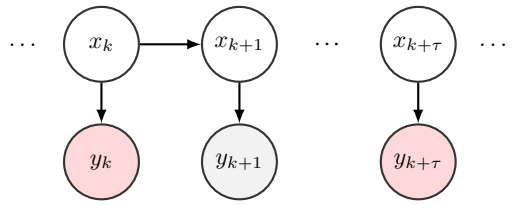

Figure 1. Graphical conditional-dependency structure of an HMM with (hidden) state variable x_k and corresponding noisy observation y_k. The red nodes are the second-order moments $\Pr[y_k, y_{k+\tau}]$ with $\tau \in \{1, 2, \ldots, \bar{\tau}\}$, that are used in Section 4 to estimate the HMM parameters. Previous work employs only the special case where the observations in the pair are consecutive: $\tau \in \{1\}$.

The consequences of restricting, effectively, to only short substrings in the observed data is double-edged: On the one hand, attractive properties include that *i)* they have very low computational cost – in practice, they are orders of magnitudes faster than conventional *maximum likelihood* (ML) schemes –, and *ii)* they are, under suitable assumptions, strongly consistent and the associated algorithms do not suffer from local optima. The main disadvantage on the other hand, compared to standard ML estimation, is the loss in statistical efficiency – information available in the observed data is neglected.

In this paper, we aim to reduce this gap in statistical efficiency while preserving the two attractive properties mentioned above. Our core idea is simple: *include a user-defined number $\bar{\tau}$ of lagged (i.e., non-consecutive) tuples in these procedures.* For example, in a pair-based method of moments (e.g., Vanluyten et al., 2008; Lakshminarayanan & Raich, 2010; Kontorovich et al., 2013; Subakan et al., 2015; Mattila et al., 2017; Huang et al., 2018), this corresponds to including correlations on the form $\Pr[y_k, y_{k+\tau}]$ with $\tau = 1, 2, \ldots, \bar{\tau}$, as illustrated in Fig. 1.

Despite the simplicity and intuitive appeal of the idea, it is not straight-forward to extend previous methods: care has to be taken to respect the two attractive properties mentioned above. In particular, this means avoiding non-convex problem formulations. Our key contribution is demonstrating how several, both pair- and triplet-based, methods of moments can be extended with this idea, while preserving the attractive properties of the methods they extend.

1.1. Main Results and Outline

In the main text, we considering pair-based (which we also refer to as second-order) methods of moments. This allows us to keep the presentation clear and concise; we can demonstrate the key idea, without being impeded by the tensor notation that is inherent in higher-order methods. In the supplementary material (Mattila, 2020, pp. 81–104), we also treat triplet-based (third-order) methods of moments.

In summary, the main results of this paper are:

- We derive expressions for non-consecutive HMM moments, and demonstrate how these can be incorporated in existing methods of moments (where the extensions only introduce steps invoking convex optimization);
- Our algorithms involve only a single sweep through the HMM dataset (in contrast to iterative ML algorithms that process the full dataset in each iteration), and take more information from the observed data into account than the previous methods they extend;
- Theoretically, we analyze the two principal attractive properties of our proposed estimators. First, we show that the computational complexity scales only linearly with the number of additional lags considered, and that the dominating cost is *independent* of the number of data-samples. Second, we prove that the estimators are strongly consistent;
- Numerical demonstrations of *i)* the improved accuracy of the proposed extensions compared to non-lagged methods of moments, and *ii)* the attractive run-times (up to two orders of magnitude faster than EM) on synthetic data;
- A numerical evaluation of the performance on a real-world financial time-series dataset.

The paper is organized as follows. Section 2 presents preliminaries related to HMMs and assumptions. In Section 3, we derive expressions for non-consecutive HMM moments. These are incorporated in second-order methods of moments in Section 4. A discussion of related work is provided in Section 5. Numerical experiments are performed in Section 6.

The supplementary material (Mattila, 2020, pp. 81–104) contains extensions of third-order methods, detailed proofs and additional numerical experiments.

2. Preliminaries

In this section, we first define the notation used in this paper, and subsequently outline necessary preliminaries for HMMs – complete treatments are available in, e.g., (Rabiner, 1989; Cappé et al., 2005; Krishnamurthy, 2016).

2.1. Notation

All vectors are column vectors unless transposed. The vector of all ones is denoted $\mathbb{1}$. The element at row i and column j of a matrix is $[\cdot]_{ij}$, and the element at position i of a vector is $[\cdot]_i$. Inequalities $(>, \geq, \leq, <)$ between vectors or matrices are interpreted elementwise. The vector operator $\mathrm{diag}(\cdot) : \mathbb{R}^n \to \mathbb{R}^{n \times n}$ gives the matrix where the vector

has been put on the diagonal, and all other elements are zero. $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm of a matrix. The indicator function $\mathrm{I}\{\cdot\}$ takes the value 1 if the expression $\cdot$ is fulfilled and 0 otherwise.

2.2. Hidden Markov Models

We consider a discrete-time *hidden Markov model* (HMM). It comprises a finite state Markov chain on the state space $\mathcal{X} = \{1, 2, \ldots, X\}$ with time-homogeneous $X \times X$ transition probability matrix

$$[P]_{ij} = \Pr[x_{k+1} = j | x_k = i]. \tag{3}$$

We denote by $\pi_0 \in \mathbb{R}^X$ and $\pi_\infty \in \mathbb{R}^X$, the initial and the stationary distributions, respectively, of the HMM (which exist under appropriate assumptions). An HMM is said to be *stationary* if $\pi_0 = \pi_\infty$.

We consider a discrete-valued observation process for the HMM.[1] Hence, observations are sampled from the finite set $\mathcal{Y} = \{1, 2, \ldots, Y\}$ according to the $X \times Y$ observation probability matrix

$$[B]_{ij} = \Pr[y_k = j | x_k = i]. \tag{4}$$

Note that both P and B are row-stochastic matrices (i.e., the elements on each row are non-negative and sum to one).

2.3. Problem Formulation

In order to employ an HMM for, e.g., filtering or prediction, its model parameters have first to be specified or estimated. The learning problem for HMMs that we consider is:

Problem 1. Given a sequence $y_1, \ldots, y_N$ of observations generated by a stationary HMM of known state and observation dimensions X and Y, estimate its parameters P and B.

