Maximum likelihood estimator and Kullback–Leibler information in misspecified Markov chain models

Priscilla E. Greenwood
University of British Columbia

Wolfgang Wefelmeyer
University of Siegen

Abstract

Suppose we have specified a parametric model for the transition distribution of a Markov chain, but that the true transition distribution does not belong to the model. Then the maximum likelihood estimator estimates the parameter which maximizes the Kullback–Leibler information between the true transition distribution and the model. We prove that the maximum likelihood estimator is asymptotically efficient in a nonparametric sense if the true transition distribution is unknown.

1 Introduction

Suppose we observe $X_0, \dots, X_n$ from an ergodic Markov chain on an arbitrary state space. We have specified a parametric model $Q_\vartheta(x, dy)$ for the transition distribution, and an initial distribution $\eta_0(dx)$. Consider the following two situations:

1. We believe, erroneously, that the model is correct, and use the maximum likelihood estimator for estimating the parameter.

2. We know that the model is incorrect, and want to fit a transition distribution from the model to the true transition distribution, using Kullback–Leibler information as 'distance'.

Both situations lead to the same problem: Let $Q(x, dy)$ be the true transition distribution and $\eta(dx)$ the true initial distribution. Write $\pi(dx)$ for the invariant distribution. Suppose that $Q_\vartheta(x, dy)$ has a density $q_\vartheta(x, y)$ with respect to some dominating measure $m(x, dy)$. By the *Kullback–Leibler information* we mean the expectation $\pi Q \log q_\vartheta$ of $\log q_\vartheta(x, y)$ under the joint invariant distribution $\pi \otimes Q = \pi(dx)Q(x, dy)$ of two successive

[1] Work supported by NSERC, Canada.

[2] *AMS 1991 subject classifications:* 62G05, 62G20, 62M05.

[3] *Key words and Phrases:* Efficient estimation, Kullback–Leibler information, Markov chain, maximum likelihood estimator, misspecified model.

observations. Write $k(\pi \otimes Q)$ for the parameter ϑ which maximizes the Kullback–Leibler information. A natural estimator for $\pi \otimes Q$ is the *empirical measure*

$$E_n = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_{(X_{i-1}, X_i)}.$$

Then $k(E_n)$ is an estimator for the *maximum Kullback–Leibler information functional* $k(\pi \otimes Q)$. By definition of the functional, $\vartheta = k(E_n)$ maximizes

$$E_n \log q_\vartheta = \frac{1}{n} \sum_{i=1}^{n} \log q_\vartheta (X_{i-1}, X_i).$$

This means that $k(E_n)$ is the *maximum likelihood estimator.*

We see that in both situations we are led to use the maximum likelihood estimator for estimating the maximum Kullback–Leibler information functional. Here we will be interested in the following question: Is the maximum likelihood estimator efficient for this functional if the model is misspecified? We expect that it is efficient according to the following heuristic argument. The empirical distribution E_n is efficient for $\pi \otimes Q$ in a certain sense. See Greenwood and Wefelmeyer (1995), extending Penev (1991) who considers estimating the invariant distribution π. If one were to prove that k is differentiable in a suitable sense, efficiency of the maximum likelihood estimator would follow.

We do not pursue this approach here. Instead, we use that the maximum likelihood estimator solves an estimating equation and admits the following stochastic expansion. For simplicity, suppose that ϑ varies in a compact subset of the real line. Write ℓ'_ϑ and ℓ''_ϑ, respectively, for the first and second derivative of $\log q_\vartheta$ with respect to ϑ. From now on we write $k(Q)$ for $k(\pi \otimes Q)$. Since $\vartheta = k(Q)$ maximizes $\pi Q \log q_\vartheta$, we have $\pi Q \ell'_{k(Q)} = 0$. Let $\vartheta = \hat{\vartheta}$ be a consistent solution of

$$n^{-1/2} \sum_{i=1}^{n} \ell'_\vartheta (X_{i-1}, X_i) = o_P(1),$$

for instance the maximum likelihood estimator. A Taylor expansion gives

$$n^{1/2} \left(\hat{\vartheta} - k(Q) \right) = -(\pi Q \ell''_{k(Q)})^{-1} n^{-1/2} \sum_{i=1}^{n} \ell'_{k(Q)} (X_{i-1}, X_i) + o_P(1).$$

Using $\pi Q \ell''_{k(Q)} = \partial_{\vartheta = k(Q)} \pi Q \ell'_\vartheta = -D$, say, we could avoid writing a second derivative of q_ϑ and get

$$n^{1/2} \left(\hat{\vartheta} - k(Q) \right) = D^{-1} n^{-1/2} \sum_{i=1}^{n} \ell'_{k(Q)} (X_{i-1}, X_i) + o_P(1), \tag{1.1}$$

and hence asymptotic normality of the estimator by a central limit theorem for Markov chains.

