Penalized Maximum Likelihood Estimation for Gaussian hidden Markov Models

Grigory Alexandrovich[1]

Fachbereich Mathematik und Informatik, Philipps-Universität Marburg, Germany.

Abstract

The likelihood function of a Gaussian hidden Markov model is unbounded, which is why the maximum likelihood estimator (MLE) is not consistent. A penalized MLE is introduced along with a rigorous consistency proof.

Keywords: Gaussian hidden Markov models, penalized maximum likelihood, consistency

1. Introduction

Hidden Markov models (HMMs) build a wide class of general-purpose models for describing weakly dependent stochastic processes. A HMM is a bivariate stochastic process $(X_i, Y_i)_{i\in\mathbb{Z}}$ where $(X_i)_{i\in\mathbb{Z}}$ is a Markov chain and $(Y_i)_{i\in\mathbb{Z}}$ is given $(X_i)_{i\in\mathbb{Z}}$ a sequence of independent random variables, such that given $(X_i)_{i\in\mathbb{Z}}$ Y_i depends only on X_i. The process $(Y_i)_{i\in\mathbb{Z}}$ can be observed, whereas the Markov chain $(X_i)_{i\in\mathbb{Z}}$ cannot (is hidden). We consider the case where the cardinality of the image space of the state variables (state space) is finite and transition probabilities $P(X_i = k \mid X_{i-1} = l)$ do not depend on i. In this case every irreducible Markov chain has a unique strictly positive stationary distribution, which is a left eigenvector of the t.p.m. with eigenvalue 1, see Zucchini and MacDonald [15] and the distribution of the Markov chain can be described by the transition probability matrix (t.p.m.).

HMMs have many application areas such as speech-, face-, handwriting recognition, biological sequence analysis, earthquakes prediction, finance etc..

Often the state-dependent distributions $Y_i \mid X_i = k$ are determined by a finite-dimensional euclidean parameter, like in the case of Gaussian HMMs. Then the law of the process $(Y_i, X_i)_{i\in\mathbb{Z}}$ is determined by the t.p.m. and the vector of state-dependent parameters.

[1]Address for correspondence: Grigory Alexandrovich, Philipps-Universität Marburg, Fachbereich Mathematik und Informatik, Hans-Meerwein-Strasse. D-35032 Marburg, Germany, Email: alexandrovich@mathematik.uni-marburg.de, Fon: + 49 6421 28 25462

An important task in the context of HMMs is estimation of the underlying parameter, which is often solved by maximizing the log-likelihood function. In the case of Gaussian HMMs however, a direct maximization has a theoretical drawback since the objective function is unbounded. Consider a two-state HMM and an estimator with $\hat{\mu}_1 = Y_1$, $|\hat{\Sigma}_1| = \varepsilon$, $\hat{\mu}_2 \in \mathbb{R}^d$ arbitrary, $\hat{\Sigma}_2 = I$, $\hat{\Phi}$ aperiodic and irreducible. Then the likelihood function tends to infinity as $\varepsilon \to 0$ and hence the MLE is not consistent.

The same problem exists in the i.i.d. mixture setting and was addressed by several authors. Two basic strategies for overcoming the unboundedness were studied in the literature: restricted optimization and penalization of the likelihood. In the first case a lower bound on the variances or their ratios is imposed, see e.g. Hathaway [7]. In the second case a term which penalizes small variances (Ciuperca et al. [6], Chen et al. [5], Chen and Tan [4]) or ratios of variances (Tanaka [13]) is added to the log-likelihood. The second approach has some advantages over the first one - there is no tuning constant to choose and the penalty function actually disappears with increasing sample size.

Although the unboundedness has no serious impact on the practice, since maximization algorithms, like EM, search for local maxima and converge only seldom against degenerate solutions, it should be desirable to eliminate this theoretical drawback by introducing a consistent estimator.

The state-dependent parameters of a HMM can be consistently estimated by maximizing the marginal mixture log-likelihood, or equivalently the HMM likelihood under independence assumption (IMLE) under some thechnical conditions, see Lindgren [10] and references therein. One necessary condition is $\lim_{\theta \to \partial\Theta} \varphi(y, \theta) = 0$ except on a zero-measure set, independent of the limit of θ. This condition is violated in our case as indicated above.

In the present work, a two-stage procedure is proposed for a consistent estimation of the parameters of a Gaussian HMM. In the first stage, the parameters of the marginal distribution of the observed process are estimated by maximizing a penalized mixture likelihood. Some ideas from Chen et al. [5] are used, where consistency of a penalized MLE for Gaussian mixtures is shown. The main difficulty during the generalization of that result is a more complicated large deviations behaviour of HMM samples.

In the second stage, the full HMM likelihood is maximized over a neighbourhood of the estimates from the stage 1. Since this neighbourhood is regular and contains the true parameter of the HMM for n large enough, the consistency result from Leroux [9] can be applied. The maximization in each step can be done with the EM algorithms for Gaussian mixture models and for HMMs respectively.

No simulation study is given in the current work, since trial runs showed unpenalized maximization to work very well, if the initial guests are not chosen consciously poor. The aim of present paper is a theoretical one - to prove the existence of a consistent penalized MLE for Gaussian hidden models.

2. The model and main results

In what follows θ_0 denotes the true parameter of the HMM, θ_0^{mix} the true parameter of the marginal mixture and F the true marginal distribution function. Y_1^n is a shorthand for $(Y_1, \ldots, Y_n)$. The matrix Φ_0 is assumed to be irreducible. Following notations will be used:

$\theta_k = (\mu_k, \sigma_k^2)$	parameters of the k'th state
$d_c(x,y) = \sum_{s=1}^r \lvert\arctan(x_s) - \arctan(y_s)\rvert$	metric on $\mathbb{R}^r$
$\mathcal{T} = \{\Phi \in \mathbb{R}^{K\times K},\ \Phi_{i,\cdot} \in \mathcal{S}^{K-1},\ 1 \le i \le K\}$	transition probability matrices
$\mathcal{S}^{K-1} = \{(\alpha_1, \ldots, \alpha_K) \in \mathbb{R}^K,\ \sum_{i=1}^K \alpha_i = 1,\ \alpha_i \ge 0\}$	distributions on $\{1, \ldots, K\}$

Table 1: **Notations**

Definition 1. Let $(X_i, Y_i)_{i\in\mathbb{Z}}$ be a stochastic process, where $(Y_i)_{i\in\mathbb{Z}}$ are independent given $(X_i)_{i\in\mathbb{Z}}$, which is a homogeneous first order Markov chain. Furthermore

$$X_i \in \{1, \ldots, K\}, \tag{1}$$

$$Y_i \mid (X_j)_{j\in\mathbb{Z}} = Y_i \mid X_i, \tag{2}$$

$$Y_i \mid X_i = k = N(\mu_{0k}, \sigma_{0k}^2). \tag{3}$$

The process $(X_i, Y_i)_{i\in\mathbb{Z}}$ is called a *Gaussian hidden Markov model (HMM)*. In the special case where $(X_i)_{i\in\mathbb{Z}}$ are independent, the process $(X_i, Y_i)_{i\in\mathbb{Z}}$ is called a *Gaussian mixture model*.

The set of possible HMM parameters will be denoted by

$$\Theta^{full} = \{(\Phi, \mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2) \mid \Phi \in \mathcal{T}, \quad \mu_j \in \mathbb{R}, \quad \sigma_j^2 \in (0,\infty),\ j = 1 \ldots K\}.$$

The set parameters of a Gaussian mixture for the first stage of the algorithm will be denoted by

$$\Theta^{mix} = \{(\pi, \mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2) \mid \pi \in \mathcal{S}^{K-1}, \quad \mu_j \in \mathbb{R}, \quad \sigma_j^2 \in (0,\infty),\ j = 1 \ldots K\}.$$

$\mu_k(\theta), \sigma_k(\theta)$ denote the coordinate projections on the state-dependent parameters for $1 \le k \le K$. This dependence on θ will be suppressed whenever no confusion can occur. The compactification of both sets is done by adding limits of Cauchy sequences with respect to d_c as in Kiefer and Wolfowitz [8], and is denoted by $\bar{\Theta}^{full}$ and $\bar{\Theta}^{mix}$. Let $\alpha = (\alpha_1, \ldots, \alpha_K)$ be an initial state distribution, $\varphi(y, \mu, \sigma^2)$ the density of the normal distribution with mean μ and variance σ^2:

$$\varphi(y, \mu, \sigma^2) = (2\pi)^{-\frac{1}{2}} \sigma^{-1} \exp(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}).$$

For $\theta \in \Theta^{full}$ the function

$$L_n^{full}(\theta; Y_1, \ldots, Y_n) = \sum_{x_1=1}^K \ldots \sum_{x_n=1}^K \alpha_{x_1} \varphi(Y_1, \theta_{x_1}) \prod_{i=2}^n \phi_{x_{i-1},\, x_i} \varphi(Y_i, \theta_{x_i}) \tag{4}$$

is called the *likelihood function* for $Y_1, \ldots, Y_n$. For $\theta \in \Theta^{mix}$ the function

$$L_n^{mix}(\theta; Y_1, \ldots, Y_n) = \prod_{i=1}^{n} \sum_{j=1}^{K} \pi_j \varphi(Y_i, \theta_j) = \prod_{i=1}^{n} f(Y_i, \theta), \tag{5}$$

where $f(y, \theta) = \sum_{j=1}^{K} \pi_j \varphi(y, \theta_j)$, is called the *marginal-mixture-likelihood function* for $Y_1, \ldots, Y_n$.