In order to guarantee that this problem is well-posed, the following two assumptions are standard:

Assumption 1. The transition and observation matrices are elementwise strictly positive – that is, $P > 0$ and $B > 0$.

Assumption 2. The transition and observation matrices P and B, respectively, are full row rank.

Assumption 1 is a common assumption in statistical inference for HMMs (e.g., Baum & Petrie, 1966; Cappé et al., 2005) and serves as a proxy for ergodicity of the HMM. It implies that the underlying Markov chain is ergodic (irreducible and aperiodic). Assumption 2 is related to identifiability and is standard in methods of moments for HMMs (e.g., Hsu et al., 2012; Gassiat et al., 2016) – see (Huang et al., 2018) for a discussion on how it can be relaxed. The assumption of a stationary HMM can be fulfilled by discarding the first few data points; if the chain is mixing (Assumption 1), then it forgets its initial condition geometrically fast.

[1] For clarity of presentation, we consider only discrete observation spaces in the main text. We treat continuous-valued observation processes in the supplementary material (Mattila, 2020, pp. 81–104).

3. Moments in HMMs

In this section, we define moment conditions that we will employ to compute estimates of the HMM parameters via the method of moments. There are a number of potential candidates. In this work, the crucial quantities are the pair- and triplet-wise correlations between observations (albeit we postpone the treatment of the latter to the supplementary material; Mattila, 2020, pp. 81–104).

3.1. Second-Order Moments

Define the *lag-τ second-order moments* $M_2(k, \tau) \in \mathbb{R}^{Y \times Y}$ of the HMM as the matrices:

$$[M_2(k, \tau)]_{ij} \overset{\text{def.}}{=} \Pr[y_k = i, y_{k+\tau} = j], \tag{5}$$

where $i, j = 1, \ldots, Y$ and $\tau \geq 0$. In words: as the joint probabilities of pairs of observations spaced a distance τ apart in time.[2]

In terms of terminology, the reader should note that when we speak of *non-lagged* moments, we mean specifically the case that only τ up to $\tau = 1$ is used in an estimation procedure. In this case, the two observations in the pair (5) are at most consecutive: y_k and y_{k+1}. When we refer to *higher-order* or *lagged* moments, we mean that a whole range of values $\tau = 0, 1, 2, \ldots$ is employed. In this case, there is also a lag between the first observation y_k and the second observation $y_{k+\tau}$.

It is readily verified that the matrices (5) can be expressed in terms of the HMM parameters as

$$M_2(k, \tau) = B^T \operatorname{diag}((P^T)^k \pi_0) P^\tau B, \tag{6}$$

for $\tau > 0$, and

$$M_2(k, 0) = \operatorname{diag}(B^T (P^T)^k \pi_0). \tag{7}$$

For a stationary HMM (i.e., $\pi_0 = \pi_\infty$), these matrices do not depend on absolute time k and relations (6) and (7) reduce to

$$M_2(\tau) = B^T \operatorname{diag}(\pi_\infty) P^\tau B, \tag{8}$$

for $\tau > 0$, and

$$M_2(0) = \operatorname{diag}(B^T \pi_\infty), \tag{9}$$

respectively.

[2] The case $\tau = 0$ actually corresponds to *first-order moments* $[M_1(k)]_i \overset{\text{def.}}{=} \Pr[y_k = i]$, where $M_1(k) \in \mathbb{R}^Y$. However, for notational convenience in Section 4, we express these as a special case of second-order moments: $M_2(k, 0) = \operatorname{diag}(M_1(k))$.

4. Extension of Second-Order Methods of Moments to Include Non-Consecutive Pairs

In this section, we extend second-order methods of moments (e.g., Vanluyten et al., 2008; Lakshminarayanan & Raich, 2010; Kontorovich et al., 2013; Subakan et al., 2015; Mattila et al., 2017; Huang et al., 2018) to include non-consecutive pairs of observations. That is, pairwise probabilities $\Pr[y_k, y_{k+\tau}]$ for τ up to some number $\bar{\tau}$ of lags (specified by the user). The section is concluded with a discussion on statistical properties, and indications for how to choose $\bar{\tau}$.

4.1. Moment-Matching from Second-Order Moments

The full learning problems for HMMs from *consecutive* moments $\Pr[y_k, y_{k+1}]$ using relation (8) is:

$$\begin{aligned} \min_{\substack{\hat{\pi}_\infty \in \mathbb{R}^X, \hat{P} \in \mathbb{R}^{X\times X}, \\ \hat{B} \in \mathbb{R}^{X\times Y}}} \quad & \|\hat{M}_2(1) - \hat{B}^T \operatorname{diag}(\hat{\pi}_\infty)\hat{P}\hat{B}\|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & \hat{\pi}_\infty \geq 0, \quad \mathbb{1}^T\hat{\pi}_\infty = 1, \\ & \hat{P} \geq 0, \quad \hat{P}\mathbb{1} = \mathbb{1}, \\ & \hat{B} \geq 0, \quad \hat{B}\mathbb{1} = \mathbb{1}, \\ & \hat{\pi}_\infty = \hat{P}^T\hat{\pi}_\infty, \end{aligned} \tag{10}$$

where $\hat{M}_2(1)$ is an empirical estimate of $M_2(1)$ and the first constraints are due to $\hat{P}, \hat{\pi}_\infty, \hat{B}$ representing probabilities, and the last since $\hat{\pi}_\infty$ is a stationary distribution of $\hat{P}$.

This problem is non-convex and previous methods either use alternating (block-coordinate descent) schemes (e.g., Vanluyten et al., 2008; Lakshminarayanan & Raich, 2010; Huang et al., 2018) that:

i) fix $\hat{B}$ and optimize for $\hat{P}, \hat{\pi}_\infty$,

ii) fix $\hat{P}, \hat{\pi}_\infty$ and optimize for $\hat{B}$, and repeat;

or decouple the problem (e.g., Kontorovich et al., 2013; Subakan et al., 2015; Mattila et al., 2017) to first estimate B separately[3] and then perform only step *i)*.

Below, we demonstrate how step *i)* can be improved by including lagged moments – the complementary step (estimating B) is identical. Our aim is hence, in this section, to estimate the transition matrix P:

Assumption 3. The observation matrix B is given.

[3]In *parametric-output HMMs*, as a first step, the output parameters are estimated via a general mixture model learner, and as a second step, the identification of the transition matrix P becomes identification of a known-sensor HMM. These HMMs are most natural to consider in general observation spaces, which include *Gaussian HMMs*. We provide an extended discussion in the supplementary material (Mattila, 2020, pp. 81–104).