In the i.i.d. case, Huber (1967) gives conditions for (1.1) to hold for the maximum likelihood estimator. They do not involve a second derivative of q_ϑ. Ogata (1980) translates the argument to Markov chains. Weaker conditions are given by Pollard (1985) for the i.i.d. case, and by Hosoya (1989) for stationary linear processes. See also Andrews and Pollard (1994). McKeague (1984) and Kutoyants (1988) prove asymptotic normality of the maximum likelihood estimator for misspecified diffusion models.

It is the purpose of this paper to show that the stochastic expansion (1.1) leads to an optimality property of the estimator $\hat{\vartheta}$ when the true transition distribution may not be in the parametric model, or even in a local neighborhood.

Our result is new even in the i.i.d. case, for which the proofs simplify considerably. The only other result on efficiency of an estimator under a fixed distribution outside a misspecified model we are aware of is Theorem 5 in Beran (1977). He uses the Hellinger distance to fit a parametric model for i.i.d. observations. He shows that the minimum Hellinger distance functional is a boundedly differentiable function of the true density, and estimates the functional by replacing the density by an estimator. He proves that the estimator is efficient under misspecification, and robust at the parametric model. Bounded differentiability of minimum Hellinger distance functionals is also obtained by Yang (1991) and Ying (1992).

We use arguments similar to those in Greenwood and Wefelmeyer (1995). There we considered efficiency of empirical estimators $n^{-1}\sum_{i=1}^n f(X_{i-1},X_i)$ with f bounded. Here the function which arises is $f=\ell'_{k(Q)}$ which will, in general, be unbounded.

The result given here generalizes easily to finite-dimensional parameters and to higher-order Markov chains. It probably remains true for observations coming from more general time series, at least if they are locally asymptotically normal in an appropriate sense.

2 An optimality property of the maximum likelihood estimator

Let $X_0,\dots,X_n$ be observations from a Markov chain with values in a measurable state space E. Let $Q(x,dy)$ denote the transition distribution, and $\eta(dx)$ the initial distribution.

As usual, write $\eta Q(dy)=\int\eta(dx)Q(x,dy)$ and $\eta\otimes Q(dx,dy)=\eta(dx)Q(x,dy)$. For a function $f(x)$ write $\eta f=\int\eta(dx)f(x)$ and $Qf(x)=Q_xf=\int Q(x,dy)f(y)$. For a function $f(x,y)$ of two arguments, we interpret Qf as $Qf(x)=Q_xf=\int Q(x,dy)f(x,y)$ and Q^kf as $Q^{k-1}Qf$.

Assumption 1. *The Markov chain is positive Harris recurrent with invariant distribution π.*

Let $\|f\|=(\pi f^2)^{1/2}$ denote the norm of $L_2(\pi)$, and $\|R\|=\sup\{\|Rf\|:\|f\|=1\}$ the corresponding operator norm of a transition kernel $R(x,dy)$. Set $\Pi(x,dy)=\pi(dy)$.

Assumption 2. *We have $\|Q^j - \Pi\| \to 0$ for $j \to \infty$.*

We wish to prove that any estimator $\hat{\vartheta}$ with stochastic approximation (1.1) is efficient in the sense of a Hájek–LeCam convolution theorem described as follows. The initial distribution η is fixed. We regard the collection of transition distributions on E as a parameter space for the distributions governing the data. Under Assumption 1, the model is *locally asymptotically normal* at the true transition distribution Q, fixed above, in the following sense. Let

$$H = \left\{ h : E^2 \to \mathbb{R} \text{ bounded, measurable, } Q_x h = 0 \text{ for all } x \in E \right\}.$$

