Stochastics: An International Journal of Probability and Stochastic Processes,
Vol. 00, No. 00, December 2006, 1–13

Optimality of estimators for misspecified semi-Markov models

URSULA U. MÜLLER†, ANTON SCHICK[1]‡ and WOLFGANG WEFELMEYER*§

† Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA
‡ Department of Mathematical Sciences, Binghamton University, Binghamton, NY 13902-6000, USA
§ Mathematisches Institut, Universität zu Köln, Weyertal 86-90, 50931 Köln, Germany

(*v3.1 released December 2006*)

Suppose we observe a geometrically ergodic semi-Markov process and have a parametric model for the transition distribution of the embedded Markov chain, for the conditional distribution of the inter-arrival times, or for both. The first two models for the process are semiparametric, and the parameters can be estimated by conditional maximum likelihood estimators. The third model for the process is parametric, and the parameter can be estimated by an unconditional maximum likelihood estimator. We determine heuristically the asymptotic distributions of these estimators and show that they are asymptotically efficient. If the parametric models are not correct, the (conditional) maximum likelihood estimators estimate the parameter that maximizes the Kullback–Leibler information. We show that they remain asymptotically efficient in a nonparametric sense.

Keywords: Hellinger differentiability, local asymptotic normality, asymptotically linear estimator, Markov renewal process.

AMS Subject Classification: Primary: 62M09; secondary: 62F12, 62G20.

1. Introduction

For i.i.d. observations, Daniels [6] and Huber [20] show that the maximum likelihood estimator of a misspecified parametric model estimates the parameter that maximizes the Kullback–Leibler (KL) information, and determine its asymptotic distribution. Weaker conditions are given by Pollard [33]. For applications see also White [35], Müller [30], and Doksum, Ozeki, Kim and Neto [7]. Analogous results are obtained for parametric Markov chain models by Ogata [31], for parametric time series by Hosoya [19] and by Andrews and Pollard [1], and for parametric diffusion models by McKeague [28] and Kutoyants [25]. We refer also to the monograph of Kutoyants [26]. Applications to time series models in econometrics are studied by White [36] and Sin and White [34], and in the monograph of White [37].

Greenwood and Wefelmeyer [15] prove that the maximum likelihood estimator of a misspecified parametric Markov chain model is efficient in a nonparametric sense. Related efficiency results for misspecified parametric time series are in Dahlhaus and Wefelmeyer [5]. Here we outline corresponding results for semi-Markov processes. We consider both parametric and semiparametric misspecified models. The arguments are heuristic; sufficient regularity conditions can be obtained as in the above references.

Suppose we observe a semi-Markov process Z_t, $t \geq 0$, with values in an arbitrary measurable space E, on a time interval $0 \leq t \leq n$. Let $(X_0, T_0), (X_1, T_1), \ldots$ denote the embedded Markov renewal process. Its transition distribution factors as

$$S(x, dy, du) = Q \otimes R(x, dy, du) = Q(x, dy)R(x, y, du),$$

[1]Supported in part by NSF Grant DMS0405791
*Corresponding author. Email: wefelm@math.uni-koeln.de

Stochastics: An International Journal of Probability and Stochastic Processes
ISSN: 1744-2508 print/ISSN 1744-2516 online
http://www.tandf.co.uk/journals
DOI: 10.1080/1744250YYxxxxxxxx

where $Q(x, dy)$ is the transition distribution of the embedded Markov chain $X_0, X_1, \ldots$, and $R(x, y, du)$ is the conditional distribution of the inter-arrival time $U_j = T_j - T_{j-1}$ given $X_{j-1} = x$ and $X_j = y$.