Remark 1. It should be noted that this special case is additionally motivated by any application in which the sensor is designed by the user. Consider, for example, a target-tracking system. The sensor specifications can be determined prior to deployment in controlled trials, which gives the operator knowledge about the sensor equipment (i.e., the observation matrix B). The maneuvering strategy of the tracked target (i.e., the transition matrix P) is unknown and has to be estimated.

4.2. Estimating the Transition Matrix with Non-Consecutive Lags

Even in a known-sensor HMM (Assumption 3) or, equivalently, in step *i)* above, the moment-matching problem is non-convex[4]:

$$\begin{aligned} \min_{\hat{\pi}_\infty \in \mathbb{R}^X, \hat{P} \in \mathbb{R}^{X\times X}} \quad & \|\hat{M}_2(1) - B^T \operatorname{diag}(\hat{\pi}_\infty)\hat{P}B\|_{\mathrm{F}}^2 \\ \text{s.t.} \quad & \hat{\pi}_\infty \geq 0, \quad \mathbb{1}^T\hat{\pi}_\infty = 1, \\ & \hat{P} \geq 0, \quad \hat{P}\mathbb{1} = \mathbb{1}, \\ & \hat{\pi}_\infty = \hat{P}^T\hat{\pi}_\infty, \end{aligned} \tag{11}$$

due to the products between $\hat{P}$ and $\hat{\pi}_\infty$.

Including non-consecutive lags via (8) and jointly minimizing all moment conditions leads to the objective function:

$$\sum_{\tau=1}^{\bar{\tau}} \|\hat{M}_2(\tau) - B^T \operatorname{diag}(\hat{\pi}_\infty)\hat{P}^\tau B\|_{\mathrm{F}}^2, \tag{12}$$

subject to the same constraints. The additional non-convexity (due to the higher-order powers of the transition matrix) makes it computationally demanding to compute a global solution. In particular, non-convex optimization commonly relies on local-search heuristics (Jain & Kar, 2017) much alike those employed in standard ML estimation for HMMs.

It has been shown (e.g., Kontorovich et al., 2013; Mattila et al., 2017) that the *consecutive* problem (11) can be reformulated as a convex optimization problem. Below, we extend this approach to the *non-consecutive* problem (12) and propose a novel sequential method that involves only convex (quadratic) optimization.

Step 1. Estimating Second-Order Moments The left-hand sides of equations (8) and (9) are readily estimated from data via the empirical estimator

$$[\hat{M}_2(\tau)]_{ij} \overset{\text{def.}}{=} \frac{1}{N-\tau} \sum_{k=1}^{N-\tau} \mathrm{I}\{y_k = i, y_{k+\tau} = j\}, \tag{13}$$

[4]Unsurprisingly, also the likelihood of a known-sensor HMM (Assumption 3) can be multi-modal and cause problems for local-search ML algorithms – see the example in Section 6.1.

for $\tau = 0, 1, \ldots, \bar{\tau}$. The key step in a method of moments is decomposing these empirical estimates into the (unknown) system parameters on the right-hand sides (8) and (9), which we demonstrate next.

Step 2. Matching Moments In equation (8), denote

$$A(\tau) \stackrel{\text{def.}}{=} \text{diag}(\pi_\infty) P^\tau, \tag{14}$$

for $\tau \geq 0$.[5] This implies, by definition, that

$$A(\tau + 1) = A(\tau) P, \tag{15}$$

and that we can rewrite equation (8) as

$$M_2(\tau) = B^T A(\tau) B. \tag{16}$$

As mentioned above, let the number of lagged pairs in the estimation procedure be $\bar{\tau} \geq 1$. We propose the following convex (quadratic) procedure to perform moment matching:

(i) Minimize the mismatch in equation (9) by solving

$$\begin{aligned} \min_{\hat{\pi}_\infty \in \mathbb{R}^X} \quad & \|\hat{M}_2(0) - \text{diag}(B^T \hat{\pi}_\infty)\|_{\text{F}}^2 \\ \text{s.t.} \quad & \hat{\pi}_\infty \geq 0, \quad \mathbb{1}^T \hat{\pi}_\infty = 1, \end{aligned} \tag{17}$$

and set[6]

$$\hat{A}(0) = \text{diag}(\hat{\pi}_\infty). \tag{18}$$

(ii) For $\tau = 1, \ldots, \bar{\tau}$, minimize the mismatch in equation (8) by solving

$$\begin{aligned} \min_{\hat{P}(\tau) \in \mathbb{R}^{X \times X}} \quad & \|\hat{M}_2(\tau) - B^T \hat{A}(\tau - 1) \hat{P}(\tau) B\|_{\text{F}}^2 \\ \text{s.t.} \quad & \hat{P}(\tau) \geq 0, \quad \hat{P}(\tau) \mathbb{1} = \mathbb{1}, \end{aligned} \tag{19}$$

and set

$$\hat{A}(\tau) = \hat{A}(\tau - 1) \hat{P}(\tau). \tag{20}$$

In essence, the optimization problems (17) and (19) minimize the discrepancy between the empirical estimate of $\hat{M}_2(\tau)$ and its analytical expression.[7] The output of algorithm (17)–(20) is a sequence $\hat{A}(0), \ldots, \hat{A}(\bar{\tau})$, and involves solving $\bar{\tau} + 1$ convex (quadratic) optimization problems that are *independent of the data-size* N.

[5] The matrix $A(\tau)$ can be interpreted as the joint state distribution lagged τ time steps apart: $[A(\tau)]_{ij} = \Pr[x_k = i, x_{k+\tau} = j]$.

[6] In (18), $\hat{\pi}_\infty$ denotes the minimizing argument of (17); and similarly for (19) and (20) below.

[7] We employ the Frobenius norm for simplicity – other choices (or, weightings) could improve accuracy (see, e.g., Gourieroux & Monfort, 1995).