Now penalty functions for the first stage of the procedure are defined similar to Chen et al. [5] .

Definition 2. A function $p_n : \Theta^{mix} \to \mathbb{R}$ with following properties

1. $p_n(\theta) = \sum_{k=1}^{K} \tilde{p}_n({\sigma^2}_k)$.
2. At any fixed θ, with $\sigma_k^2 > 0$, $k = 1, \ldots, K$, we have $p_n(\theta) = o(n)$, and $\sup_\theta \max\{0, p_n(\theta)\} = o(n)$.
3. p_n is differentiable and as $n \to \infty$, $p_n'(\theta) = o(n^{\frac{1}{2}})$ at any fixed σ_k^2, with $\sigma_k^2 > 0$, $k = 1, \ldots, K$.
4. For large enough n, $\tilde{p}_n(\sigma^2) \leq \sqrt{n}(\log n)^2 \log \sigma^2$, when $\sigma^2 < cn^{-2}$ for some $c > 0$.
5. For every $\varepsilon > 0$ holds $\sup_{\{\theta \mid \sigma^2(\theta) > \varepsilon\}} |\tilde{p}_n(\theta)| = o(n)$.

is called a *penalty function.*

These requirements are very similar to those from Chen et al. [5] and Chen and Tan [4]. The last condition was missing in the cited works, although it was implicitly assumed. The main difference lies in the fourth condition, which is linked to Lemma 11 below and is imposed to control the damaging effect of observations near degenerate components. Lemma 11 generalizes Lemma 1 from Chen and Tan [4] and is the most challenging part of the proof. The original proof relies on a Bernstein inequality for i.i.d. observations from Serfling [12], which is however not applicable for dependent observations. A more recent result from Merlevède et al. [11] was used instead.

The requirements are not very restrictive, for example the following function $\tilde{p}_n(\sigma^2) = -n^{-1}\mathrm{tr}(\sigma^{-2})$ fulfils them.

Definition 3. Let

$$\hat{\theta}_n^{pIMLE} = \operatorname*{argmax}_{\theta \in \Theta^{mix}} \log L_n^{mix}(\theta; Y_1, \ldots, Y_n) + p_n(\theta) \tag{6}$$

For ease of notation let $\nu(\theta) = (\mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2)(\theta)$ for $\theta \in \Theta^{mix} \cup \Theta^{full}$ be the coordinate projection on the state-dependent parameters. For a mixture parameter $\theta' \in \Theta^{mix}$ and a $\delta > 0$ let

$$\Theta^{full}(\theta', \delta) = \{\theta \in \Theta^{full} \mid ||\nu(\theta), \nu(\theta')||_2 \leq \delta\}.$$

The *penalized maximum likelihood estimator (pMLE)* of θ is defined by

$$\hat{\theta}_n^{pMLE} = \underset{\theta \in \Theta^{full}(\hat{\theta}_n^{pIMLE},\, \delta)}{\operatorname{argmax}} \log L_n^{full}(\theta; Y_1, \ldots, Y_n) + p_n(\theta) \tag{7}$$

for a penalty function p_n.

Now we are ready to establish the main result of this paper, namely the consistency of the penalized maximum likelihood estimator for Gaussian hidden Markov models. The consistency is formulated in terms of the convergence in quotient topology (see Leroux [9]]).

Definition 4. For a parameter $\theta \in \Theta^{full}$, the *equivalence class* $\tilde{\theta}$ is defined by

$$\tilde{\theta} = \{\theta' \in \Theta^{full} \mid (\theta'_{X_i})_{i\in\mathbb{Z}} \overset{d}{=} (\theta_{X_i})_{i\in\mathbb{Z}}\},$$

that is the set of the parameters which induce the same law for the process $(\theta_{X_i})_{i\in\mathbb{Z}}$ as θ.

Convergence in quotient topology means that every open subset of the parameter space, that contains the equivalence class of θ_0, must for large n, contain the equivalence class of $\hat{\theta}_{pMLE}$.

Theorem 5. *$\hat{\theta}_{pMLE}$ converges to θ_0 in quotient topology with probability one for every positive $\delta > 0$ in definition of $\hat{\theta}_n^{pMLE}$.*

Next theorem states the asymptotic equivalence between the penalized MLE and the maximizer of the full HMM likelihood over a restricted parameter space, where the variance parameters are bounded away from the zero. This allows us to transfer some results from the restricted case to the penalized one.

Theorem 6 (Asymptotic equivalence)**.** *Let $\hat{\theta}_R = \operatorname{argmax}_{\Theta^{full}} \log L_n^{full}(\theta; Y_1^n)$, s.t. ${\sigma^2}_k \geq \varepsilon$, for $k \in \{1, \ldots, K\}$ for some small ε, such that ${\sigma^2}_{0k} > \varepsilon$, for $k \in \{1, \ldots, K\}$, then*

$$\sqrt{n}(\hat{\theta}_n^{pMLE} - \hat{\theta}_R) \xrightarrow{P} 0. \tag{8}$$

Proof. We expand $\nabla \log L_n^{full}(\hat{\theta}_n^{pMLE}) = \nabla \log L_n^{full}(\hat{\theta}_R) + \nabla^2 \log L_n^{full}(\tilde{\theta})(\hat{\theta}_n^{pMLE} - \hat{\theta}_R)$, where $\tilde{\theta}$ lies on the line segment between $\hat{\theta}_R$ and $\hat{\theta}_n^{pMLE}$. Since the true parameter lies in the interior of the feasible set, we have $\nabla \log L_n^{full}(\hat{\theta}_R) = 0$. So we obtain $\nabla \log L_n^{full}(\hat{\theta}_n^{pMLE}) = \nabla^2 \log L_n^{full}(\tilde{\theta})(\hat{\theta}_n^{pMLE} - \hat{\theta}_R)$. Furthermore, since $\hat{\theta}_n^{pMLE}$ and $\hat{\theta}_R$ are both consistent [2], we have $\tilde{\theta} \to \theta_0$. Hence by the consistency of $\tilde{\theta}$ and Lemma 2 from Bickel et al. [2] it holds: $\frac{1}{n}\nabla^2 \log L_n^{mix}(\tilde{\theta}) \xrightarrow{P} -I_0$, where I_0 is a non-random matrix (the Fisher-Information) and by the continuous mapping theorem $n\nabla^2 \log L_n^{full}(\tilde{\theta})^{-1} \xrightarrow{P} -{I_0}^{-1}$. Combining these facts yields

$$\sqrt{n}(\hat{\theta}_n^{pMLE} - \hat{\theta}_R) = \overbrace{n\nabla^2 \log L_n^{full}(\tilde{\theta})^{-1}}^{\to -{I_0}^{-1}} \frac{1}{\sqrt{n}} \nabla \log L_n^{full}(\hat{\theta}_n^{pMLE}).$$

[2] $\hat{\theta}_R$ satisfies conditions stated by Leroux

Finally $\frac{1}{\sqrt{n}}\nabla \log L_n^{full}(\hat{\theta}_n^{pMLE}) \xrightarrow{P} 0$, since $\nabla \log L_n^{full}(\hat{\theta}_n^{pMLE}) = -\nabla p_n(\hat{\theta}_n^{pMLE})$ and $p_n(\hat{\theta}_n^{pMLE}) = o(\sqrt{n})$ a.s. by construction. □

The following result establishes the asymptotic normality of the penalized MLE.