We assume that the embedded Markov chain is stationary. We write $P_1(dx)$, $P_2(dx, dy)$ and $P_3(dx, dy, du)$ for the stationary laws of X_{j-1}, (X_{j-1}, X_j) and (X_{j-1}, X_j, U_j), respectively. Of course, $P_2 = P_1 \otimes Q$ and $P_3 = P_2 \otimes R = P_1 \otimes Q \otimes R$. Set $N = \max\{j : T_j \le n\}$. We note that studying a semi-Markov process is equivalent to studying the embedded Markov renewal process. The latter is a Markov chain. Observing the semi-Markov process up to time n is equivalent to observing the embedded Markov renewal process up to the random time N.

Natural estimators for P_1, P_2 and P_3 are the empirical distributions

$$\mathbb{P}_1 = \frac{1}{N}\sum_{j=1}^{N} \delta_{X_{j-1}}, \qquad \mathbb{P}_2 = \frac{1}{N}\sum_{j=1}^{N} \delta_{(X_{j-1},X_j)}, \qquad \mathbb{P}_3 = \frac{1}{N}\sum_{j=1}^{N} \delta_{(X_{j-1},X_j,U_j)}.$$

where δ_x denotes the Dirac measure at a point x.

Let Θ be an open subset of $\mathbb{R}^d$. We consider the following three models for the semi-Markov process. In *Model Q* we assume a parametric form $Q = Q_\vartheta$, $\vartheta \in \Theta$, of the transition distribution of the embedded Markov chain. These models are also considered in Greenwood, Müller and Wefelmeyer [11]. In *Model R* we assume a parametric form $R = R_\vartheta$, $\vartheta \in \Theta$, of the conditional distribution of the inter-arrival times. In *Model S* we assume parametric forms $Q = Q_\vartheta$ and $R = R_\vartheta$, $\vartheta \in \Theta$, for both. Of course, the last model covers the case that Q and R carry different parameters. We assume that $Q_\vartheta(x, dy)$ has a density $q_\vartheta(x, y)$ with respect to some dominating measure $\mu(dy)$, and $R_\vartheta(x, y, du)$ has a density $r_\vartheta(x, y, u)$ with respect to some dominating measure $\nu(du)$.

If Model Q holds, then the transition distribution of the semi-Markov process is semiparametric, $S = Q_\vartheta \times R$, with R an infinite-dimensional nuisance parameter. A natural estimator of ϑ is the *partial maximum likelihood estimator* $\hat\vartheta_Q$, which maximizes

$$\mathbb{P}_2[\log q_\vartheta] = \frac{1}{N}\sum_{j=1}^{N} \log q_\vartheta(X_{j-1}, X_j).$$

Suppose that Model Q is misspecified, and that the true transition distribution of the embedded Markov chain is Q. Then $\mathbb{P}_2[\log q_\vartheta]$ is an empirical version of the *KL information* $P_2[\log q_\vartheta]$. Let $K_Q(P_2)$ denote the parameter that maximizes $P_2[\log q_\vartheta]$. We call K_Q a *KL functional*. Note that the partial maximum likelihood estimator is the empirical version of the KL functional, $\hat\vartheta_Q = K_Q(\mathbb{P}_2)$. Since Model Q is misspecified, the semi-Markov model is nonparametric. The empirical distribution $\mathbb{P}_2$ is efficient for P_2 in a certain sense. If the KL functional is smooth, i.e. compactly differentiable in an appropriate sense, it follows that $\hat\vartheta_Q = K_Q(\mathbb{P}_2)$ is efficient for $K_Q(P_2)$. We will not use this approach in this paper. Instead we derive, in Section 3, a stochastic expansion of $\hat\vartheta_Q$, and determine its influence function. We also show that the KL functional K_Q is pathwise differentiable, and determine its canonical gradient. To keep the exposition simple, we do not give regularity conditions for these results. They can be adapted e.g. from those of Greenwood and Wefelmeyer [15]. It turns out that the canonical gradient equals the influence function of $\hat\vartheta_Q$. By the characterisation of efficient estimators in Section 2, this shows that $\hat\vartheta_Q$ is efficient in the nonparametric semi-Markov model. We also show that $\hat\vartheta_Q$ remains efficient when Model Q is true. The advantage of our approach is that we do not need to check compact differentiability of K_Q and a corresponding efficiency property of $\mathbb{P}_2$.