Step 3. Reconstructing the Transition Matrix P In order to construct an estimate of the transition matrix P, we utilize the shift-relation (15) in a least-squares fit (incorporating the information from every lag):

$$\begin{aligned} \min_{\hat{P} \in \mathbb{R}^{X \times X}} \quad & \left\| \begin{bmatrix} \hat{A}(0) \\ \vdots \\ \hat{A}(\bar{\tau} - 1) \end{bmatrix} \hat{P} - \begin{bmatrix} \hat{A}(1) \\ \vdots \\ \hat{A}(\bar{\tau}) \end{bmatrix} \right\|_{\text{F}}^2 \\ \text{s.t.} \quad & \hat{P} \geq 0, \quad \hat{P} \mathbb{1} = \mathbb{1}. \end{aligned} \tag{21}$$

4.2.1. Summary of Second-Order Algorithm

To summarize, the proposed algorithm involves *i)* estimating the moment matrices using the estimator (13), *ii)* solving optimization problems (17) and (19), and *iii)* estimating P by solving the least-squares problem (21). As discussed in Section 4.1, the observation matrix is either separately or iteratively estimated.

The special case $\bar{\tau} = 1$ reduces the algorithm to the form (11) which is considered in (e.g., Vanluyten et al., 2008; Lakshminarayanan & Raich, 2010; Kontorovich et al., 2013; Subakan et al., 2015; Mattila et al., 2017; Huang et al., 2018). The proposed method can, as such, be seen as an extension by including non-consecutive lags (i.e., more information from the observed data) in the estimation procedure.

4.3. Analysis of the Proposed Second-Order Algorithm

In this section, we analyze the multiple-lag method of moments estimator proposed above.

4.3.1. Computational Cost

In terms of computational cost, the procedure involves *i)* $\bar{\tau} + 1$ sliding-window estimators (13), and *ii)* solving $\bar{\tau} + 1$ convex (quadratic) optimization problems of size X^2 (that do not depend on the data-size N). In other words, *the dominating computational cost of the procedure is independent of* N, and scales linearly with the number of lags $\bar{\tau}$ considered. In comparison, each iteration of EM has a complexity $\mathcal{O}(X^2 N)$. This can be prohibitively expensive for large N, especially when many iterations are required for convergence.

4.3.2. Consistency

In terms of statistical properties of the proposed method, the main theoretical result of this section is the following.

Theorem 1. *Assume that the observations are generated by an HMM of known state dimension, whose transition and observation matrices satisfy Assumptions 1, 2, and 3. Then, the estimate of the transition matrix P resulting from the algorithm in Section 4.2 is strongly consistent – that is, the*

estimate will converge to the true value with probability one as the number of samples $N \to \infty$.

The theorem follows by showing that the moment matrices $\hat{M}_2(\tau)$ converge (using a law of large numbers; Cappé et al. 2005), and subsequently that the $\hat{A}(\tau)$'s and $\hat{P}$ converge (invoking convergence in minimization; Rockafellar & Wets 1998).[8]

The importance of Theorem 1 is to assure that the proposed computationally tractable algorithm – recall that a direct approach for including non-consecutive lags results in the non-convex problem (12) – is statistically sound; as the data-size grows, we will obtain the true transition matrix.

4.3.3. Choice of Number of Lags

Our second theoretical result serves as a guide for choosing the number of lags $\bar{\tau}$ to include. The matrix $A(\tau)$ defined in (14) converges to $\pi_\infty \pi_\infty^T$ (at a geometric rate determined by the second largest eigenvalue of the underlying Markov chain):

Proposition 1. *Consider an HMM satisfying Assumptions 1 and 2. The corresponding matrix $A(\tau)$, defined in (14), converges as*

$$A(\tau) - A(\tau+1) = \mathcal{O}(\tau^{m_2-1}|\lambda_2|^\tau), \qquad (22)$$

where λ_2 is the second largest eigenvalue (modulus) of the transition matrix P and m_2 its algebraic multiplicity.

This means that $A(\tau+1) \approx A(\tau)$ for large τ, and that increasing $\bar{\tau}$ has diminishing returns since the additional rows introduced in (21) provide incrementally less new information. Hence, a guideline is to choose $\bar{\tau}$ on the order of the time-constant of the Markov chain underlying the HMM. This is illustrated by numerical experiments in Section 6.

5. Related Work

The most widely used methods for parameter estimation in HMMs are schemes that iteratively aim to maximize the likelihood of observed data; the *expectation-maximization* (EM), or Baum-Welch, algorithm as well as direct optimization via Newton-Raphson and variants (Cappé et al., 2005; Krishnamurthy, 2016). These local-search procedures require careful initialization to avoid problems with local optima (see Section 6.1). Moreover, the computational cost can be prohibitively high for large datasets.

An alternative to ML estimation is the method of moments, which was originally devised to identify mixtures of univariate Gaussians (Pearson, 1894) by selecting the parameters of the distribution so as to equate empirical estimates of various-order moments. It has since then been generalized and adapted to a vast range of model structures (e.g., Hansen, 1982; Gourieroux & Monfort, 1995; Hall, 2005). In terms of HMMs, methods of moments are commonly based on occurrence-probabilities of different substrings of observation sequences. These have recently received much interest due to their computational attractiveness, as well as consistency guarantees under suitable assumptions. Many recent methods (Vanluyten et al., 2008; Lakshminarayanan & Raich, 2010; Anandkumar et al., 2012; Hsu et al., 2012; Kontorovich et al., 2013; Subakan et al., 2015; Anandkumar et al., 2014; Mattila et al., 2017; Huang et al., 2018) are based on low-order correlations between *consecutive* observations.

Methods based on pairwise occurrences include (Vanluyten et al., 2008; Lakshminarayanan & Raich, 2010; Kontorovich et al., 2013; Subakan et al., 2015; Mattila et al., 2017; Huang et al., 2018). A concern with these methods is that HMMs are, in general, not identifiable from only their pairwise (second-order) moments (Chang, 1996; Anandkumar et al., 2012; Huang et al., 2018) and (Mattila, 2020, Appendix 4.F). A notable exception is parametric-output HMMs (e.g., Gaussian HMMs). (Kontorovich et al., 2013) demonstrated that these can be treated in a sequential fashion. First, the observation likelihoods are estimated using a mixture model learner, and subsequently, the transition probabilities are estimated in a known-sensor HMM identification setting. Our contribution improves the estimation of the transition probabilities in this framework.