Theorem 7 (Asymptotic normality)**.**

$$\sqrt{n}(\hat{\theta}_n^{pMLE} - \theta_0) \xrightarrow{d} N(0, I_0^{-1}), \tag{9}$$

where $-I_0 = \lim_{n\to\infty} \frac{1}{n}\nabla^2 \log L_n^{full}(\theta_0, Y_1, \ldots, Y_n)$.

Proof. This statement follows from the asymptotic equivalence between $\hat{\theta}_n^{pMLE}$ and $\hat{\theta}_R$ and the fact, that $\hat{\theta}_R$ satisfies the assumptions of Theorem 1 in Bickel et al. [2]. The assumptions are:

(A1) The transition probability matrix is ergodic.

(A2) The elements of Φ and the stationary distribution are twice differentiable w.r.t θ.

(A3) Let $\theta = (\theta_1, \ldots, \theta_r)$. There exists a $\delta > 0$, such that (i) for all $1 \leq i \leq r$ and all $k \in \{1, \ldots, K\}$

$$\mathbb{E}_0\left[\sup_{|\theta-\theta_0|<\delta} |\frac{\partial}{\partial\theta_i} \log \varphi(Y_1; \mu_k, {\sigma^2}_k)|^2\right] < \infty,$$

(ii) for all $1 \leq i, j \leq r$ and all $k \in \{1, \ldots, K\}$

$$\mathbb{E}_0\left[\sup_{|\theta-\theta_0|<\delta} |\frac{\partial^2}{\partial\theta_i\partial\theta_j} \log \varphi(Y_1; \mu_k, {\sigma^2}_k)|\right] < \infty,$$

(iii) for all $j = 1, 2$, all $1 \leq i_l \leq r$, $l = 1, \ldots, j$, and all $k \in \{1, \ldots, K\}$

$$\int \sup_{|\theta-\theta_0|<\delta} |\frac{\partial^j}{\partial\theta_{i_1} \ldots \partial\theta_{i_j}} \varphi(Y_1; \mu_k, {\sigma^2}_k)| dy < \infty,$$

(A4) There exists a $\delta > 0$ such that with

$$\rho_0(y) = \sup_{|\theta-\theta_0|<\delta} \max_{1\leq k_1,k_2\leq K} \frac{\varphi(y|\mu_{k_1}, {\sigma^2}_{k_1})}{\varphi(y|\mu_{k_2}, {\sigma^2}_{k_2})},$$

$\mathbb{P}(\rho_0(Y_1) = \infty | X_1 = k) < 1$ for all $k \in \{1, \ldots, K\}$.

(A5) θ_0 is an interior point of Θ

(A6) The maximum likelihood estimator is strongly consistent.

(A1) is part of our assumptions. The elements of Φ are part of the parameter vector and the initial distribution doesn't depend on θ, so (A2) is satisfied too. The conditions (A3) and (A4) are satisfied since φ is the normal density and ${\sigma^2}_k > 0$ for $k \in \{1, \ldots, K\}$. Furthermore (A5) follows also from ${\sigma^2}_k > 0$ for $k \in \{1, \ldots, K\}$. Finally (A6) holds, since $\hat{\theta}_R$ satisfies the regularity conditions from Leroux [9]. □

3. Proofs

3.1. Preliminary analytic results

Several technical statements follow.

Proposition 8. *Let* $(\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$ *set*

$$\tilde{A} = \tilde{A}(\mu, \sigma^2) = \{y \in \mathbb{R} \mid \frac{(y-\mu)^2}{\sigma^2} \leq (\log \sigma^2)^2\}. \tag{10}$$

Then

$$\varphi(y, \mu, \sigma^2) \leq \begin{cases} \sigma^{-1} & y \in \tilde{A} \\ \exp -\frac{(y-\mu)^2}{4\sigma^2} & \textit{otherwise.} \end{cases} \tag{11}$$

Proof. First we note $\varphi(y, \mu, \sigma^2) \leq \sigma^{-1}$ for every $y \in \mathbb{R}$, so the first inequality is obvious. For $y \notin \tilde{A}$ we have that $\frac{(y-\mu)^2}{\sigma^2} > (\log \sigma^2)^2$. Therefore

$$\begin{aligned} \varphi(y, \mu, \sigma^2) &\leq \frac{1}{\sqrt{\sigma^2}} \exp\big(-\frac{(y-\mu)^2}{\sigma^2}/4\big) \exp\big(-\frac{(y-\mu)^2}{\sigma^2}/4\big) \\ &< \frac{1}{\sqrt{\sigma^2}} \exp\big(-\frac{1}{4}(\log \sigma^2)^2\big) \exp\big(-\frac{(y-\mu)^2}{\sigma^2}/4\big) \\ &= \exp\Big(-\frac{1}{2}\big(\log \sigma^2 + (\log \sigma^2)^2/2\big)\Big) \exp(-\frac{(y-\mu)^2}{\sigma^2}/4) \\ &\leq \exp\big(-\frac{(y-\mu)^2}{\sigma^2}/4\big). \end{aligned} \tag{12}$$

□

Proposition 9. *Let* $\mu_1, \mu_2 \in \mathbb{R}$ *and* $\sigma_1^2, \sigma_2^2 \in (0, \infty)$ *with* $\sigma_1^2 \leq \sigma_2^2 \leq \varepsilon$*, for some* $0 < \varepsilon < e^{-1/4}$*. Suppose that* $y \in \mathbb{R}$ *is such that*

$$\frac{(y-\mu_1)^2}{\sigma_1^2} > (\log \sigma_1^2)^2, \qquad \frac{(y-\mu_2)^2}{\sigma_2^2} \leq (\log \sigma_2^2)^2.$$

then

$$\varphi(y, \mu_1, \sigma_1^2) < \varphi(y, \mu_2, \sigma_2^2).$$

Proof. From the properties of y we have

$$\begin{aligned} \frac{1}{\sqrt{\sigma_1^2}} \exp\{-\frac{1}{2}\frac{(y-\mu_1)^2}{\sigma_1^2}\} &< \frac{1}{\sqrt{\sigma_1^2}} \exp\{-\frac{1}{2}(\log \sigma_1^2)^2\}, \\ \frac{1}{\sqrt{\sigma_2^2}} \exp\{-\frac{1}{2}\frac{(y-\mu_2)^2}{\sigma_2^2}\} &\geq \frac{1}{\sqrt{\sigma_2^2}} \exp\{-\frac{1}{2}(\log \sigma_2^2)^2\}. \end{aligned}$$

Thus, it sufficies to shows that the function

$$f(z) = \frac{1}{z}\exp\{-\frac{1}{2}\log(z^2)^2\}, \qquad z > 0,$$

is increasing near zero. The first derivative is given by

$$f'(z) = -\frac{1}{z^2}\exp\{-\frac{1}{2}\big(\log(z^2)\big)^2\}\left[1 + 4\log(z)\right],$$

which is > 0 for $z < e^{-1/4}$. □

Lemma 10. *Let Y be a random variable in $\mathbb{R}$ with a bounded density w.r.t. the Lebesgue measure. Given $\delta > 0$ there is a τ_0, such that for any $\mu \in \mathbb{R}$ and $\sigma^2 \in (0, \infty)$ with $\sigma^2 < \tau_0$, we have*

$$\mathbb{P}\big(Y \in \tilde{A}(\mu, \sigma^2)\big) < \delta,$$

where $\tilde{A}(\mu, \sigma^2)$ is defined in (10).

Proof. The Lebesgue length of $\tilde{A}(\mu, \sigma^2)$ is given by $2\sigma\,|\log\sigma^2|$, which tends to zero as $\sigma^2 \to 0$. The statement follows since Y has a bounded Lebesgue density. □

3.2. Bounds on the number of points near degenerate components

The following lemma is a generalization of Lemma 1 from Chen et al. [5] for Gaussian hidden Markov processes. It bounds the number of observations of such a process which are located in neighbourhoods of degenarate components. These observations have a high contribution to the likelihood and will be ruled out by penalty function.