This space plays the role of local parameter space. For $h \in H$ set

$$Q_{nh}(x, dy) = Q(x, dy) \left(1 + n^{-1/2} h(x, y)\right).$$

Write P_n and P_{nh} for the joint distribution of $X_0, \dots, X_n$ if Q and Q_{nh}, respectively, are true. Then, under Assumption 1,

$$\log dP_{nh} / dP_n = n^{-1/2} \sum_{i=1}^{n} h(X_{i-1}, X_i) - \frac{1}{2} \pi Q h^2 + o_{P_n}(1) \tag{2.1}$$

and

$$n^{-1/2} \sum_{i=1}^{n} h(X_{i-1}, X_i) \Rightarrow N_h, \tag{2.2}$$

where N_h is normal with mean 0 and variance $\pi Q h^2$. This nonparametric version of local asymptotic normality is due to Penev (1991).

The $L_2(\pi \otimes Q)$-closure of H is

$$\overline{H} = \{ h \in L_2(\pi \otimes Q) : Q_x h = 0 \text{ for all } x \in E \}.$$

Call a functional $k(Q)$ *differentiable* at Q with *canonical gradient* g if $g \in \overline{H}$ and

$$n^{1/2} \left(k(Q_{nh}) - k(Q) \right) \to \pi Q h g, \qquad h \in H. \tag{2.3}$$

Call an estimator $\hat{k}$ *regular* for k at Q with *limit* L if

$$n^{1/2} \left(\hat{k} - k(Q_{nh}) \right) \Rightarrow L \text{ under } P_{nh}, \qquad h \in H.$$

The convolution theorem says that $L = M + N$, where M is independent of N, and N is normal with mean 0 and variance $\pi Q g^2$. This justifies calling $\hat{k}$ *efficient* if its limit distribution is N.

Call an estimator $\hat{k}$ *asymptotically linear* for k at Q with *influence function* f if

$$n^{1/2} \left(\hat{k} - k(Q) \right) = n^{-1/2} \sum_{i=1}^{n} f(X_{i-1}, X_i) + o_{P_n}(1). \tag{2.4}$$

We have the following characterization.

Proposition. *Under Assumption 1, an estimator is regular and efficient for a differentiable functional k at Q if and only if it is asymptotically linear with influence function equal to the canonical gradient of k.*

For an appropriate version of the convolution theorem and the Proposition see Greenwood and Wefelmeyer (1990).

Now we turn to our misspecified model. It is determined by a parametric family $Q_\vartheta(x, dy)$ of transition distributions and an initial distribution $\eta_0(dx)$ on the state space E.

Assumption 3. *The parameter space Θ is compact. For $x \in E$, the transition distributions $Q_\vartheta(x, dy)$ have a density $q_\vartheta(x, y)$ with respect to some dominating measure $m(x, dy)$, and q_ϑ is measurable. There is a unique $\vartheta = k(Q)$ in the interior of Θ which maximizes the Kullback–Leibler information $\pi Q \log q_\vartheta$. The function $\log q_\vartheta$ is twice differentiable in the following sense: For $\vartheta \in \Theta$ and $\tau \to \vartheta$,*

$$\log q_\tau = \log q_\vartheta + (\tau - \vartheta)(\ell'_\vartheta + r_\tau) \tag{2.5}$$

with $\ell'_\vartheta \in L_2(\pi \otimes Q)$ and $r_\tau \to 0$ in $L_2(\pi \otimes Q)$. For $\vartheta \to k = k(Q)$,

$$\ell'_\vartheta = \ell'_k + (\vartheta - k)(\ell''_k + s_\vartheta) \tag{2.6}$$

with $\ell''_k \in L_2(\pi \otimes Q)$ and $s_\vartheta \to 0$ in $L_2(\pi \otimes Q)$. Further,

$$D := -\pi Q \ell''_k > 0.$$

For $f \in L_2(\pi)$ define

$$(Af)(x, y) = \sum_{j=0}^{\infty} (Q_y^j f - Q_x^{j+1} f). \tag{2.7}$$

Theorem. *Under Assumptions 1 and 2 on the true transition distribution and Assumption 3 on the parametric model, any estimator $\hat\vartheta$ fulfilling*