A popular branch of triplet-based (third-order) methods (e.g., Mossel & Roch, 2005; Hsu et al., 2012; Anandkumar et al., 2012; 2014) originates in (Chang, 1996). In these methods, it is shown that the moment-matching problem can be performed via a spectral (i.e., eigenvalue) factorization. It is noted in (Anandkumar et al., 2012) that the properties of the method hold for higher-order (i.e., non-consecutive) lags, but this is not explored at any depth – the topic of the present paper. Due to how the estimates are computed (via a spectral factorization), it is possible that the parameter estimates, especially in the small-sample regime, land outside the feasible parameter space (i.e., probabilities may not be non-negative or sum-to-one). Various methods have been proposed to address these problems; from truncation and projection to exterior point methods (Shaban et al., 2015). These problems are alleviated by making the estimates more accurate according to our proposed extensions (details can be found in the supplementary material; Mattila, 2020, pp. 81–104).

There also exists a number of related methods (e.g., Hjalmarsson & Ninness, 1998; Anderson, 1999; Vidyasagar, 2006; Vanluyten, 2008; Andersson & Rydén, 2009; Finesso et al., 2010; Cybenko & Crespi, 2011; Vidyasagar, 2011;

[8] A complete proof can be found in the supplementary material (Mattila, 2020, pp. 81–104).

Hsu et al., 2012; Vidyasagar, 2014; Tran et al., 2016) that estimate alternative parametrizations of an HMM and/or are based on longer substrings. In this work, we limit the scope to the methods discussed in the previous two paragraphs due to their recent popularity and since they *i)* estimate the standard parametrization of an HMM, and *ii)* use only substrings of length two and three (that are easier to estimate from a fixed-size dataset;[9] see Huang et al., 2018). It should, however, be noted that (Hjalmarsson & Ninness, 1998; Andersson & Rydén, 2009) employ lagged pairwise correlations, but do not enforce non-negativity constraints on the HMM parameters.

Restricting to short substrings in the data (approximate summary statistics) is key to the computational gain of these methods. However, it also leads to a loss of accuracy due to not taking all information available in the data into account. In order to reduce the resulting gap in statistical efficiency (compared to iterative ML estimation) various methods have been proposed. Thanks to the strong consistency and low computational cost of the method of moments estimate, it can effectively be used as a first approximation of the ML estimate. In (Kontorovich et al., 2013; Mattila et al., 2017; Balle et al., 2014; Zhang et al., 2016), estimates from methods of moments are used as initializers for ML schemes, and in (Tran et al., 2016) an iterative reweighing scheme is proposed. These hybrid approaches can, however, still present computational challenges. In this paper, we avoid potentially costly iterative schemes completely in favor of including more information in the moment conditions. However, employing the resulting estimate as an initializer for an ML schemes is, of course, still a valid option.

6. Numerical Evaluation

In this section, we evaluate the proposed second-order method of moments in numerical experiments. More details and numerical experiments for third-order methods can be found in the supplementary material (Mattila, 2020, pp. 81–104). All simulations in this section were run in MATLAB R2017a on a 3.1 GHz CPU.

6.1. Convergence of Local-Search Procedures

We begin by illustrating the potential algorithmic issues that standard local-search procedures aiming to maximize the likelihood function can suffer from. In order to be able to visualize the likelihood surface, we consider a known-sensor HMM (Assumption 3) with $X = 2$; that is, the observation matrix B is known and we aim to estimate the two unknown parameters of the transition matrix P.

[9] The number of occurrence-probabilities grows exponentially with substring length; there are Y^2 pairwise probabilities, Y^3 tripletwise, etc.

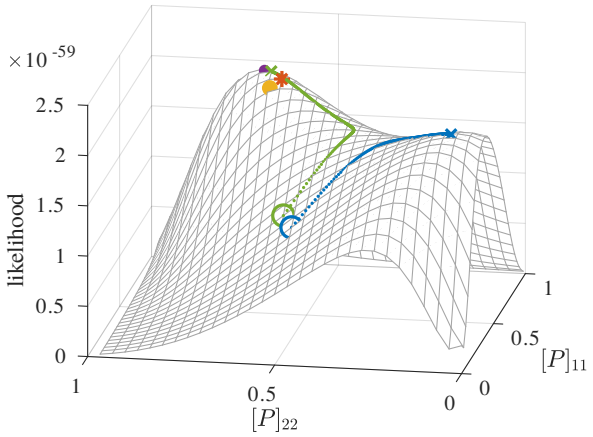

Figure 2. The likelihood function of data realized by a known-sensor HMM (Assumption 3). The true parameters are marked with a red star. The iterates from a local-search procedure started in two different initial points (marked with blue and green circles) converge to two different local optima; only the green trajectory computes the maximum likelihood estimate. Estimates from the method of moments (Section 4.2) are marked with solid purple and yellow circles for $\bar{\tau} = 1$ and $\bar{\tau} = 3$, respectively.

The standard MATLAB implementation of the *expectation-maximization* (EM) algorithm for HMMs (‘`hmmtrain`’) was employed (modified to account for Assumption 3). It was initialized in two different starting points, marked with large hollow circles in Fig. 2. As can be seen in the figure, even though the starting points are close, the iterates (in blue and green) converge to two different local optima – only the green iterates converge to the ML estimate (i.e., the global optimum). In the figure, we have marked the true parameter values with a red star.

In contrast, the method of moments (Section 4.2) does not suffer from bad initializations due to it invoking only convex optimization. The estimates computed using $\bar{\tau} = 1$ and $\bar{\tau} = 3$ are marked with purple and yellow solid circles, respectively.

6.2. Performance on Synthetic Data

We now evaluate how different parameters influence the algorithm in controlled trials on synthetic data. We consider an $X = Y = 3$ known-sensor HMM.

Remark 2. Due to the data-size N, the system dimensions X and Y can be of modest scale and still cause severe computational issues – recall that each EM-iteration has complexity $\mathcal{O}(X^2N)$. In gene-sequencing applications, the state and observation spaces are related to the four nucleotides A, T, C and G of *deoxyribonucleic acid* (DNA) and the datasets are enormous (e.g., the human genome consists of roughly 10^9 nucleotides) – see (Durbin, 1998; Vidyasagar, 2014).