Lemma 11. *Let $(Y_n)_{n\geq 1}$ be a stationary Gaussian hidden Markov process with K states and parameter vector $(\Phi, \mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2)$. Let F_n be the empirical distribution function of $Y_1, \ldots, Y_n$, and M denote an upper bound for the marginal mixture density. Then almost sure there exists $N \in \mathbb{N}$, such that*

$$\sup_y [F_n(y+\tau) - F_n(y)] \leq \frac{(\log n)^2}{\sqrt{n}} + 2M\tau + \frac{1}{n}$$

for all $n \geq N$ and $\tau \in [0, e^{-1}]$.

Proof of Lemma 11. For $\tau = 0$ the statement is trivial. Let $\tau \in (0, e^{-1}]$ and $1 \leq k, i \leq n$ we define $\eta_k = F^{-1}(\frac{k}{n})$. We have

$$\begin{aligned}
&\sup_y [F_n(y+\tau) - F_n(y)] \\
&\quad \leq \max_k [F_n(\eta_k + \tau)] - F_n(\eta_{k-1}) \\
&\quad \leq \max_k [\{F_n(\eta_k + \tau) - F_n(\eta_{k-1})\} - \{F(\eta_k + \tau) - F(\eta_{k-1})\}] \\
&\quad + \max_k \{F(\eta_k + \tau) - F(\eta_{k-1})\}.
\end{aligned} \tag{13}$$

To bound the second term in (13), by the Mean Value Theorem we obtain

$$\begin{aligned} F(\eta_k+\tau)-F(\eta_{k-1}) &= F(\eta_k+\tau)-F(\eta_k)+n^{-1} \\ &\le M\tau+n^{-1} =: \delta_n(\tau). \end{aligned} \tag{14}$$

It remains to find an appropriate bound for

$$\Delta_{n,k}^{\tau} = |\{F_n(\eta_k+\tau)-F_n(\eta_{k-1})\}-\{F(\eta_k+\tau)-F(\eta_{k-1})\}|.$$

Write

$$\begin{aligned} n\Delta_{n,k}^{\tau} &= |\sum_{i=1}^{n} \mathbf{1}_{\{Y_i\le\eta_k+\tau\}} - \mathbf{1}_{\{Y_i\le\eta_{k-1}\}} - \{F(\eta_k+\tau)-F(\eta_{k-1})\}| \\ &= |\sum_{i=1}^{n} Z_{i,k}^{\tau} - \{F(\eta_k+\tau)-F(\eta_{k-1})\}|. \end{aligned}$$

where $Z_{i,k}^{\tau} = \mathbf{1}_{\{Y_i\le\eta_k+\tau\}} - \mathbf{1}_{\{Y_i\le\eta_{k-1}\}}$.

From the Bernstein inequality in Lemmas 13 and 15 in the Appendix, there exist positive constants $\gamma, C_1, C_2, C_3, C_4, V$ and $n_0 \in \mathbb{N}$ depending only on the true parameter vector $(\Phi_0, \mu_{01}, \ldots, \mu_{0K}, \sigma_{01}^2, \ldots, \sigma_{0K}^2)$ of the HMM such that

$$\mathbb{P}(|\Delta_{n,j}^{\tau}| \ge x) \le n\exp\left(-\frac{n^{\gamma}x^{\gamma}}{C_1}\right) + \exp\left(-\frac{n^2x^2}{C_2(1+nV)}\right) + \exp\left(-\frac{n^2x^2}{C_3 n}\exp\frac{(nx)^{\gamma(1-\gamma)}}{C_4(\log\{xn\})^{\gamma}}\right) \tag{15}$$

for every $x \in \mathbb{R}$, $j = 1, \ldots, n$ and $\tau \in (0, e^{-1}]$. Setting $x = \frac{(\log n)^2}{2\sqrt{n}}$ gives

$$\begin{aligned} \mathbb{P}(|\Delta_{n,k}^{\tau}| \ge \frac{(\log n)^2}{2\sqrt{n}}) \le\ & n\,\exp\left(-\frac{n^{\frac{\gamma}{2}}(\log n)^{2\gamma}}{2^{\gamma}C_1}\right) \\ &+\ \exp\left(-\frac{n(\log n)^4}{4C_2(1+nV)}\right) \\ &+\ \exp\left(-\frac{(\log n)^4}{4C_3}\exp\frac{\{n^{\frac{1}{2}}(\log n)^2/2\}^{\gamma(1-\gamma)}}{C_4(\log\{(\log n)^2 n^{\frac{1}{2}}/2\})^{\gamma}}\right). \end{aligned}$$

Therefore we get that for every $n \ge n_0$, $j = 1, \ldots, n$ and $\tau \in (0, e^{-1}]$,

$$\mathbb{P}\Big(|\Delta_{n,k}^{\tau}| \ge \frac{(\log n)^2}{2\sqrt{n}}\Big) \le cn^{-3} \tag{16}$$

for some constant c. Let $r_n = \frac{(\log n)^2}{2M\sqrt{n}}$. We have that

$$\begin{aligned} &\mathbb{P}\big(\max_{k=1\ldots n}|\Delta_{n,k}^{r_n}| \ge \frac{(\log n)^2}{2\sqrt{n}}\big) \le \mathbb{P}\Big(\cup_{k=1}^{n}\{|\Delta_{n,k}^{r_n}| \ge \frac{(\log n)^2}{2\sqrt{n}}\}\Big) \\ \le &\sum_{k=1}^{n}\mathbb{P}(|\Delta_{n,k}^{r_n}| \ge \frac{(\log n)^2}{2\sqrt{n}}) < cn^{-2}. \end{aligned} \tag{17}$$

By Borel-Cantelli, a.s. there is an N_1, such that

$$\max_{k=1\dots n} |\Delta_{n,k}^{r_n}| \le \frac{(\log n)^2}{2\sqrt{n}}, \qquad n \ge N_1.$$

Therefore, by (13) and (14) and monotonicity,

$$\sup_{\tau\in(0,r_n]} \sup_y |F_n(y+\tau) - F_n(y)| \le \sup_y |F_n(y+r_n) - F_n(y)|$$
$$\le \frac{(\log n)^2}{2\sqrt{n}} + \delta_n(r_n) \le \frac{(\log n)^2}{\sqrt{n}} + 1/n, \qquad n \ge N_1,$$

which shows the estimate for all $\tau \in (0, r_n]$.
Next consider $\tau \in [r_n, e^{-1}]$. Now we define a finite grid over $[r_n, e^{-1}]$ by $\tau_0 = r_n$ and $\tau_{k+1} = 2\tau_k$, where $k \le \lfloor \log_2 \frac{2Me^{-1}\sqrt{n}}{(\log n)^2} \rfloor =: k_n < \log n$ for n large enough. If $\tau_{k_n} < e^{-1}$, we add the point $\tau_{k_n+1} = e^{-1}$ to the grid, hence we assume w.l.o.g. $\tau_{k_n} = e^{-1}$. Let

$$D_n = \bigcup_{k=1}^{k_n} \Big\{ \sup_y F_n(y+\tau_k) - F_n(y) \ge \frac{(\log n)^2}{2\sqrt{n}} + \delta_n(\tau_k) \Big\}.$$

From (13), (14) and (16) we obtain

$$\sum_{n=1}^{\infty} \mathbb{P}(D_n) \le \sum_{n=1}^{\infty}\sum_{j=1}^{k_n} \mathbb{P}\Big(\big\{ \sup_y F_n(y+\tau_j) - F_n(y) \ge \frac{(\log n)^2}{2\sqrt{n}} + M\tau_j + \frac{1}{n} \big\} \Big)$$
$$\le \sum_{n=1}^{\infty}\sum_{k=1}^{k_n} \mathbb{P}\Big(\max_{j=1\dots n} |\Delta_{n,j}^{\tau_k}| \ge \frac{(\log n)^2}{2\sqrt{n}} \Big)$$
$$\le \sum_{n=1}^{\infty} c \log n\, n^{-2} < \infty.$$

where we estimate the maximal probability as in (17). We conclude by Borel-Cantelli $\mathbb{P}(D_n \text{ i.o.}) = 0$. Since for every $\tau \in [r_n, e^{-1}]$ there exist two grid points such that $\tau \in [\tau_j, \tau_{j+1}]$, a.s. there is an N_2 such that

$$\sup_y F_n(y+\tau) - F_n(y) \le \sup_y F_n(y+\tau_{j+1}) - F_n(y) \le \frac{(\log n)^2}{2\sqrt{n}} + 2M\tau + \frac{1}{n}$$

for all $n \ge N_2$ and $\tau \in [\tau_j, \tau_{j+1}]$, where we used $\tau_{j+1} \le 2\tau$. □

Remark: The rate in the lemma above can be improved from $\sqrt{n}(\log n)^2$ to $\sqrt{n}(\log n)^{1+q}$ For any $q > 0$. But the higher one is still sufficient for the proof.