$$n^{1/2}\left(\hat\vartheta - k(Q)\right) = D^{-1} n^{-1/2} \sum_{i=1}^{n} \ell'_{k(Q)}(X_{i-1}, X_i) + o_{P_n}(1) \tag{2.8}$$

is regular and efficient for the maximum Kullback–Leibler information functional k at Q. The canonical gradient of k and the influence function of $\hat\vartheta$ are both (setting $k = k(Q)$)

$$g(x, y) = D^{-1}\left(\ell'_k(x, y) - Q_x \ell'_k + (AQ\ell'_k)(x, y)\right) \tag{2.9}$$

with A defined in (2.7). *The asymptotic variance of $\hat{\vartheta}$ is*

$$\pi Q g^2 = D^{-2} \left(\pi Q \ell_k'^2 + 2 \sum_{j=1}^{\infty} \iint \pi(dx) Q(x,dy) \ell_k'(x,y) Q_y^j \ell_k' \right).$$

The maximum likelihood estimator fulfills (2.8) under appropriate conditions, as detailed in the Introduction.

The proof of the Theorem is based on Lemmas 1, 5 and 6 in Section 3 below. Definition (2.3) of differentiability requires that $k(Q_{nh})$ is defined for n sufficiently large. This is shown in Lemma 4.

Proof. The Proposition characterizes regular and efficient estimators. By Lemma 6, the functional k is differentiable at Q with canonical gradient g given by (2.9). It remains to show that $\hat{\vartheta}$ has influence function equal to g. Write $k = k(Q)$. Rewrite (2.8) as

$$n^{1/2}(\hat{\vartheta} - k) = D^{-1} n^{-1/2} \sum_{i=1}^{n} \left(\ell_k'(X_{i-1}, X_i) - Q(X_{i-1}, \ell_k') + Q(X_{i-1}, \ell_k') \right) + o_{P_n}(1).$$

We have $\pi Q \ell_k' = 0$ by Lemma 5. From the martingale approximation in Lemma 1, applied for $f = Q\ell_k'$, we obtain

$$n^{-1/2} \sum_{i=1}^{n} Q(X_{i-1}, \ell_k') = n^{-1/2} \sum_{i=1}^{n} (AQ\ell_k')(X_{i-1}, X_i) + o_{P_n}(1).$$

From definition (2.9) of g we see that

$$n^{1/2}(\hat{\vartheta} - k) = n^{-1/2} \sum_{i=1}^{n} g(X_{i-1}, X_i) + o_{P_n}(1).$$

Hence $\hat{\vartheta}$ is asymptotically linear in the sense of (2.4) with influence function equal to the canonical gradient g, and the Theorem is proved. □

If Q happens to lie in the parametric model, say $Q = Q_\vartheta$, then $k(Q_\vartheta) = \vartheta$. In this case, $Q_{\vartheta x} \ell_\vartheta' = 0$ for all $x \in E$, and $\sum_{i=1}^n \ell_\vartheta'(X_{i-1}, X_i)$ is a martingale. The gradient reduces to

$$g_\vartheta(x,y) = D_\vartheta^{-1} \ell_\vartheta'(x,y)$$

with $D_\vartheta = -\pi_\vartheta Q_\vartheta \ell_\vartheta''$, and the asymptotic variance reduces to the inverse of the Fisher information,

$$\pi_\vartheta Q_\vartheta g_\vartheta^2 = D_\vartheta^{-2} \pi_\vartheta Q_\vartheta \ell_\vartheta'^2 = 1 \Big/ \pi_\vartheta Q_\vartheta \ell_\vartheta'^2 .$$

3 Lemmas

We will make use of the following martingale approximation of Gordin and Lifšic (1978, Remark 3); see also Durrett (1991, p. 375). The idea goes back to Gordin (1969).

Lemma 1. *Under Assumption 2, for $f \in L_2(\pi)$,*

$$n^{-1/2}\sum_{i=1}^{n}(f(X_{i-1})-\pi f) = n^{-1/2}\sum_{i=1}^{n}(Af)(X_{i-1},X_i)+o_{P_n}(1)$$

with A defined in (2.7).