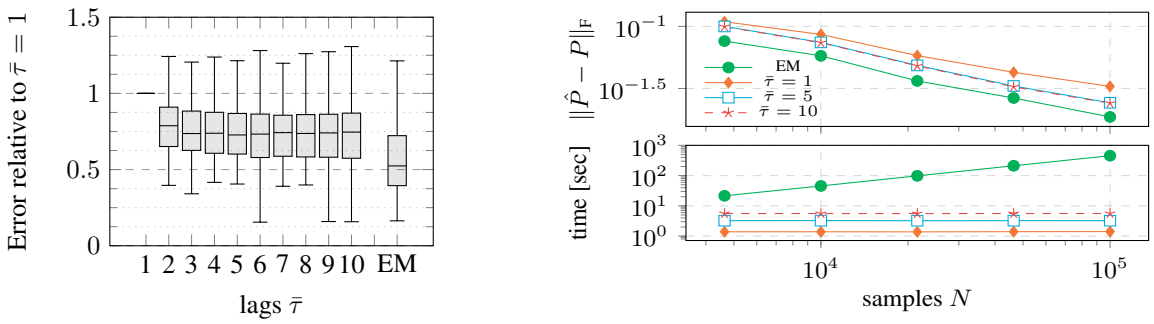

Figure 3. Left: Relative error of the proposed multiple-lag method of moments estimator for different lags $\bar{\tau}$ at $N = 10^5$ samples. Each box[11] is the result of 100 independent simulations. *Right:* Errors and run-times for various sample sizes N and lags $\bar{\tau}$. Each point is the average of 100 independent simulations. The figures demonstrate that including non-consecutive lags increases the accuracy of the estimate, while preserving the orders-of-magnitude faster run-times compared to the standard EM algorithm (note the logarithmic scale).

As before, the standard MATLAB implementation of EM was employed (modified to account for Assumption 3), but now initialized in the true parameter values (to avoid convergence to bad local optima).

6.2.1. Number of Lags $\bar{\tau}$

In the left plot of Fig. 3, we generated 100 independent realizations of $N = 10^5$ observations from the HMM. The errors, in the estimate of P, relative to a non-lagged ($\bar{\tau} = 1$) method of moments (e.g., Vanluyten et al., 2008; Lakshminarayanan & Raich, 2010; Kontorovich et al., 2013; Subakan et al., 2015; Mattila et al., 2017) are plotted for varying $\bar{\tau}$, as well as for the EM algorithm.

In the figure, there is on average a 25% improvement in error by including multiple lags. The second largest eigenvalue of the underlying Markov chain was 0.6, implying a time-constant of $1/(1-0.6) = 2.5$. As was noted in Proposition 1, the benefits of including higher-order lags become negligible as $A(\tau)$ reaches stationarity. In the plot, this can be seen by the stagnation of improvement after $\bar{\tau} \approx 3$ (roughly the time-constant).

6.2.2. Sample-size N

In the right plots of Fig. 3, we show the average run-times and errors of 100 independent realizations for a varying number of samples N. The attractive run-time of the method of moments is clearly visible in the bottom plot: at 10^5 samples, there are almost two orders of magnitude of difference compared to EM. Moreover, since only a single (counting) pass through the data is required, the run-time is roughly constant in N. The improved accuracy is visible in the top plot by the reduced gap to EM when including lagged pairs.

[11] Displayed are the median, the lower and upper quartiles, as well as, with whiskers, the (smallest) largest data value which is (larger) smaller than the (lower) upper quartile plus 1.5 times the inter-quartile-range.

Again, the diminishing returns (Proposition 1) are apparent by the overlap of the $\bar{\tau} = 5$ and $\bar{\tau} = 10$ curves.

6.3. Estimating Regimes in Financial Markets

We now illustrate the performance of the extended second-order method on real-world data where, in contrast to before, *i)* the observation process is continuous-valued and *ii)* the analog of Assumption 3 does not hold directly. The specific details on how to extend the method in Section 4 (using the decoupling approach of Kontorovich et al., 2013) to *unknown*-sensor HMMs on general observation spaces are provided in the supplementary material (Mattila, 2020, Appendix 4.D).

6.3.1. Background

Regime-switching market models based on HMMs (e.g., Turner et al., 1989; Hamilton, 1990; Rydén et al., 1998; Bulla, 2011; Dias et al., 2015; Kole & van Dijk, 2017) help economists interpret and analyze past and current market conditions, as well as perform market forecasting for use in, e.g., portfolio allocation (Yin & Zhou, 2004; Elliott et al., 2010; Nystrup et al., 2018). Crucial to the applicability of the models are their accuracy, as well as the time required for model estimation and recalibration (which becomes increasingly important in high-frequency applications where the datasets are large and sampling times are on the order of milliseconds).

6.3.2. Setup

As in (Rydén et al., 1998; Kole & van Dijk, 2017; Nystrup et al., 2018), we consider a two-regime (state) HMM which models log-returns as (conditionally) Gaussian distributed. We consider the publicly available weekly closing price of the Apple stock (`AAPL`) during 38 years; from 19th December 1980 to 25th April 2018. We chose the number of lags as $\bar{\tau} = 10$, since we expect trends to persist on a time-scale