3.3. Proof of Theorem 5 in case $K = 2$

Proof. It is sufficient to show the consistency of $\hat{\theta}_n^{pIMLE}$ for the state dependent parameters. Then the consistency of $\hat{\theta}_n^{pMLE}$ follows from the result in Leroux [9], since the

maximization in stage 2 is carried out over a regular set, which contains the true parameter.

We show the consistency of $\hat{\theta}_n^{pIMLE}$ for the case $K = 2$ since the general K follows analogously. We follow Chen and Tan [4] in the proof structure and divide the parameter space in a finite number of subsets, one of which is regular. Step by step we show by applying Lemma 11 and classical techniques $\hat{\theta}_n^{pIMLE}$ to lie outside any of the irregular subsets.

Let $K = 2$ and assume w.l.o.g. $\sigma_1^2 \leq \sigma_2^2$. We divide the parameter space Θ^{mix} into three disjoint subsets.

$$\begin{aligned}
\Gamma_1 &= \{\, \theta \in \Theta^{mix} \,|\, \sigma_1^2 \leq \sigma_2^2 \leq \varepsilon_0 \,\}, \\
\Gamma_2 &= \{\, \theta \in \Theta^{mix} \,|\, \sigma_1^2 \leq \tau_0, \sigma_2^2 \geq \varepsilon_0 \,\}, \\
\Gamma_3 &= \Theta^{mix} \setminus \Gamma_1 \cup \Gamma_2.
\end{aligned}$$

For each $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \in \mathbb{R} \times \mathbb{R} \times (0, \infty) \times (0, \infty)$ we define the intervals subsets as in (10),

$$\tilde{A}_1 = \tilde{A}(\mu_1, \sigma_1^2), \qquad \tilde{A}_2 = \tilde{A}(\mu_2, \sigma_2^2).$$

Set

$$A_1 = \{i \mid Y_i \in \tilde{A}_1\}, \qquad A_2 = \{i \mid Y_i \in \tilde{A}_2\}, \tag{18}$$

and $M = \sigma_1^{-1}$. Further set

$$H_0 = \lim_n \frac{1}{n} \log L_n^{mix}(\theta_0^{mix}; Y_1^n), \tag{19}$$

which exists and is finite, see Lindgren [10]. The scalars ε_0 and τ_0 are chosen to satisfy

1. $2\sqrt{2}\varepsilon_0^{\frac{1}{2}} |\log \varepsilon_0| < e^{-1}, \qquad |\varepsilon_0^{\frac{1}{2}} \log \varepsilon_0 \log \varepsilon_0^{-\frac{1}{2}}| \leq 1/2.$
2. $0 < \tau_0 \leq \varepsilon_0,$
3. $-\log \varepsilon_0 - (\log \varepsilon_0)^2 \leq 4(H_0 - 2),$
4. $\varepsilon_0 < \sigma_{01}^2,$
5. $\mathbb{P}(Y_1 \in \tilde{A}_1^c \cap \tilde{A}_2^c) \geq \frac{1}{2}$ for $\theta \in \Gamma_1$.

The first part of condition 1 is necessary for applying Lemma 11, the second part is possible since $\varepsilon^{\frac{1}{2}} \log \varepsilon \log \varepsilon^{-\frac{1}{2}} \to 0$ as $\varepsilon \to 0$. The second condition ensures the order of the components. The third condition bounds the effect of observations, which will be ruled out by the log-likelihood at the true parameter. The existence of ε_0 and τ_0 which satisfy the first four conditions is obvious. The fifth condition can be achieved by applying Lemma 10.

Step 1. We shall show that

$$\sup_{\theta\in\Gamma_1}\big(\log L_n^{mix}(\theta;Y_1^n)+p_n(\theta)\big)-\log L_n^{mix}(\theta_0^{mix};Y_1^n)-p_n(\theta_0^{mix})\to-\infty. \tag{20}$$

To this end, we shall show that a.s. there is an N, such that for $n\geq N$ we have that

$$\log L_n^{mix}(\theta;Y_1^n)+p_n(\theta)\leq n(H_0-1),\qquad n\geq N. \tag{21}$$

the conclusion then follows together with (19). To show (21), for a set $S\subset\{1,\dots,n\}$ with $n(S)$ elements let

$$L_n^{mix}(\theta;S)=\prod_{j\in S}f(Y_j,\theta),$$

and write

$$\log L_n^{mix}(\theta;Y_1^n)+p_n(\theta)=\Big(\log L_n^{mix}(\theta;A_1)+\tilde{p}_n(\sigma_1^2)\Big)+\Big(\log L_n^{mix}(\theta;A_1^c\cap A_2)+\tilde{p}_n(\sigma_2^2)\Big)+\log L_n^{mix}(\theta;A_1^c\cap A_2^c). \tag{22}$$

We shall bound each term on the right seperately in order to achieve (21). Since $\sigma_1^2\leq\sigma_2^2$ we have that $f(y,\theta)\leq\sigma_1^{-1}$ for any y, and hence that $\log L_n^{mix}(\theta;A_1)\leq n(A_1)\log{\sigma_1}^{-1}$. First we assert for $\varepsilon_0\geq\sigma_1^2>n^{-2}$ with the help of Lemma 11

$$\begin{aligned}\log L_n^{mix}(\theta;A_1)&\leq n(A_1)\log\sigma_1^2\leq\big(\sqrt{n}(\log n)^2-nM\sigma_1\log\sigma_1^2+1\big)\log{\sigma_1}^{-1}\\&=\sqrt{n}(\log n)^2\log{\sigma_1}^{-1}-nM\sigma_1\log\sigma_1+\log{\sigme_1}^{-1}=:h_n(\sigma_1^2)\end{aligned}$$

and

$$\sup_{\sigma_1^2\in[n^{-2},\,\varepsilon_0]}h_n(\sigma_1^2)\leq\sqrt{n}(\log n)^2\log n-nM\varepsilon_0^{1/2}\log\varepsilon_0+\log n<n/4. \tag{23}$$

The right hand side of the last display is less than a fraction of n for n large and ε small enough. Now suppose $\sigma_1^2\leq n^{-2}$, then from Property 4 of the penalty $\tilde{p}_n$ and Lemma 11, a.s. for large enough n, we obtain the bound

$$\begin{aligned}&\log L_n^{mix}(\theta;A_1)+\tilde{p}_n(\sigma_1^2)\leq n(A_1)\log{\sigma_1}^{-1}+\sqrt{n}(\log n)^2\log\sigma_1^2\\&\leq\big(\sqrt{n}(\log n)^2-nM\sigma_1\log\sigma_1^2+1\big)\log{\sigma_1}^{-1}+\sqrt{n}(\log n)^2\log\sigma_1^2\\&=\sqrt{n}(\log n)^2\log\sigma_1+\log{\sigma_1}^{-1}-nM\sigma_1\log\sigma_1^2\log{\sigma_1}^{-1}\\&\leq n/4,\end{aligned} \tag{24}$$

since $\sqrt{n}(\log n)^2\log\sigma_1+{\sigma_1}^{-1}$ is negative, $\sigma_1^2\leq\varepsilon_0$ and ε_0 is chosen to satisfy the second part of condition 1 above. Similarly, for $y\in A_1^c\cap A_2$, from Lemma 9 we have that $f(y,\theta)\leq\log{\sigma_2}^{-1}$, and hence that $L_n^{mix}(\theta;A_1^c\cap A_2)\leq n(A_2)\log{\sigma_2}^{-1}$, and similarly as in (24) we obtain a.s. for large enough n that

$$\begin{aligned}\sup_{\sigma_1^2\in[n^{-2},\,\varepsilon_0]}\log L_n^{mix}(\theta;A_1^c\cap A_2)+\tilde{p}_n(\sigma_2^2)\leq n/4\\ \sup_{\sigma_1^2\in(0,\,n^{-2})}\log L_n^{mix}(\theta;A_1^c\cap A_2)\leq n/4\end{aligned} \tag{25}$$