The maximum Kullback–Leibler information functional $k(Q)$ maximizes $\pi Q \log q_\vartheta$. We will see that $\pi Q \log q_\vartheta$ has derivative $\pi Q \ell'_{k(Q)}$ at $\vartheta = k(Q)$, and we obtain the gradient of $k(Q)$ from the equation $\pi Q \ell'_{k(Q)} = 0$. To prove that $k(Q)$ is differentiable, we must know that $\pi Q \ell'_\vartheta$ varies smoothly with Q. We do not wish to assume that ℓ'_ϑ is bounded. A reasonable condition on ℓ'_ϑ is $\pi \otimes Q$-square integrability which leads to the question whether a functional of the transition distribution of the form πf is differentiable if f is π-square integrable. In particular, we must prove that π varies continuously with Q. This is a stability property of the invariant distribution. Our arguments follow those of Kartashov (1985a), (1985b), who proves a different version of stability. The replacement of R by A is also implicit in Penev (1991). Set $I(x,dy) = \varepsilon_x(dy)$. Recall the notation $\Pi(x,dy) = \pi(dy)$.

Lemma 2. *Under Assumption 2, the operator $I - Q + \Pi$ on $L_2(\pi)$ has a bounded inverse*

$$R = I + \sum_{j=1}^{\infty}(Q^j - \Pi). \tag{3.1}$$

Each transition distribution Q' with $\|Q'-Q\| < \|R\|^{-1}$ has a unique invariant distribution $\pi' = \pi\left(I-(Q'-Q)R\right)^{-1}$ such that uniformly for $f \in L_2(\pi)$ with $\|f\| \le 1$,

$$\pi' f - \pi f = \pi(Q'-Q)Rf + O(\|Q'-Q\|^2). \tag{3.2}$$

In particular, there are an $\varepsilon > 0$ and a c such that $\|Q'-Q\| < \varepsilon$ implies for $f \in L_2(\pi)$ with $\|f\| \le 1$,

$$|\pi' f - \pi f| \le c\|Q'-Q\|. \tag{3.3}$$

The operator R in (3.2) *can be replaced by the operator A defined in* (2.7).

Proof. By Assumption 2, $\|Q^j - \Pi\| \le ba^j$, and the operator R given by (3.1) is the bounded inverse of $I - Q + \Pi$. For $\|Q'-Q\| < \|R\|^{-1}$, the operator $I-(Q'-Q)R$ has a bounded inverse S. Set $\pi' = \pi S$. We show that π' is the unique invariant distribution of Q'. First apply S^{-1} to $\pi' = \pi S$ and obtain

$$\pi = \pi'\left(I-(Q'-Q)R\right) = \pi' - \pi'(Q'-Q)R.$$

Then apply R^{-1} to obtain

$$\pi = \pi R^{-1} = \pi'(I - Q + \Pi) - \pi'(Q' - Q) = \pi' - \pi' Q' + \pi.$$

Hence $\pi' Q' = \pi'$. Reversing the steps, we see that the invariant distribution is unique. Assertion (3.2) now follows by von Neumann expansion of S. Since R is bounded, (3.2) implies (3.3).

It remains to show that R can be replaced by A. Recall definitions (2.7) and (3.1) of the operators A and R and use

$$(Q'_x - Q_x)(Q^j - \Pi)f = \int (Q' - Q)(x, dy) \int \left(Q^j(y, dz) - Q^{j+1}(x, dz)\right) f(z)$$

to write

$$\pi(Q' - Q)Rf = \pi(Q' - Q)Af.$$

Hence (3.2) holds also with R replaced by A. □

Lemma 3. *Under Assumption 2, for $h \in H$ and n sufficiently large, Q_{nh} has a unique invariant distribution π_{nh}, and uniformly for functions $f(x, y)$ with $\pi Q f^2 \leq 1$,*

$$n^{1/2}(\pi_{nh} Q_{nh} f - \pi Q f) \to \pi Q h(f - Qf + AQf). \tag{3.4}$$

Proof. By definition of Q_{nh}, the sequence $n^{1/2}\|Q_{nh} - Q\|$ is bounded. Recall that $\pi(Qf)^2 \leq \pi Q f^2$. Use Lemma 2 with f replaced by Qf and $Q_x h = 0$ for all x to obtain

$$\begin{aligned} n^{1/2}(\pi_{nh} Q_{nh} f - \pi Q f) &= \pi Q h f + n^{1/2}(\pi_{nh} - \pi) Q f + (\pi_{nh} - \pi) Q h f \\ &\to \pi Q h f + \pi Q h A Q f = \pi Q h (f - Qf + AQf). \end{aligned}$$

The next three lemmas involve the misspecified model. The argument follows, in part, Beran (1977), who considers the i.i.d. case and Hellinger distance.

