COMPOSITE LIKELIHOOD UNDER HIDDEN MARKOV MODEL

Jiahua Chen[†], Yi Huang[†], and Peiming Wang[‡]

[†]*Department of Statistics, University of British Columbia, Vancouver, BC, Canada V6T 1Z2*
[‡]*Department of Finance, Auckland University of Technology, Auckland 1142, New Zealand*

Supplementary Material

Abstract: In this appendix, we provide technical details on asymptotic properties of the composite likelihood estimators in the corresponding paper. All statements of the asymptotic results and the related definitions can be found in the main paper. We have included some key information in the appendix to make it nearly self-contained.

Suppose we have a set of observations $\{Y_t\}_{t=1}^T$ from a hidden Markov model in which the conditional distribution of Y_t given its hidden state is $f(y;\theta)$. The marginal joint distribution of Y_t, Y_{t+1} has density function

$$f(y_t, y_{t+1}; \Psi) = \int f(y_t;\theta) f(y_{t+1};\theta^*)\, \mathrm{d}\Psi \tag{1}$$

for some mixing distribution Ψ. We define the composition log likelihood after some regularization as

$$\ell_{cl}(\Psi) = \sum_{t=1}^{T-1} \log f(y_t, y_{t+1}; \Psi) + C \sum_{i,\,j} \log \pi_{ij}. \tag{2}$$

Given a random sample of n observations from $f(y_1, y_2; \Psi)$, the corresponding MLE $\hat{\Psi}$ for Ψ is consistent under simple conditions on $f(y;\theta)$; see Kiefer and Wolfowitz (1956). In this paper, $y(1{:}T) = \{Y_t\}_{t=1}^T$ is instead a time series generated according to an HMM. We show that a nearly identical proof is applicable under a set of high-level assumptions. These assumptions can be easily verified from conditions similar to those in Kiefer and Wolfowitz (1956).

A1 (Identifiability). $F(y_1, y_2; \Psi_1) = F(y_1, y_2; \Psi_2)$ for all (y_1, y_2) if and only if $\Psi_1 = \Psi_2$.

A2 (Compactness). The space of the mixing distribution Ψ can be expanded to form a compact metric space $\mathcal{M}$; and $f(y_1, y_2; \Psi)$ can be continuously extended to $\mathcal{M}$.

A3 (Jensen's inequality). Let Ψ_0 be the true mixing distribution under the HMM. After extension, for any $\Psi \in \mathcal{M}$ and $\Psi \neq \Psi_0$, there exists an $\epsilon > 0$ such that

$$E_{\Psi_0}\{\log[f(Y_1, Y_2; \Psi, \epsilon)/f(Y_1, Y_2; \Psi_0)]\} < 0$$

where $f(Y_1, Y_2; \Psi, \epsilon) = \max\big[1, \sup\{f(Y_1, Y_2; \Psi') : \|\Psi' - \Psi\| \le \epsilon\}\big]$.

A4 The time series $y(1{:}T)$ is from a finite-state HMM such that its hidden Markov chain is irreducible and the series is in equilibrium.

One may notice that A3 implies A1. At the same time, A1 together with some continuity and integration conditions implies A3; retaining A1 makes the proof easier to understand. Establishing A2 and A3 were major steps in the proof of Kiefer and Wolfowitz (1956). A2 is violated when the kernel distribution is two-parameter normal. The consistent result under a normal kernel will be addressed with additional steps. The norm $\|\cdot\|$ is the distance between two mixing distributions. The high-level assumptions are helpful for avoiding non-innovative and tedious details in the proof. With A4, the essential steps in Kiefer and Wolfowitz (1956) remain valid under the HMM.

Theorem 1: *Under A1–A4, the maximum composite likelihood estimator of Ψ is strongly consistent as $T \to \infty$.*

Proof: A key step is to show the validity of the law of large numbers for $\log f(Y_1, Y_2; \Psi, \epsilon)$. Leroux (1992) shows that when the HMM is in equilibrium and the Markov chain is irreducible (A4), $Y(1{:}T)$ is ergodic. Together with A3, this implies

$$T^{-1} \sum_{t=1}^{T-1} \log f(Y_t, Y_{t+1}; \Psi, \epsilon) \to E_{\Psi_0} \log f(Y_t, Y_{t+1}; \Psi, \epsilon)$$

almost surely. With the addition of A2, it implies that for any $\epsilon > 0$, there is a $\delta > 0$ such that

$$\sup\{\sum_{t=1}^{T-1} \log f(Y_t, Y_{t+1}; \Psi) : \|\Psi - \Psi_0\| > \epsilon\} < \sum_{t=1}^{T-1} \log f(Y_t, Y_{t+1}; \Psi_0) - \delta T$$

almost surely. The regularization term $\sum \log \pi_{ij} = O(\log T)$ when $\min \pi_{ij} \ge T^{-1}$. The smoothness of $f(x;\theta)$ in θ implies