Further,

$$\begin{aligned}\log L_n^{mix}(\theta; A_1^c \cap A_2^c) &\leq \sum_{j \in A_1^c \cap A_2^c} \log \left[\exp(\log {\sigma_2}^{-1} - \frac{1}{2}(\log \sigma_2^2)^2) \right] \\ &\leq \sum_{j \in A_1^c \cap A_2^c} -\frac{1}{2}\log \varepsilon_0 - \frac{1}{2}(\log \varepsilon_0)^2 \\ &\leq n(H_0 - 2).\end{aligned} \tag{26}$$

Here, for the first inequality we recall that the function $\frac{1}{z}\exp\{-\frac{1}{2}\log(z^2)^2\}$ is monotone increasing near zero, as shown in the proof of Lemma 9. Let us argue for the last inequality in (26). In case $H_0 < 2$, we assumed that $-\log \varepsilon_0 - (\log \varepsilon_0)^2 \leq 4(H_0 - 2)$, so that in this case we obtain

$$\begin{aligned}&\sum_{j \in A_1^c \cap A_2^c} -\frac{1}{2}\log \varepsilon_0 - \frac{1}{2}(\log \varepsilon_0)^2 \\ \leq &n(A_1^c \cap A_2^c)\, 2\, (H_0 - 2) \leq n(A_1^c \cap A_2^c)\, (H_0 - 2).\end{aligned}$$

In case $H_0 \geq 2$ we use the trivial bound $-\log \varepsilon_0 - (\log \varepsilon_0)^2 \leq 2(H_0 - 2)$, and get

$$\sum_{j \in A_1^c \cap A_2^c} -\frac{1}{2}\log \varepsilon_0 - \frac{1}{2}(\log \varepsilon_0)^2 \leq n(A_1^c \cap A_2^c)\, (H_0 - 2)$$

as well. By condition 5 and the ergodic theorem, we get $n(A_1^c \cap A_2^c)/n \geq 1/2$ a.s., which gives the last estimate in (26). Now (21) follows from (22), (23), (24), (25) and (26).

Step 2. Next, we show that

$$\sup_{\theta \in \Gamma_2} \Big(\log L_n^{mix}(\theta; Y_1^n) + p_n(\theta) \Big) - \log L_n^{mix}(\theta_0^{mix}) - p_n(\theta_0^{mix}) \to -\infty. \tag{27}$$

In the following, the parameters μ_i, σ_i^2 will depend on θ, $i = 1, 2$, which we suppress in the notation. Define the set of indices $A_1 = A(\mu_1, \sigma_1^2)$ as in (18). We recall following bounds from the proof of Lemma 8

$$\varphi(y, \mu_1, \sigma_1^2) \leq \begin{cases} {\sigma_1}^{-1} \exp(-\frac{(\mu_1 - y)^2}{4\sigma_1^2}) & y \in \tilde{A}_1 \\ \exp(-\frac{(\mu_1 - y)^2}{4\sigma_1^2}) & \text{otherwise} \end{cases}.$$

Following Chen and Tan [4] we define a sub-density

$$g(y, \theta) = \pi_1 \exp(-\frac{(\mu_1 - y)^2}{4\sigma_1^2}) + \pi_2 \varphi(y, \mu_2, \sigma_2^2).$$

the function g is bounded by $\varepsilon_0^{-\frac{1}{2}}$ on Γ_2. Following statements hold for every $\theta \in \Gamma_2$:

$$\log f(Y_i,\theta) \le \log g(Y_i,\theta) + \mathbf{1}_{\{i\in A_1\}} \log {\sigma_1}^{-1},$$
$$\log L_n^{mix}(\theta) \le n(A) \log {\sigma_1}^{-1} + \sum_{i=1}^{n} \log g(Y_i,\theta),$$
$$\mathbb{E}_{\theta_0^{mix}} \log g(Y,\theta)/f(Y,\theta_0^{mix}) \le \log \mathbb{E}_{\theta_0^{mix}} g(Y,\theta)/f(Y,\theta_0^{mix}) < 0,$$
$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{g(Y_i,\theta)}{f(Y_i,\theta_0^{mix})} \to \mathbb{E}_{\theta_0^{mix}} \log \frac{g(Y,\theta)}{f(Y,\theta_0^{mix})} < 0.$$

Now by using $\mathbb{E}\sup_{\theta\in U_\epsilon(\theta')} \varphi(Y,\theta) < \infty$ for a sufficiently small neighborhood $U_\epsilon(\theta')$ of a $\theta' \in \Gamma_2$ and considering the compactification of Γ_2 by taking limits with respect to d_c, we apply the classical technique, see Wald [14], to obtain

$$\lim_{n\to\infty} \sup_{\theta\in\Gamma_2} \frac{1}{n}\sum_{i=1}^{n} \log \frac{g(Y_i,\theta)}{f(Y_i,\theta_0^{mix})} =: -\kappa(\tau_0) < 0$$

Where $\kappa(\tau_0)$ is a decreasing function, since larger τ_0 makes Γ_2 larger. Hence for a small enough $\tau_0 \le \varepsilon_0$

$$\begin{aligned}
&\sup_{\theta\in\Gamma_2} \log L_n^{mix}(\theta) + p_n(\theta) - \log L_n^{mix}(\theta_0^{mix}) - p_n(\theta_0^{mix}) \\
&\qquad \le \sup_{\theta\in\Gamma_2} n(A) \log {\sigma_1}^{-1} + p_n(\theta) + \sum_{i=1}^{n} \log g(Y_i,\theta) - \log L_n^{mix}(\theta_0^{mix}) - p_n(\theta_0^{mix}) \\
&\qquad \le \sup_{\theta\in\Gamma_2} \left(\sqrt{n}(\log n)^2 - nM\sigma_1 \log \sigma_1^2 + 1\right) \log {\sigma_1}^{-1} + p_n(\theta) \\
&\qquad + \sup_{\theta\in\Gamma_2} \sum_{i=1}^{n} \log \frac{g(Y_i,\theta)}{f(Y_i,\theta_0^{mix})} - p_n(\theta_0^{mix}) \\
&\qquad \le \kappa(\varepsilon_0)n/2 - n\kappa(\varepsilon_0) = -\kappa(\varepsilon_0)n/2 - p_n(\theta_0^{mix}) \to -\infty.
\end{aligned}$$

We conclude $\hat{\theta}_n^{pIMLE} \in \Gamma_3$ which is regular and contains the true parameter θ_0^{mix}, so $\hat{\theta}_n^{pIMLE}$ is consistent for parameters of the stationary mixture.

The feasible set $\Theta^{full}(\hat{\theta}_n^{pIMLE},\ \delta)$ in stage 2 of the calculation of $\hat{\theta}_n^{pMLE}$ contains a.s. the true parameter θ_0 and the consistency result from Leroux [9] can be applied. It completes the proof of the theorem. □

4. Conclusion

In the presented paper the existence of a consistent, asymptotically normal estimator for Gaussian hidden Markov models was proved. Ideas from the articles Chen et al. [5] and Chen and Tan [4] were used and generalized.

The proof was restricted to the one-dimensional case. The multivariate case could be proved if an analogon of Lemma 11 for more than one dimension would exist. In the i.i.d. setting Chen and Tan [4] in Lemma 2 use an ascription to the univariate case in order derive such a statement. But this ascription is flawed. However, there exists an alternative approach, based on a uniform Law of Iterated Logarithm for VC classes, see Alexandrovich [1]. Unfortunately such an approach is currently not available for dependent observations.

5. Acknowledgements

I would like to thank Prof. Dr. Hajo Holzmann for his help in the writing of this manuscript.

References

[1] Alexandrovich, G. (2014). A note on the article 'Inference for multivariate normal mixtures' by J. Chen and X. Tan. *Preprint submitted to Journal of Multivariate Analysis*.

[2] Bickel, J. P., Ritov, Y. and Rydèn, T. (1998). Asymptotic Normality of the maximum likelihood estimator for general Hidden Markov Models. *The Annals of Statistics*, **26** 1614–1635.

[3] Billingsley, P. (1986). *Probability and Measure*. John Wiley & Sons.

[4] Chen, J. and Tan, X. (2009). Inference for multivariate normal mixtures. *Journal of Multivariate Analysis*, **100** 1367–1383.

[5] Chen, J., Tan, X. and Zhang, R. (2008). Inference for normal mixtures in mean and variance. *Statistica Sinica*, **18** 443–465.