The other two models are treated analogously. If Model R holds, then the transition distribution of the semi-Markov process is semiparametric, $S = Q \otimes R_\vartheta$, with Q an infinite-dimensional nuisance parameter. A natural estimator of ϑ is the *partial maximum likelihood estimator* $\hat\vartheta_R$, which maximizes

$$\mathbb{P}_3[\log r_\vartheta] = \frac{1}{N}\sum_{j=1}^{N} \log r_\vartheta(X_{j-1}, X_j, U_j).$$

Suppose that Model Q is misspecified, and that the true conditional distribution of the inter-arrival times is R. Then $\mathbb{P}_3[\log r_\vartheta]$ is an empirical version of $P_3[\log r_\vartheta]$. Again we call the latter *KL information*. We denote by $K_R(P_3)$ the parameter that maximizes $P_3[\log r_\vartheta]$, and we call K_R a *KL functional*. Then $\hat{\vartheta}_R = K_R(\mathbb{P}_3)$. In Section 4 we derive a stochastic expansion of $\hat{\vartheta}_R$ and the canonical gradient of K_R and show that $\hat{\vartheta}_R$ is efficient in the nonparametric semi-Markov model. We also show that $\hat{\vartheta}_R$ remains efficient when Model R is true.

If Model S holds, then the transition distribution of the semi-Markov process is parametric, $S_\vartheta = Q_\vartheta \otimes R_\vartheta$. Set

$$s_\vartheta(x, y, u) = q_\vartheta(x, y) r_\vartheta(x, y, u).$$

A natural estimator of ϑ is the *maximum likelihood estimator* $\hat{\vartheta}_S$, which maximizes

$$\mathbb{P}_3[\log s_\vartheta] = \mathbb{P}_2[\log q_\vartheta] + \mathbb{P}_3[\log r_\vartheta] = \frac{1}{N}\sum_{j=1}^{N} \log q_\vartheta(X_{j-1}, X_j) + \frac{1}{N}\sum_{j=1}^{N} \log r_\vartheta(X_{j-1}, X_j, U_j).$$

Suppose that Model Q is misspecified, and that the true transition distribution of the embedded Markov renewal process is $S = Q \otimes R$. Then $\mathbb{P}_3[\log s_\vartheta]$ is an empirical version of $P_3[\log s_\vartheta]$. Again we call the latter *KL information*. We denote by $K_S(P_3)$ the parameter that maximizes $P_3[\log s_\vartheta]$, and we call K_S a *KL functional*. Then $\hat{\vartheta}_S = K_S(\mathbb{P}_3)$. In Section 5 we derive a stochastic expansion of $\hat{\vartheta}_S$ and the canonical gradient of K_S and show that $\hat{\vartheta}_S$ is efficient in the nonparametric semi-Markov model. We also show that $\hat{\vartheta}_S$ remains efficient when Model S is true. Section 6 contains some additional comments.

2. Characterization of efficient estimators

We assume that the embedded Markov chain is positive Harris recurrent and geometrically ergodic in $L_2(P_2)$. We make the usual assumption that the conditional distribution of the inter-arrival times does not charge zero. We also assume that the mean inter-arrival time $m = EU_j$ is finite. Then

$$n/N \to m \quad a.s. \tag{1}$$

For a function $f \in L_2(P_3)$ we have the strong law of large numbers

$$\frac{1}{N}\sum_{j=1}^{N} f(X_{j-1}, X_j, U_j) \to P_3[f] \quad \text{a.s.} \tag{2}$$

For a function $f \in L_2(P_3)$ with $Sf = 0$ we have the martingale central limit theorem

$$n^{-1/2}\sum_{j=1}^{N} f(X_{j-1}, X_j, U_j) \Rightarrow m^{-1/2}(P_3[f^2])^{1/2} Y, \tag{3}$$

where Y denotes a standard normal random variable.