$$E_{\Psi_0} \log f(Y_t, Y_{t+1}; \Psi) - E_{\Psi_0} \log f(Y_t, Y_{t+1}; \Psi_0) \to 0$$

when $\Psi \to \Psi_0$. Thus, if none of $\pi_{ij} = 0$ in Ψ_0, then $\sum \log \pi_{ij} = O(1)$. If some $\pi_{ij} = 0$ in Ψ_0, we can find a Ψ in a T^{-1}-distance neighborhood such that its $\sum \log \pi_{ij} = O(\log T)$ and

$$E_{\Psi_0} \log f(Y_t, Y_{t+1}; \Psi) \ge E_{\Psi_0} \log f(Y_t, Y_{t+1}; \Psi_0) - \delta/2. \quad (3)$$

Clearly, these discussions imply

$$\sup\{\ell_{cl}(\Psi) : \|\Psi - \Psi_0\| > \epsilon\} < \sup\{\ell_{cl}(\Psi) : \|\Psi - \Psi_0\| < T^{-1}\}.$$

Hence, the MLE of Ψ is in an ϵ-neighborhood of Ψ_0 almost surely for some $\epsilon > 0$. This implies the strong consistency and completes the proof.

Notably, if the kernel distribution of the HMM is $N(\mu, \sigma^2)$, A2 is not satisfied. Thus, Theorem 1 is not directly applicable. For this special case, we have defined

$$\ell_{pcl}(\Psi) = \ell_{cl}(\Psi) - T^{-1/2} \sum_{i=1}^{N} \{\log(\sigma_i^2/\hat{\sigma}_0^2) + \hat{\sigma}_0^2/\sigma_i^2\}, \tag{4}$$

with a regularization term on σ^2. This regularization term confines the MLE into a restricted space of Ψ on which A1–A4 are satisfied.

Lemma 1 *Suppose $\{Y_t\}_{t=1}^{\infty}$ is generated by an HMM with a normal kernel. Let $\hat{\Psi}$ be the maximum point of the specially regularized composite likelihood $\ell_{pcl}(\Psi)$ given in (4). Then, there exists a small enough ϵ such that as $T \to \infty$,*

$$P\Big(\min_{1 \le j \le N} \hat{\sigma}_j^2 > \epsilon\Big) \to 1.$$

The proof is omitted because it is similar to that of Chen et al. (2008) but much more involved.

Corollary 1 *Under the conditions of Lemma 1, the maximum point $\hat{\Psi}$ of $\ell_{pcl}(\Psi)$ defined in (4) is consistent for Ψ.*

We next investigate the asymptotic normality. Consider the situation where Ψ is a smooth function of some identifiable vector γ with true mixing distribution Ψ_0 corresponding to γ_0 that is in the interior of its corresponding parameter space. We further assume that $\pi_{ij} \neq 0$ at γ_0. Bickel et al. (1998) imposed practically the same conditions on Ψ for MLE under the FL. The asymptotic normality under more relaxed conditions is likely tedious and is more appropriate as a separate research project.

For notational simplicity, we introduce $g(x_t; \gamma) = \log f(y_t, y_{t+1}; \Psi)$ and assume that it is twice differentiable with respect to γ. That is, we simplify $\log f$ into g, (y_t, y_{t+1}) to x_t, and highlight that Ψ is in fact determined by γ. Its derivatives with respect to γ are conveniently written as $g'(\cdot)$ and $g''(\cdot)$.

The regularized composite likelihood is written as

$$\ell_{cl}(\gamma) = \sum_{t=1}^{T-1} g(x_t; \gamma) + \sum_{i,j} \log \pi_{ij}.$$

The consistency of the CL MLE of Ψ leads to the consistency of the CL MLE $\hat{\gamma}$. The consistency and γ_0 being an interior point imply that $\ell'_{cl}(\hat{\gamma}) = 0$. Let the implied MLE be $\hat{\pi}_{ij}$. We must have

$$0 = \ell'_{cl}(\hat{\gamma}) = \sum_{t=1}^{T-1} g'(x_t; \gamma_0) + \sum_{t=1}^{T-1} g''(\gamma_0)(\hat{\gamma} - \gamma_0)\{1 + o_p(1)\} + \sum_{i,j} \partial(\log \hat{\pi}_{ij})/\partial\gamma.$$

Clearly, $\hat{\gamma}$ is asymptotically normal when $T^{-1/2}\sum g'(x_t;\gamma_0)$ is, and $T^{-1}\sum g''(x_t;\gamma_0)$ has a positive definite matrix limit. The regularization term has no effect on the limiting distribution if

$$\sum_{i,j} \partial(\log \hat{\pi}_{ij})/\partial\gamma = o_p(T^{-1/2})$$

uniformly in a neighborhood of γ_0. This is implied when γ_0 is an interior point.