[6] Ciuperca, G., Ridolfi, A. and Idier, J. (2003). Penalized maximum likelihood estimator for normal mixtures. *Scandinavian Journal of Statistics*, **30** 645–59.

[7] Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Annals of Statistics*, **13** 795–800.

[8] Kiefer, J. and Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, **27** 887–906.

[9] Leroux, B. G. (1990). Maximum-likelihood estimation for hidden markov models. *Stochastic Processes and their Applications*, **40** 127–143.

[10] Lindgren, G. (1978). Markov regime models for mixed distributions and switching regressions. *Scand. J. Statistics* 81–91.

[11] Merlevède, F., Peligrad, M. and Rio, E. (2011). A bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, **151** 435–474.

[12] Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Willey, New York.

[13] Tanaka, K. (2009). Strong consistency of the maximum likelihood estimator for finite mixtures of location-scale distributions when penalty is imposed on the ratios of the scale parameters. *Scandinavian Journal of Statistics*, **36** 171–184.

[14] Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics*, **20** 595–601.

[15] Zucchini, W. and MacDonald, I. (2009). *Hidden Markov Models for Time Series*. Chapman & Hall.

A. Mixing properties and Bernstein inequality

Here we deduce the Bernstein-type inequality (15) from Theorem 1 from Merlevède et al. [11]. Let us start by formulating a simplified version of that result.

Definition 12. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $\mathcal{M}_1, \mathcal{M}_2 \subset \mathcal{A}$ sub-sigma-fields, $\mathcal{Z} = (Z_i)_{i\in\mathbb{Z}}$ real valued random variables.

1. The *α-dependence coefficient* between $\mathcal{M}_1$ and $\mathcal{M}_2$ is defined by
$$\alpha(\mathcal{M}_1, \mathcal{M}_2) = \sup\{|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)| : A \in \mathcal{M}_1, B \in \mathcal{M}_2\} \tag{28}$$
2. For the sequence $(Z_i)_{i\in\mathbb{Z}}$ the *α-mixing (or strong-mixing) coefficient* is a function $\mathbb{N} \to \mathbb{R}^+$ defined by
$$\alpha_{\mathcal{Z}}(g) = \sup_{k\in\mathbb{N}} \alpha(\sigma(Z_i,\ -\infty < i \le k),\ \sigma(Z_i,\ k+g \le i < \infty)) \tag{29}$$

The conditions that are needed for the Bernstein inequality are the following. There exist positive constants a, b, γ_1 and c, $\gamma_2 > 0$ such that
$$\alpha(g) \le ae^{-cg^{\gamma_1}}, \tag{B1}$$
$$\sup_i \mathbb{P}(|Z_i| > z) \le e^{1-(z/b)^{\gamma_2}}, \tag{B2}$$

From Merlevède et al. [11], we have the following result.

Lemma 13. *Let $(Z_i)_{i\in\mathbb{Z}}$ be a sequence of centered real valued random variables, which satisfy Assumptions (B1) and (B2). Set $S_j = \sum_{i=1}^{j} Z_i$. Then there exist constants V, γ, C_1, C_2, C_3 and C_4 depending only on the constants a, b, γ_1 and c, $\gamma_2 > 0$ involved in Assumptions (B1) and (B2), such that for all $x > 0$,*

$$\mathbb{P}(\sup_{j\le n}|S_j| \ge x) \le n\exp\left(-\frac{x^\gamma}{C_1}\right) + \exp\left(-\frac{x^2}{C_2(1+nV)}\right) + \exp\left(-\frac{x^2}{C_3 n}\exp\frac{x^{\gamma(1-\gamma)}}{C_4(\log x)^\gamma}\right).$$

In order to deduce (15) from this result, we need to show that given a univariate Gaussian HMM $\mathcal{Y} = (Y_i)_{i\in\mathbb{Z}}$, the conditions (B1) and (B2) hold true for

$$\tilde{Z}^\tau_{i,k} = \mathbf{1}_{\{Y_i\le\eta_k+\tau\}} - \mathbf{1}_{\{Y_i\le\eta_{k-1}\}} - \big(F(\eta_k+\tau) - F(\eta_{k-1})\big), \tag{30}$$

where the constants a, b, γ_1 and c, $\gamma_2 > 0$ do not depend on k and τ. Since

$$\big|\tilde{Z}^\tau_{i,k}\big| \le 2 + 2M, \qquad \forall\ \tau \in (0, e^{-1}],\ 1 \le k \le n,\ n \ge 1,$$

this is evidently possible for (B2) and the constants b and γ_2. For (B1), we first consider the HMM itself. For lack of easy reference, we prove the following well-known result.

Proposition 14. *Let $\mathcal{Y} = (Y_i)_{i\in\mathbb{Z}}$ be a stationary Gaussian Hidden Markov process with a finite state space. Then $\alpha(g) = \mathcal{O}(\rho^g)$ for some $0 < \rho < 1$.*

Proof. Since the process is assumed to be stationary, it suffices to show that

$$\sup\{|\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)| : A \in \sigma(Y_i;\ i \le 0),\ B \in \sigma(Y_i; i \ge g)\} \le c\rho^g \tag{31}$$

for some $c > 0$, $0 < \rho < 1$. First we prove (31) for certain algebras and then show that the sets, which satisfy (31) form a monotone class. An application of the monotone class theorem (e.g. Theorem 8.9 in Billingsley [3]) then completes the proof. We consider the following algebras

$$\mathcal{F}_0 = \{(Y_{i_1}, \dots, Y_{i_m}) \in B \mid B \in \mathcal{B}^m, -\infty < i_1, \dots, i_m < 0, m \in \mathbb{N}\},$$
$$\mathcal{F}_1 = \{(Y_{j_1}, \dots, Y_{j_l}) \in B \mid B \in \mathcal{B}^l, g \le j_1, \dots, j_l < \infty, l \in \mathbb{N}\}.$$

It is easy to see, that $\mathcal{F}_0$ and $\mathcal{F}_1$ are really algebras and generate $\sigma^2(Y_i, -\infty < i \le 0)$ and $\sigma^2(Y_i, g \le i < \infty)$ respectively. Now we assume $A \in \mathcal{F}_0$ and $B \in \mathcal{F}_1$, that is there exist Borel sets B_1 and B_2 so that, $A = \{(Y_{i_1}, \dots, Y_{i_m}) \in B_1\}$ and $B = \{(Y_{j_1}, \dots, Y_{j_l}) \in B_2\}$ for some integer-vectors $(i_1, \dots, i_m)$ and $(j_1, \dots, j_l)$.

For $y \in \mathbb{R}$ we define $\tilde{\mathbb{P}}(y) = \operatorname{diag}(\varphi(y, \mu_1, \sigma_1^2), \dots, \varphi(y, \mu_K, \sigma_K^2))$. With $\mathbb{1}$ we denote a

column-vector of dimension K with 1 at every entry. Now we have

$$\mathbb{P}(A)\mathbb{P}(B) = \int\limits_{B_1} \delta\tilde{\mathbb{P}}(y_1)\prod_{p=2}^{m}\Phi^{i_p-i_{p-1}}\tilde{\mathbb{P}}(y_p)\mathbb{1}dy \int\limits_{B_2} \delta\tilde{\mathbb{P}}(y_1)\prod_{p=2}^{l}\Phi^{i_p-i_{p-1}}\tilde{\mathbb{P}}(y_p)\mathbb{1}dy$$

$$= \int\limits_{B_1\times B_2} \delta\tilde{\mathbb{P}}(y_1)\prod_{p=2}^{m}\Phi^{i_p-i_{p-1}}\tilde{\mathbb{P}}(y_p)\mathbb{1}\delta\tilde{\mathbb{P}}(y_1')\prod_{p=2}^{l}\Phi^{j_p-j_{p-1}}\tilde{\mathbb{P}}(y_p')\mathbb{1}dydy'$$

$$\mathbb{P}(A\cap B) = \int\limits_{B_1\times B_2} \delta\tilde{\mathbb{P}}(y_1)\prod_{p=2}^{m}\Phi^{i_p-i_{p-1}}\tilde{\mathbb{P}}(y_p)\Phi^{j_1-i_m}\tilde{\mathbb{P}}(y_1')\prod_{p=2}^{l}\Phi^{j_p-j_{p-1}}\tilde{\mathbb{P}}(y_p')\mathbb{1}dydy'$$