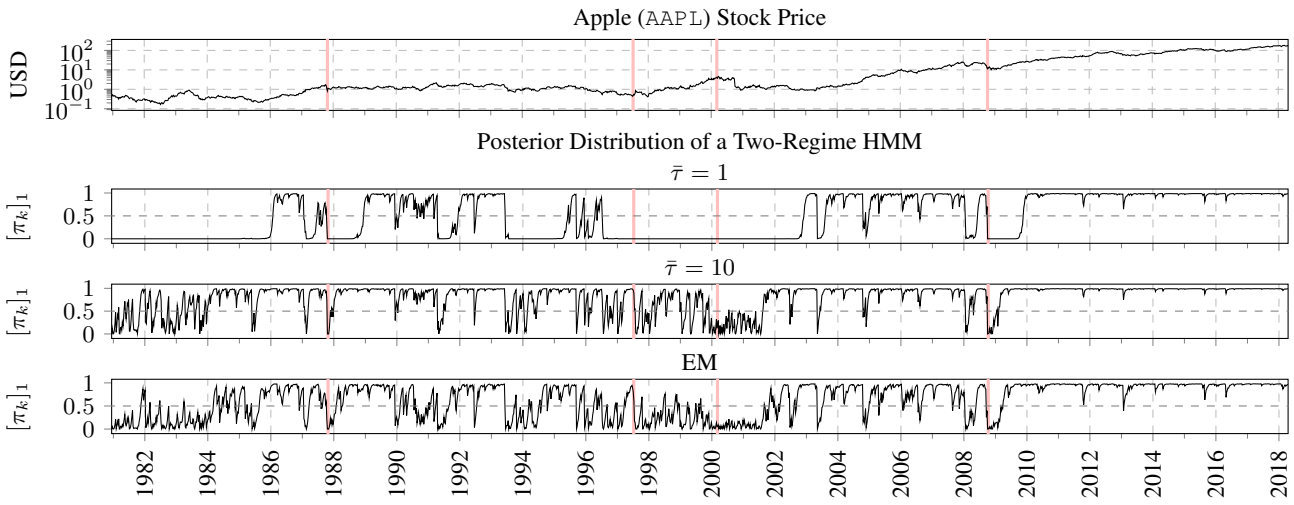

Figure 4. Top: The stock price of Apple (AAPL) in USD. *Bottom (three):* The posterior distribution (23) of a two-regime HMM with weekly log-returns as observations. The vertical red lines mark: the Black Monday ('87), Steve Jobs's return to Apple ('97), the burst of the dotcom-bubble ('00), and the global financial crisis ('08). The resemblance of the filter distributions computed with the models estimated using EM and the multiple-lag method ($\bar{\tau}=10$) should be noted.

of months. As benchmarks, we compare against EM and a non-lagged method of moments ($\bar{\tau} = 1$).

6.3.3. Results

All three methods ($\bar{\tau} = 1$, $\bar{\tau} = 10$, EM) yield transition matrices and observation likelihoods that can be interpreted as "bull-and-bear" market models: state 1 corresponds to optimistic market conditions ("bull"), and state 2 corresponds to pessimistic and volatile conditions ("bear").

To evaluate the qualities of the resulting models, we computed the posterior distribution

$$[\pi_k]_i = \Pr[x_k = i | y_1, \ldots, y_k], \quad i \in \{1, 2\}, \tag{23}$$

using an HMM filter (see, e.g., Krishnamurthy, 2016) with the parameters resulting from each one of the three methods. The posterior distribution (23) can be used to analyze and monitor market conditions in real-time (e.g., Bulla, 2011; Kole & van Dijk, 2017).

The resulting posterior distributions are plotted in Fig. 4. From the posteriors, it is, for example, possible to infer the change in market conditions related to the Black Monday ('87), Steve Jobs's return to Apple ('97), the burst of the dotcom-bubble ('00), and the start of the global financial crisis ('08), which we have marked in Fig. 4 with red vertical lines. The improved accuracy can be clearly seen by the high resemblance of the posterior distributions corresponding to the multiple-lag method ($\bar{\tau} = 10$) and that of EM. The distribution corresponding to the non-lagged method ($\bar{\tau} = 1$), in comparison, lacks many features.

7. Conclusion

In this paper, we have demonstrated how recent methods of moments for HMMs can be extended by incorporating non-consecutive observation tuples (pairs and/or triplets – see the supplementary material for details on the latter; Mattila, 2020, pp. 81–104). Our proposed extensions allow us to extract more information from the observed data, yielding more accurate estimates, while preserving the attractive computational and statistical properties of this type of estimator. The algorithms require only a single pass through the dataset (in contrast, the standard EM algorithm has a computational cost $\mathcal{O}(X^2N)$ per iteration). In practice, this is reflected in orders-of-magnitude faster run-times. Moreover, due to the non-iterative nature of these methods, their run-time can be better predicted in advance, making them suitable for real-time applications. We demonstrated improved accuracy and run-times in numerical experiments on both synthetic and real-world data.

In future work, it would be of interest to apply the proposed methodology for estimation of Markov-switched autoregressive models and jump Markov linear systems (e.g., Krishnamurthy & Rydén, 1998; Krishnamurthy, 2016; Cappé et al., 2005). It is also worthwhile studying what is to be gained by employing an optimal weighting in (17), (19) and (21), as well as quantifying the sample-efficiency through concentration inequalities for dependent random variables (e.g., Kontorovich & Ramanan, 2008), or analyzing the asymptotic covariance (e.g., Hansen, 1985). Moreover, employing method of moments estimates as initializers for iterative ML schemes is a promising hybrid that can combine the advantages of both approaches.

Acknowledgements

This research was supported in part by the Swedish Research Council (NewLEADS, 2016-06079, 2018-03438), the Wallenberg AI, Autonomous Systems and Software Program (WASP), the U.S. Army Research Office under grant W911NF-19-1-0365, the U.S. Air Force Office of Scientific Research under grant FA9550-18-1-0007, and the National Science Foundation under grant 1714180.

References

Anandkumar, A., Hsu, D., and Kakade, S. M. A method of moments for mixture models and hidden Markov models. In *Proceedings of the 25th Conference on Learning Theory (COLT'12)*, pp. 33.1–33.34, 2012.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

Anderson, B. D. O. The realization problem for hidden Markov models. *Mathematics of Control, Signals and Systems*, 12(1):80–120, 1999.

Andersson, S. and Rydén, T. Subspace estimation and prediction methods for hidden Markov models. *The Annals of Statistics*, 37(6B):4131–4152, 2009.

Balle, B., Hamilton, W. L., and Pineau, J. Methods of moments for learning stochastic languages: unified presentation and empirical comparison. In *Proceedings of the 31st International Conference on Machine Learning (ICML'14)*, pp. 1386–1394, 2014.

Baum, L. E. and Petrie, T. Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563, 1966.

Bulla, J. Hidden Markov models with t components. Increased persistence and other aspects. *Quantitative Finance*, 11(3):459–475, 2011. doi: 10.1080/14697681003685563.

Cappé, O., Moulines, E., and Rydén, T. *Inference in Hidden Markov Models*. Springer, 2005. ISBN 0387402640.

Chang, J. T. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Mathematical Biosciences*, 137(1):51–73, 1996.

Cybenko, G. and Crespi, V. Learning hidden Markov models using nonnegative matrix factorization. *IEEE Transactions on Information Theory*, 57(6):3963–3970, 2011. ISSN 0018-9448, 1557-9654.

Dias, J. G., Vermunt, J. K., and Ramos, S. Clustering financial time series: new insights from an extended hidden Markov model. *European Journal of Operational Research*, 243(3):852–864, 2015.

Durbin, R. *Biological Sequence Analysis: Probabalistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

Elliott, R. J., Siu, T. K., and Badescu, A. On mean-variance portfolio selection under a hidden Markovian regime-switching model. *Economic Modelling*, 27(3):678–686, 2010.