In order to characterize efficient estimators for functionals of semi-Markov models, we consider a family Q_δ, $\delta \in \Delta$, of transition distributions of the embedded Markov chain, and a family R_δ, $\delta \in \Delta$, of conditional distributions of the inter-arrival time. Here Δ is a possibly infinite-dimensional set, the *parameter space*. We fix $\delta \in \Delta$ and set $Q = Q_\delta$, $R = R_\delta$ and

$$V = \{v \in L_2(P_2) : Qv = 0\}, \qquad W = \{w \in L_2(P_3) : Rw = 0\}.$$

Note that V and W can be viewed as orthogonal subspaces of $L_2(P_3)$. We assume that the parametrization is smooth in the following sense. There is a linear space K, the *tangent space* of Δ, and a bounded linear operator $D = (D_Q, D_R) : K \to V \times W$, and for each $k \in K$ there is a sequence δ_{nk} in Δ such that $Q_{nk} = Q_{\delta_{nk}}$ is Hellinger differentiable at Q with derivative $D_Q k \in V$,

$$P_1\Big[\int \Big(dQ_{nk}^{1/2} - dQ^{1/2} - \frac{1}{2}n^{-1/2}D_Q k\, dQ^{1/2}\Big)^2\Big] \to 0,$$

and $R_{nk} = R_{\delta_{nk}}$ is Hellinger differentiable at R with derivative $D_R k \in W$,

$$P_2\Big[\int \Big(dR_{nk}^{1/2} - dR^{1/2} - \frac{1}{2}n^{-1/2}D_R k\, dR^{1/2}\Big)^2\Big] \to 0.$$

Now write M_n for the distribution of Z_t, $0 \le t \le n$, if Q and R are in effect, and M_{nk} if Q_{nk} and R_{nk} are. By Taylor expansion and (2) and (3), we obtain *local asymptotic normality*:

$$\log \frac{dM_{nk}}{dM_n} = n^{-1/2}\sum_{j=1}^{N}(D_Q k(X_{j-1}, X_j) + D_R k(X_{j-1}, X_j, U_j)) - m^{-1}(P_2[D_Q^2 k] + P_3[D_R^2 k]) + o_p(1) \quad (4)$$

and

$$n^{-1/2}\sum_{j=1}^{N}(D_Q k(X_{j-1}, X_j) + D_R k(X_{j-1}, X_j, U_j)) \Rightarrow m^{-1/2}(P_2[D_Q^2 k] + P_3[D_R^2 k])^{1/2} Y. \quad (5)$$

For Markov chains, different proofs are in Penev [32], Bickel [2] and Greenwood and Wefelmeyer [13]; see also Bickel and Kwon [4]. For Markov step processes see Höpfner, Jacod and Ladelli [18] and Höpfner [16, 17]. A proof for nonparametric semi-Markov models is in Greenwood and Wefelmeyer [14].

We want to estimate a d-dimensional functional $\varphi : \Delta \to \mathbb{R}^d$ of the parameter δ. We call φ *differentiable* at δ with *gradient* (v_φ, w_φ) if $v_\varphi \in V^d$, $w_\varphi \in W^d$, and

$$n^{1/2}(\varphi(\delta_{nk}) - \varphi(\delta)) \to m^{-1}(P_2[v_\varphi D_Q k] + P_3[w_\varphi D_R k]), \quad k \in K. \quad (6)$$

The *canonical gradient* $(v_\varphi^*, w_\varphi^*)$ of φ is the componentwise projection of (v_φ, w_φ) onto the closure of $(DK)^d$ in $(L_2(P_3))^d$. If DK is closed in $L_2(P_3)$, we can write $(v_\varphi^*, w_\varphi^*) = (D_Q k_\varphi, D_R k_\varphi)$ for some $k_\varphi \in K$. This will be the case in Sections 3–5.