The α-mixing coefficients of a stationary process $\{X_t\}_{t=0}^{\pm\infty}$ are defined to be

$$\alpha(t) = \sup_{A\in\mathcal{F}_t^{\infty};B\in\mathcal{F}_{-\infty}^{0}} |P(AB) - P(A)P(B)|$$

for $t = 1, 2, \ldots$, where $\mathcal{F}_n^m$ is the σ-algebra generated by $\{X_t\}_{t=n}^m$. We base our normality proof on the following standard result of Ibragimov (1962).

Lemma 2 *Suppose a stationary process* $\{X_t\}_{t=0}^{\pm\infty}$ *satisfies:*

(a) $E(X_0) = 0$ *and* $E(|X_0|^3) < \infty$*;*

(b) $\sum_{t=1}^{\infty}\{\alpha(t)\}^{1/3} < \infty$ *for its* α*-mixing coefficients.*

Then, as $T \to \infty$*,*

$$\sigma^2 = \lim_{T\to\infty} T^{-1}\mathrm{Var}\left(\sum_{t=1}^{T} X_t\right) < \infty$$

and $T^{-1/2}\sum_{t=1}^{T} X_t \to N(0,\sigma^2)$ *in distribution.*

Consider the stochastic process $g'(X_t;\gamma_0)$. The equilibrium assumption A4 implies that $\{g'(X_t;\gamma_0)\}_{t=0}^{\infty}$ is stationary. The HMM with finite number of state leads to $\alpha(t)$ decaying at an exponential rate (Durrett, 2010, Page 264). Hence, condition (b) is satisfied.

The moment conditions in (a) must be verified for each $f(y;\theta)$. They are satisfied by most commonly used distributions such as the Poisson, binomial, and normal distributions. Let

$$\Sigma(\gamma_0) = \lim_{T\to\infty} T^{-1}\mathrm{Var}\left\{\sum_{t=1}^{T} g'(X_t;\gamma_0)\right\}.$$

We have

$$T^{-1}\sum_{t=1}^{T-1} g''(X_t;\gamma_0) \to I(\gamma_0)$$

with $I(\gamma_0) = E\{g''(X_t;\gamma_0\}$. The above discussion leads to

$$T^{1/2}(\hat{\gamma}-\gamma_0) = \{T^{-1}\ell''(\gamma_0)\}^{-1}\{T^{-1/2}\ell'(\gamma_0)\} + o_p(1) \to N(0,V)$$

with $V = I^{-1}(\gamma_0)\Sigma(\gamma_0)I^{-1}(\gamma_0)$.

Even though many of these conditions can be verified from the basics as discussed above, we choose to state the result under "high level" conditions as follows.

Theorem 2 *Let $\{Y_t\}_{t=0}^{\infty}$ be a time series satisfying A1–A4. Assume*

(a) Ψ in the composite likelihood (2) is a twice differentiable function of γ;

(b) the induced stochastic process $g'(X_t;\gamma_0)$ is well defined and satisfies the conditions of Lemma 2;

(c) the true value γ_0 is an interior point in the space of γ;

(d) the limit $I(\gamma_0)$ is positive definite and finite.

Then, the composite likelihood MLE of γ is asymptotically normal: $T^{1/2}(\hat{\gamma}-\gamma_0) \to N(0,V)$ with $V = I^{-1}(\gamma_0)\Sigma(\gamma_0)I^{-1}(\gamma_0)$.

Because the true variances $\sigma_j^2 > 0$, the consistency of $\hat{\Psi}$ established in Corollary 1 for the HMM with a normal kernel leads to $\hat{\sigma}_j^2 \geq \delta > 0$ in probability for some δ. Once σ_j's are constrained away from zero, the proof of Theorem 2 is equally applicable to the HMM with a normal kernel and we state the result as follows.

Corollary 2 *Under the conditions of Lemma 2, the maximum point $\hat{\Psi}$ of $\ell_{pcl}(\Psi)$ is asymptotically normal as stated in Theorem 2.*

Clearly, $\Sigma(\gamma_0)$ is a sum of infinite series. This expression is not very useful for constructing a corresponding variance estimator. This is also the case for Bickel et al. (1998). Developing easily applicable asymptotic results remains a challenging task, but it will be much simpler for the CL than for the FL.

References

Bickel, P. J., Ritov, Y., and Rydén, T. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Annals of Statistics* **26**, 1614–1635.

Durrett, R. (2010). *Probability: Theory and Examples.* Cambridge University Press, 4th edition.

Ibragimov, I. A. (1962). Some limit theorems for stationary processes. *Theory of Probability and Its Applications* **7**, 349–382. (English translation).

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.

Leroux, B. G. (1992). Maximum-likelihood estimation for hidden Markov models. *Stochastic Processes and their Applications* **40**, 127–143.