Lemma 4. *Let Assumption 2 hold. Let Θ be compact and $\log q_\vartheta$ continuous in $L_2(\pi \otimes Q)$. Then, for $h \in H$ and n sufficiently large, there exists $\vartheta = k(Q_{nh})$ for which the Kullback–Leibler information $\pi_{nh} Q_{nh} \log q_\vartheta$ is maximized. If the maximum of $\pi Q \log q_\vartheta$ is unique, then $k(Q_{nh}) \to k(Q)$.*

Proof. We show that for n sufficiently large there exists $\vartheta = k(Q_{nh})$ for which $\pi_{nh} Q_{nh} \log q_\vartheta$ is maximized. Fix n and h. Set $S = (I - (Q_{nh} - Q)R)^{-1}$. By Lemma 2 we have $\pi_{nh} = \pi S$. Hence

$$\pi_{nh} Q_{nh} \log q_\vartheta = \pi S Q_{nh} \log q_\vartheta.$$

For n sufficiently large, the operators S and Q_{nh} are continuous in $L_2(\pi)$. By Assumption 3, the function $\vartheta \to \log q_\vartheta$ is continuous in $L_2(\pi)$. Hence $\pi_{nh} Q_{nh} \log q_\vartheta$ is continuous in ϑ, and the maximum is attained.

Furthermore, since the operator A is also continuous in $L_2(\pi)$, Lemma 3 implies

$$\sup_{\vartheta\in\Theta} |\pi_{nh}Q_{nh}\log q_\vartheta - \pi Q\log q_\vartheta| \to 0.$$

Hence

$$\begin{aligned} \pi_{nh}Q_{nh}\log q_{k(Q_{nh})} &\to \pi Q\log q_{k(Q)}, \\ \pi_{nh}Q_{nh}\log q_{k(Q_{nh})} - \pi Q\log q_{k(Q_{nh})} &\to 0. \end{aligned}$$

Therefore,

$$\pi Q\log q_{k(Q_{nh})} \to \pi Q\log q_{k(Q)}. \tag{3.5}$$

Since Θ is compact, we may choose a subsequence on which $k(Q_{nh})$ converges. Let k_0 denote the limit. On the subsequence,

$$\pi Q\log q_{k(Q_{nh})} \to \pi Q\log q_{k_0}.$$

Comparing with (3.5), we obtain $\pi Q\log q_{k_0} = \pi Q\log q_{k(Q)}$. Since $k(Q)$ is unique, we have $k_0 = k(Q)$, and the proof is finished. □

Lemma 5. *Let Assumption 2 hold. Let* $\log q_\vartheta$ *be differentiable in the sense of* (2.5). *Then for all* n *and* $h\in H$ *with* $k(Q_{nh})$ *in the interior of* Θ, *the function* $\pi_{nh}Q_{nh}\log q_\vartheta$ *is differentiable at* $\vartheta = k(Q_{nh})$ *with derivative* $\pi_{nh}Q_{nh}\ell'_{k(Q_{nh})} = 0$.

Proof. Write

$$t^{-1}\pi_{nh}Q_{nh}(\log q_{\vartheta+t} - \log q_\vartheta) = \pi_{nh}Q_{nh}(\ell'_\vartheta + r_{\vartheta+t}).$$

By Lemma 3 and assumption (2.5) we have $\pi_{nh}Q_{nh}r_{\vartheta+t} \to 0$ as $t\to 0$. Hence the function $\pi_{nh}Q_{nh}\log q_\vartheta$ is differentiable in ϑ with derivative $\pi_{nh}Q_{nh}\ell'_\vartheta$. Since $k(Q_{nh})$ maximizes $\pi_{nh}Q_{nh}\log q_\vartheta$ by Lemma 4 and lies in the interior of Θ by hypothesis, the assertion follows.

Lemma 6. *Under Assumptions 2 and 3, the functional* k *is differentiable at* Q *in the sense of* (2.3) *with canonical gradient* g *defined in* (2.9).