An estimator $\hat\varphi$ is called *regular* for φ at δ with *limit* L if L is a d-dimensional random vector such that

$$n^{1/2}(\hat\varphi - \varphi(\delta_{nk})) \Rightarrow L \quad \text{under } M_{nk}, \quad k \in K.$$

The convolution theorem says that

$$L = A + m^{-1/2}(P_2[v_\varphi^* v_\varphi^{*\top}] + P_3[w_\varphi^* w_\varphi^{*\top}])^{1/2} Y_d,$$

with Y_d a d-dimensional standard normal random vector, and A a d-dimensional random vector independent of Y_d. This justifies calling $\hat\varphi$ *efficient* for φ at δ if $n^{1/2}(\hat\varphi - \varphi(\delta))$ is asymptotically normal under M_n with covariance matrix $m^{-1}(P_2[v_\varphi^* v_\varphi^{*\top}] + P_3[w_\varphi^* w_\varphi^{*\top}])$.

An estimator $\hat\varphi$ is called *asymptotically linear* for φ at δ with *influence function* (a, b) if $a \in V^d$, $b \in W^d$, and

$$n^{1/2}(\hat\varphi - \varphi(\delta)) = n^{-1/2}\sum_{j=1}^{N}(a(X_{j-1}, X_j) + b(X_{j-1}, X_j, U_j)) + o_p(1).$$

We have the following characterization. An estimator $\hat{\varphi}$ is regular and efficient for φ at δ if and only if it is asymptotically linear with influence function equal to the canonical gradient,

$$n^{1/2}(\hat{\varphi} - \varphi(\delta)) = n^{-1/2} \sum_{j=1}^{N} (v_\varphi^*(X_{j-1}, X_j) + w_\varphi^*(X_{j-1}, X_j, U_j)) + o_p(1).$$

For proofs of the convolution theorem and the characterization we refer to Bickel, Klaassen, Ritov and Wellner [3].

To prove asymptotic linearity of estimators in misspecified models, we need the following *martingale approximation*. Set $L_{2,0}(P_2) = \{f \in L_2(P_2) : P_2[f] = 0\}$. The *potential* G of the embedded Markov chain is defined by

$$Gf = \sum_{i=0}^{\infty} Q^i f, \quad f \in L_{2,0}(P_2).$$

For $f \in L_2(P_2)$ set

$$Af(x,y) = G(f - P_2[f])(y) - QG(f - P_2[f])(x) = \sum_{i=0}^{\infty} (Q^i f(y) - Q^{i+1} f(x)).$$

Then $QAf = 0$ and

$$P_2[(Af)^2] = P_2[f^2] - (P_2[f])^2 + 2 \sum_{i=1}^{\infty} P_2[(f - P_2[f])Q^i f].$$

Let $f \in L_2(P_3)$ and set $f_0 = f - Rf$. Then we obtain the stochastic expansion

$$n^{-1/2} \sum_{j=1}^{N} (f(X_{j-1}, X_j, U_j) - P_3[f]) = n^{-1/2} \sum_{j=1}^{N} (ARf(X_{j-1}, X_j) + f_0(X_{j-1}, X_j, U_j)) + o_p(1). \tag{7}$$

Note that $QARf = 0$ and $Sf_0 = 0$. Hence $ARf(X_{j-1}, X_j)$ and $f_0(X_{j-1}, X_j, U_j)$ are orthogonal martingale increments. For discrete-time processes, the martingale approximation (7) is due to Gordin [9] and Gordin and Lifšic [10]. It was discovered independently by Maigret [27], Dürr and Goldstein [8] and Greenwood and Wefelmeyer [13]. See also Section 17.4 in the monograph of Meyn and Tweedie [29]. The martingale approximation (7) and the martingale central limit theorem (3) imply that

$$n^{-1/2} \sum_{j=1}^{N} (f(X_{j-1}, X_j, U_j) - P_3[f]) \Rightarrow m^{-1/2} (P_2[(ARf)^2] + P_3[(f - Rf)^2])^{1/2} Y.$$

To calculate canonical gradients of functionals in misspecified models, we need the following *perturbation expansion*, due to Kartashov [21–23],

$$n^{1/2}(P_{2nk}[f] - P_2[f]) \to P_2[D_Q k \cdot Af], \quad k \in K. \tag{8}$$

Here P_{2nk} denotes the distribution of (X_{j-1}, X_j) if Q_{nk} is in effect. This pathwise version of the perturbation expansion suffices for our purposes. Greenwood and Wefelmeyer [13] show that it follows also from the martingale approximation (7).