We have $j_1 - i_m \geq g$ and from Theorem 8.9 in Billingsley [3] we have $\Phi^g \to \mathbb{1}\delta$ with exponential rate, that is $|\Phi^g - \mathbb{1}\delta| \leq c^*\rho^g\mathbb{1}\mathbb{1}^\mathsf{T}$. For some $c^* > 0$ and $0 < \rho < 1$. So we obtain

$$|\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)| = |\int\limits_{B_1\times B_2} \delta\tilde{\mathbb{P}}(y_1)\prod_{p=2}^{m}\Phi^{i_p-i_{p-1}}\tilde{\mathbb{P}}(y_p)\mathbb{1}\delta\tilde{\mathbb{P}}(y_1')\prod_{p=2}^{l}\Phi^{j_p-j_{p-1}}\tilde{\mathbb{P}}(y_p')\mathbb{1}dydy'$$

$$- \int\limits_{B_1\times B_2} \delta\tilde{\mathbb{P}}(y_1)\prod_{p=2}^{m}\Phi^{i_p-i_{p-1}}\tilde{\mathbb{P}}(y_p)\Phi^{j_1-i_m}\tilde{\mathbb{P}}(y_1')\prod_{p=2}^{l}\Phi^{j_p-j_{p-1}}\tilde{\mathbb{P}}(y_p')\mathbb{1}dydy'|$$

$$= |\int\limits_{B_1\times B_2} \delta\tilde{\mathbb{P}}(y_1)\prod_{p=2}^{m}\Phi^{i_p-i_{p-1}}\tilde{\mathbb{P}}(y_p)\left(\mathbb{1}\delta - \Phi^{j_1-i_m}\right)\tilde{\mathbb{P}}(y_1')\prod_{p=2}^{l}\Phi^{j_p-j_{p-1}}\tilde{\mathbb{P}}(y_p')\mathbb{1}dydy'|$$

$$\leq \underbrace{\int\limits_{B_1} \delta\tilde{\mathbb{P}}(y_1)\prod_{p=2}^{m}\Phi^{i_p-i_{p-1}}\tilde{\mathbb{P}}(y_p)dy}_{\leq \mathbb{1}^\mathsf{T}} |\left(\mathbb{1}\delta - \Phi^{j_1-i_m}\right)| \underbrace{\int\limits_{B_2} \tilde{\mathbb{P}}(y_1')\prod_{p=2}^{l}\Phi^{j_p-j_{p-1}}\tilde{\mathbb{P}}(y_p')\mathbb{1}dy'}_{\leq \mathbb{1}}$$

$$\leq c^*\rho^g K^2$$

for every A, B of the assumed form. Here we used the convention $\int f dy = (\int f_1 dy, \ldots, \int f_K dy)$ for the integral of a vector-valued function f. Now, we have that for a fixed $B \in \mathcal{F}_1$, the set M_B of sets A satisfying that inequality builds a monotone class. Indeed, let $A_1 \subset A_2 \subset \ldots \subset A$, where $A_j \in M_B$. The measure $\mathbb{P}$ is continuous from below, so $|\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)| = |\mathbb{P}(\bigcup_{j=1}^{\infty} A_j \cap B) - \mathbb{P}(\bigcup_{j=1}^{\infty} A_j)\mathbb{P}(B)| = |\lim_{j\to\infty} P(A_j\cap B) - \lim_{j\to\infty} P(A_j)\mathbb{P}(B)| = \lim_{j\to\infty} |P(A_j\cap B) - P(A_j)\mathbb{P}(B)| \leq c\rho^g$. The same argument works for $A_1 \supset A_2 \supset \ldots \supset A$, since the measure $\mathbb{P}$ is also continuous from above. So M_A is a monotone class. By the monotone class Theorem (Billingsley, Theorem 3.4) we can extend the inequality on the set $\sigma(\mathcal{F}_0)\times\mathcal{F}_1$. Now we fix an $A \in \sigma(\mathcal{F}_0)$ and the same argumentation applied to the set M_A of sets B satisfying the inequality for this A yields that also M_A is a monotone class. So finally we establish the inequality on the set $\sigma(\mathcal{F}_0)\times\sigma(\mathcal{F}_1)$. □

Lemma 15. *Given a univariate stationary Gaussian HMM, the variables $(\tilde{Z}^{\tau}_{i,k})$ in (30) satisfy the conditions (B1) and (B2), where the constants can be chosen independently of k and τ. Therefore, the Bernstein inequality in Lemma 13 applies, and all constants involved can be chosen independently of k and τ.*

Proof. We already discussed Assumption (B2) above. For (B1), since

$$\sigma(\tilde{Z}^{\tau}_{i,k};\ i \leq 0) \subset \sigma(Y_i;\ i \leq 0), \qquad \sigma(\tilde{Z}^{\tau}_{i,k};\ i \geq g) \subset \sigma(Y_i;\ i \geq g)$$

for any k and τ, the α-mixing coefficients are evidently uniformly bounded by those of the HMM. □

B. Ergodicity

Stationarity affects marginal distributions of a process, while the strong mixing property describes the dependence intensity between process parts as function of the time gap between them. In the next lemma we combine the both properties to conclude ergodicity - a property which allows us to apply a strong law of large numbers to the process.

Lemma 16. *Let $(Y_i)_{i\in\mathbb{Z}}$ be a stationary strong mixing process. Then it is also ergodic.*

Proof. Since $(Y_i)_{i\in\mathbb{Z}}$ is strong mixing, we have for every $n, g \in \mathbb{N},\ A \in \sigma^2(Y^n_{-\infty}),\ B \in \sigma^2(Y^{\infty}_{n+g}) : |\mathbb{P}(A\cap B)-\mathbb{P}(A)\mathbb{P}(B)| < c\rho^g$ for a positive constant c and $0 < \rho < 1$. Now let C be an invariant set, that is there exists a Borel set $B \in \mathcal{B}^{\mathbb{Z}}$, such that $C = \{T^{-k}Y^{\infty}_{-\infty} \in B\}$ for every $k \in \mathbb{N}$, where $T^0 = id,\ T^{-1}Y^{\infty}_{-\infty}(\omega)_n = Y_{n+1}(\omega),\ T^{-k} = T^{-(k-1)} \circ T^{-1}$. So T^{-1} is the left shift and T the right shift. According to Kolmogorov extension theorem, there is a sequence (C_n) of sets $C_n = \{Y^n_{-n} \in B_n\}$, for some cylinder set $B_n \in \mathcal{B}^{2n}$, such that $\mathbb{P}(C \,\Delta\, C_n) < 2^{-n}$, where $C \,\Delta\, C_n = \{C \setminus C_n\} \cup \{C_n \setminus C\}$ is the symmetric difference.
Now since C is invariant we have

$$\mathbb{P}(T^{-k}C \,\Delta\, C_n) = \mathbb{P}(C \,\Delta\, T^k C_n) < 2^{-n},$$

for all $k, n \in \mathbb{N}$. Furthermore $T^k C_n = \{Y^{n-k}_{-n-k} \in B_n\}$, and hence $T^k C_n \in \sigma^2(Y^{n-k}_{-n-k}) \subset \sigma^2(Y^{n-k}_{-\infty})$ and $C_n \in \sigma^2(Y^n_{-n}) \subset \sigma^2(Y^{\infty}_{-n})$. Let $k \geq 2n$, $g_{k,n} = k - 2n$, then using the strong mixing property we conclude

$$|\mathbb{P}(C_n \cap T^k C_n) - \mathbb{P}(C_n)\mathbb{P}(T^k C_n)| < c\rho^{g_{k,n}},$$

for some $c > 0$ and $0 < \rho < 1$. We summarize, for every $\varepsilon > 0$ there exist $n, k \in \mathbb{N}$, such that

1. $||\mathbb{P}(C \cap C) - \mathbb{P}(C)^2| - |\mathbb{P}(C_n \cap T^k C_n) - \mathbb{P}(C_n)\mathbb{P}(T^k C_n)|| < \frac{\varepsilon}{2}$,
2. $|\mathbb{P}(C_n \cap T^k C_n) - \mathbb{P}(C_n)\mathbb{P}(T^k C_n)| < \frac{\varepsilon}{2}$,

and therefore $|\mathbb{P}(C) - \mathbb{P}(C)^2| < \varepsilon$. Since $\varepsilon > 0$ was arbitrary, we have $\mathbb{P}(C) \in \{0, 1\}$. □