Proof. By Lemma 4 we have $k(Q_{nh}) \to k(Q)$. Since $k = k(Q)$ is in the interior of Θ by Assumption 3, so is $k(Q_{nh})$ for n sufficiently large. With Lemma 5 we obtain

$$\begin{aligned} 0 &= \pi_{nh}Q_{nh}\ell'_{k(Q_{nh})} \\ &= \pi_{nh}Q_{nh}\ell'_k + (k(Q_{nh}) - k)\,\pi_{nh}Q_{nh}(\ell''_k + s_{k(Q_{nh})}). \end{aligned} \tag{3.6}$$

By Lemma 3,

$$\pi_{nh}Q_{nh}\ell''_k \to \pi Q\ell''_k = -D.$$

By Lemma 3 and relation (2.6),

$$\pi_{nh} Q_{nh} s_{k(Q_{nh})} \to 0.$$

Solve (3.6) for $k(Q_{nh})$ to obtain

$$k(Q_{nh}) - k(Q) = D^{-1} \pi_{nh} Q_{nh} \ell'_k \left(1 + o(1)\right). \tag{3.7}$$

Since $\pi Q \ell'_k = 0$ by Lemma 5, we may use Lemma 3 with $f = \ell'_k$ to obtain

$$\begin{aligned} n^{1/2} \pi_{nh} Q_{nh} \ell'_k &= n^{1/2} (\pi_{nh} Q_{nh} - \pi Q) \ell'_k \\ &\to \pi Q h (\ell'_k - Q \ell'_k + A Q \ell'_k) = \pi Q h g. \end{aligned}$$

Relation (3.7) then implies

$$n^{1/2} \left(k(Q_{nh}) - k(Q)\right) \to D^{-1} \pi Q h g.$$

This is the assertion. □

References

Andrews, D. W. K. and Pollard, D. (1994). An introduction to functional central limit theorems for dependent stochastic processes. *Internat. Statist. Rev.* 62, 119–132.

Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *Ann. Statist.* 5, 445–463.

Durrett, R. (1991). *Probability: Theory and Examples.* Wadsworth, Belmont.

Gordin, M. I. (1969). The central limit theorem for stationary processes. *Soviet Math. Dokl.* 10, 1174–1176.

Gordin, M. I. and Lifšic, B. A. (1978). The central limit theorem for stationary Markov processes. *Soviet Math. Dokl.* 19, 392–394.

Greenwood, P. E. and Wefelmeyer, W. (1990). Efficiency of estimators for partially specified filtered models. *Stochastic Process. Appl.* 36, 353–370.

Greenwood, P. E. and Wefelmeyer, W. (1995). Efficiency of empirical estimators for Markov chains. *Ann. Statist.* 23, 132–143.

Hosoya, Y. (1989). The bracketing condition for limit theorems on stationary linear processes. *Ann. Statist.* 17, 401–418.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symp. Math. Statist. Probab.* 1, 221–233.

Kartashov, N.V. (1985a). Criteria for uniform ergodicity and strong stability of Markov chains with a common phase space. *Theory Probab. Math. Statist.* 30, 71–89.

Kartashov, N.V. (1985b). Inequalities in theorems of ergodicity and stability for Markov chains with common phase space. I. *Theory Probab. Appl.* 30, 247–259.

Kutoyants, Yu. A. (1988). On an identification problem for dynamical systems with small noise. *Izv. Akad. Nauk Armyan. SSR* 23, 270–285.

McKeague, I. W. (1984). Estimation for diffusion processes under misspecified models. *J. Appl. Probab.* 21, 511–520.

Ogata, Y. (1980). Maximum likelihood estimates of incorrect Markov models for time series and the derivation of AIC. *J. Appl. Probab.* 17, 59–72.

Penev, S. (1991). Efficient estimation of the stationary distribution for exponentially ergodic Markov chains. *J. Statist. Plann. Inference* 27, 105–123.

Pollard, D. (1985). New ways to prove central limit theorems. *Econometric Theory* 1, 295–314.

Yang, S. (1991). Minimum Hellinger distance estimation of parameters in the random censorship model. *Ann. Statist.* 19, 579–602.

Ying, Z. (1992). Minimum Hellinger-type distance estimation for censored data. *Ann. Statist.* 20, 1361–1390.