3. Model Q

In Model Q we assume a parametric model q_ϑ, $\vartheta \in \Theta \subset \mathbb{R}^d$, for the μ-density of the transition distribution of the embedded Markov chain, and consider the conditional inter-arrival time distribution as unknown. Suppose the model is misspecified, and the true transition distribution is Q. Then the KL functional $K_Q(P_2)$ maximizes $P_2[\log q_\vartheta]$, and the partial maximum likelihood estimator $\hat\vartheta_Q$ maximizes $\mathbb{P}_2[\log q_\vartheta]$. Write

$$\chi_\vartheta(x,y) = \partial_\vartheta \log q_\vartheta(x,y)$$

for the d-dimensional vector of partial derivatives of $\log q_\vartheta(x,y)$. Then $K_Q(P_2)$ solves $P_2[\chi_\vartheta] = 0$, and $\hat\vartheta_Q$ solves $\mathbb{P}_2[\chi_\vartheta] = 0$. Heuristically, by Taylor expansion,

$$\begin{aligned} 0 = \mathbb{P}_2[\chi_{\hat\vartheta_Q}] &= \frac{1}{N}\sum_{j=1}^N \chi_{\hat\vartheta_Q}(X_{j-1},X_j) \\ &= \frac{1}{N}\sum_{j=1}^N \chi_{K_Q(P_2)}(X_{j-1},X_j) + \frac{1}{N}\sum_{j=1}^N \dot\chi_{K_Q(P_2)}(X_{j-1},X_j)(\hat\vartheta_Q - K_Q(P_2)) + o_p(n^{-1/2}). \end{aligned} \tag{9}$$

Here $\dot\chi_\vartheta(x,y)$ is the $d\times d$ matrix of partial derivatives of $\chi_\vartheta(x,y)$. With (1) and (2) we obtain

$$n^{1/2}(\hat\vartheta_Q - K_Q(P_2)) = -m(P_2[\dot\chi_{K_Q(P_2)}])^{-1} n^{-1/2}\sum_{j=1}^N \chi_{K_Q(P_2)}(X_{j-1},X_j) + o_p(1). \tag{10}$$

If Model Q is correctly specified and $Q = Q_\vartheta$, then $K_Q(P_2) = \vartheta$. We also have the following relations, which are well-known in the i.i.d. case,

$$0 = \partial_\vartheta Q_\vartheta(\cdot,E) = Q_\vartheta\chi_\vartheta, \qquad 0 = \partial_\vartheta Q_\vartheta\chi_\vartheta = Q_\vartheta\chi_\vartheta\chi_\vartheta^\top + Q_\vartheta\dot\chi_\vartheta.$$

In particular, the partial Fisher information matrix for Model Q is $I_\vartheta = -P_2[\dot\chi_\vartheta] = P_2[\chi_\vartheta\chi_\vartheta^\top]$. Hence, for the correctly specified model, the partial maximum likelihood estimator $\hat\vartheta_Q$ has the stochastic expansion

$$n^{1/2}(\hat\vartheta_Q - \vartheta) = mI_\vartheta^{-1} n^{-1/2}\sum_{j=1}^N \chi_\vartheta(X_{j-1},X_j) + o_p(1).$$

This means that $\hat\vartheta_Q$ is asymptotically linear with influence function $mI_\vartheta^{-1}(\chi_\vartheta,0)$, and $n^{1/2}(\hat\vartheta_Q - \vartheta)$ is asymptotically normal with covariance matrix mI_ϑ^{-1}.