Finesso, L., Grassi, A., and Spreij, P. Approximation of stationary processes by hidden Markov models. *Mathematics of Control, Signals, and Systems*, 22(1):1–22, 2010. ISSN 0932-4194, 1435-568X. doi: 10.1007/s00498-010-0050-7.

Gales, M. and Young, S. The application of hidden Markov models in speech recognition. *Foundations and Trends in Signal Processing*, 1(3):195–304, 2007.

Gassiat, É., Cleynen, A., and Robin, S. Inference in finite state space non parametric hidden Markov models and applications. *Statistics and Computing*, 26(1-2):61–71, 2016.

Gourieroux, C. and Monfort, A. *Statistics and Econometric Models*. Cambridge University Press, 1995.

Hall, A. R. *Generalized Method of Moments*. Oxford University Press, 2005.

Hamilton, J. D. Analysis of time series subject to changes in regime. *Journal of Econometrics*, 45(1-2):39–70, 1990.

Hansen, L. P. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pp. 1029–1054, 1982.

Hansen, L. P. A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators. *Journal of Econometrics*, 30(1-2): 203–238, 1985.

Hjalmarsson, H. and Ninness, B. Fast, non-iterative estimation of hidden Markov models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'98)*, volume 4, pp. 2253–2256, 1998.

Hsu, D., Kakade, S. M., and Zhang, T. A spectral algorithm for learning hidden Markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

Huang, K., Fu, X., and Sidiropoulos, N. Learning hidden Markov models from pairwise co-occurrences with application to topic modeling. In *Proceedings of the 35th International Conference on Machine Learning (ICML'18)*, pp. 2068–2077, 2018.

Jain, P. and Kar, P. Non-convex optimization for machine learning. *Foundations and Trends in Machine Learning*, 10(3-4):142–336, 2017.

Kole, E. and van Dijk, D. How to identify and forecast bull and bear markets? *Journal of Applied Econometrics*, 32 (1):120–139, 2017.

Kontorovich, L. A. and Ramanan, K. Concentration inequalities for dependent random variables via the martingale method. *The Annals of Probability*, 36(6):2126–2158, 2008.

Kontorovich, L. A., Nadler, B., and Weiss, R. On learning parametric-output HMMs. In *Proceedings of the 30th International Conference on Machine Learning (ICML'13)*, volume 28, pp. 702–710, 2013.

Krishnamurthy, V. *Partially Observed Markov Decision Processes*. Cambridge University Press, 2016.

Krishnamurthy, V. and Rydén, T. Consistent estimation of linear and non-linear autoregressive models with Markov regime. *Journal of Time Series Analysis*, 19(3):291–307, 1998.

Lakshminarayanan, B. and Raich, R. Non-negative matrix factorization for parameter estimation in hidden Markov models. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP'10)*, pp. 89–94, 2010.

Mamon, R. S. and Elliott, R. J. *Hidden Markov Models in Finance*, volume 1. Springer, 2007.

Mamon, R. S. and Elliott, R. J. *Hidden Markov Models in Finance: Further Developments and Applications*, volume 2. Springer, 2014.

Mattila, R. *Hidden Markov Models: Identification, Inverse Filtering and Applications*. PhD thesis, KTH Royal Institute of Technology, 2020. URL https://kth.diva-portal.org/smash/get/diva2:1428900/FULLTEXT01.pdf.

Mattila, R., Rojas, C. R., Krishnamurthy, V., and Wahlberg, B. Asymptotically efficient identification of known-sensor hidden Markov models. *IEEE Signal Processing Letters*, 24(12):1813–1817, 2017.

Mossel, E. and Roch, S. Learning nonsingular phylogenies and hidden Markov models. In *Proceedings of the 37th ACM symposium on Theory of computing*, pp. 366–375, 2005.

Nystrup, P., Madsen, H., and Lindström, E. Dynamic portfolio optimization across hidden market regimes. *Quantitative Finance*, 18(1):83–95, 2018.

Pearson, K. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

Rabiner, L. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Rockafellar, R. T. and Wets, R. J. B. *Variational Analysis*. Springer, 1998.

Rydén, T., Teräsvirta, T., and Åsbrink, S. Stylized facts of daily return series and the hidden Markov model. *Journal of Applied Econometrics*, 13(3):217–244, 1998.

Shaban, A., Farajtabar, M., Xie, B., Song, L., and Boots, B. Learning latent variable models by improving spectral solutions with exterior point methods. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence (UAI'15)*, pp. 792–801, 2015.

Subakan, C., Traa, J., Smaragdis, P., and Hsu, D. Method of moments learning for left-to-right hidden Markov models. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'15)*, pp. 1–5, 2015.

Tran, D., Kim, M., and Doshi-Velez, F. Spectral M-estimation with application to hidden Markov models. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS'16)*, 2016.

Turner, C. M., Startz, R., and Nelson, C. R. A Markov model of heteroscedasticity, risk and learning in the stock market. *Journal of Financial Economics*, 25(1):3–22, 1989.

Vanluyten, B. *Realization, Identification and Filtering for Hidden Markov Models Using Matrix Factorization Techniques*. PhD thesis, Katholieke Universiteit Leuven, 2008.

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

Vidyasagar, M. A realization theory for hidden Markov models: the partial realization problem. In *Proceedings of the 17th International Symposium on Mathematical Theory of Networks and Systems*, pp. 2145–2150, 2006.

Vidyasagar, M. The complete realization problem for hidden Markov models: a survey and some new results. *Mathematics of Control, Signals, and Systems*, 23(1-3):1–65, 2011.

Vidyasagar, M. *Hidden Markov Processes: Theory and Applications to Biology*. Princeton University Press, 2014.

Xia, L., Chen, C.-C., and Aggarwal, J. K. View invariant human action recognition using histograms of 3D joints. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'12) Workshops*, pp. 20–27, June 2012.

Yang, J., Xu, Y., and Chen, C. S. Human action learning via hidden Markov model. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 27 (1):34–44, 1997.

Yin, G. G. and Zhou, X. Y. Markowitz's mean-variance portfolio selection with regime switching: from discrete-time models to their continuous-time limits. *IEEE Transactions on Automatic Control*, 49(3):349–360, 2004.

Zhang, Y., Chen, X., Zhou, D., and Jordan, M. I. Spectral methods meet EM: a provably optimal algorithm for crowdsourcing. *Journal of Machine Learning Research*, 17(102):1–44, 2016.