If the model is misspecified, then $\chi_{K_Q(P_2)}$ is not in V^d. We apply the martingale approximation (7) to (10) and see that $\hat\vartheta_Q$ is asymptotically linear with influence function $-m(P_2[\dot\chi_{K_Q(P_2)}])^{-1}(A\chi_{K_Q(P_2)},0)$. Hence $n^{1/2}(\hat\vartheta_Q - K_Q(P_2))$ is asymptotically normal with covariance matrix

$$m(P_2[\dot\chi_{K_Q(P_2)}])^{-1} P_2[A\chi_{K_Q(P_2)}A^\top\chi_{K_Q(P_2)}](P_2[\dot\chi_{K_Q(P_2)}])^{-1}.$$

Let us now prove efficiency of $\hat\vartheta_Q$, first for the correctly specified model. For $c\in\mathbb{R}^d$ set $\vartheta_{nc} = \vartheta + n^{-1/2}c$. Assume that $q_{nc} = q_{\vartheta_{nc}}$ is Hellinger differentiable at ϑ,

$$\int\int \Big(q_{nc}^{1/2}(x,y) - q_\vartheta^{1/2}(x,y) - \frac{1}{2}n^{-1/2}c^\top\chi_\vartheta(x,y)q_\vartheta^{1/2}(x,y)\Big)^2 \mu(dy)P_1(dx) \to 0. \tag{11}$$

Let $\mathcal{R}$ denote the set of all conditional inter-arrival distributions. For $w \in W$ choose a sequence R_{nw} in $\mathcal{R}$ that is Hellinger differentiable at R,

$$P_2\Big[\int \Big(dR_{nw}^{1/2} - dR^{1/2} - \frac{1}{2}n^{-1/2}w\,dR^{1/2}\Big)^2\Big] \to 0. \tag{12}$$

Then the assumptions of Section 2 hold with $\Delta = \Theta \times \mathcal{R}$, $K = \mathbb{R}^d \times W$, $D_Q(c,w) = c^\top \chi_\vartheta$, $D_R(c,w) = w$. The functional to be estimated is $\varphi(\vartheta, R) = \vartheta$. By orthogonality of V and W, its canonical gradient is obtained from (6) as $(c_\vartheta^\top \chi_\vartheta, 0)$ with $d \times d$ matrix c_ϑ determined by

$$c = m^{-1}c_\vartheta^\top P_2[\chi_\vartheta \chi_\vartheta^\top]c = m^{-1}c_\vartheta^\top I_\vartheta c, \quad c \in \mathbb{R},$$

i.e. $c_\vartheta = mI_\vartheta^{-1}$. Hence the canonical gradient of ϑ is $mI_\vartheta^{-1}(\chi_\vartheta, 0)$ and equals the influence function of $\hat{\vartheta}$, which is therefore efficient for the correctly specified model.

Suppose now that the model is misspecified, and let $\mathcal{Q}$ be the set of all transition distributions of the embedded Markov chain. Let Q denote the true transition distribution. For $v \in V$ choose a sequence Q_{nv} in $\mathcal{Q}$ that is Hellinger differentiable at Q,

$$P_1\Big[\int \Big(dQ_{nv}^{1/2} - dQ^{1/2} - \frac{1}{2}n^{-1/2}v\,dQ^{1/2}\Big)^2\Big] \to 0. \tag{13}$$

Then the assumptions of Section 2 hold with $\Delta = \mathcal{Q} \times \mathcal{R}$, $K = V \times W$, $D_Q(v,w) = v$, $D_R(v,w) = w$. The functional to be estimated is $\varphi(Q,R) = K_Q(P_2)$. Heuristically,

$$0 = P_{2nv}[\chi_{K_Q(P_{2nv})}] = P_{2nv}[\chi_{K_Q(P_2)}] + P_{2nv}[\dot{\chi}_{K_Q(P_2)}](K_Q(P_{2nv}) - K_Q(P_2)) + o_p(n^{-1/2}).$$

With $P_{2nv}[\dot{\chi}_{K_Q(P_2)}] \to P_2[\dot{\chi}_{K_Q(P_2)}]$ we obtain

$$K_Q(P_{2nv}) - K_Q(P_2) = -(P_2[\dot{\chi}_{K_Q(P_2)}])^{-1}P_{2nv}[\chi_{K_Q(P_2)}] + o_p(n^{-1/2}).$$

The perturbation expansion (8) yields

$$n^{1/2}P_{2nv}[\chi_{K_Q(P_2)}] = n^{1/2}(P_{2nv} - P_2)[\chi_{K_Q(P_2)}] \to P_2[vA\chi_{K_Q(P_2)}]. \tag{14}$$

Hence

$$n^{1/2}(K_Q(P_{2nv}) - K_Q(P_2)) \to -(P_2[\dot{\chi}_{K_Q(P_2)}])^{-1}P_2[vA\chi_{K_Q(P_2)}], \quad v \in V,$$

and the canonical gradient of K_Q is obtained from (6) as $-m(P_2[\dot{\chi}_{K_Q(P_2)}])^{-1}(A\chi_{K_Q(P_2)}, 0)$ and equals the influence function of $\hat{\vartheta}_Q$, which is therefore efficient for the misspecified model.

4. Model R

Model R is completely analogous to Model Q, with interchanged roles of the transition distribution Q of the embedded Markov chain, and the conditional inter-arrival time distribution R. Specifically, in Model R we assume a parametric model r_ϑ, $\vartheta \in \Theta \subset \mathbb{R}^d$, for the ν-density of the conditional inter-arrival time, and consider the transition distribution of the embedded Markov chain as unknown. Suppose the model is misspecified, and the true conditional inter-arrival time distribution is R. Then the KL functional $K_R(P_3)$ maximizes $P_3[\log r_\vartheta]$, and the partial maximum likelihood estimator $\hat{\vartheta}_Q$ maximizes $\mathbb{P}_3[\log r_\vartheta]$. Write

$$\varrho_\vartheta(x,y,u) = \partial_\vartheta \log r_\vartheta(x,y,u)$$

for the d-dimensional vector of partial derivatives of $\log r_\vartheta(x,y,u)$. Then $K_R(P_3)$ solves $P_3[\varrho_\vartheta]=0$, and $\hat\vartheta_R$ solves $\mathbb{P}_3[\varrho_\vartheta]=0$. Heuristically, by Taylor expansion,

$$\begin{aligned} 0=\mathbb{P}_3[\varrho_{\hat\vartheta_R}] &= \frac{1}{N}\sum_{j=1}^N \varrho_{\hat\vartheta_R}(X_{j-1},X_j,U_j) \\ &= \frac{1}{N}\sum_{j=1}^N \varrho_{K_R(P_3)}(X_{j-1},X_j,U_j) + \frac{1}{N}\sum_{j=1}^N \dot\varrho_{K_R(P_3)}(X_{j-1},X_j,U_j)(\hat\vartheta_R-K_R(P_3))+o_p(n^{-1/2}). \end{aligned} \tag{15}$$

Here $\dot\varrho_\vartheta(x,y,u)$ is the $d\times d$ matrix of partial derivatives of $\varrho_\vartheta(x,y,u)$. With (1) and (2) we obtain

$$n^{1/2}(\hat\vartheta_R-K_R(P_3)) = -m(P_3[\dot\varrho_{K_R(P_3)}])^{-1}n^{-1/2}\sum_{j=1}^N \varrho_{K_R(P_3)}(X_{j-1},X_j,U_j)+o_p(1). \tag{16}$$

If Model R is correctly specified and $R=R_\vartheta$, then $K_R(P_3)=\vartheta$. We also have the following relations,

$$0=\partial_\vartheta R_\vartheta(\cdot,\cdot,\mathbb{R})=R_\vartheta\varrho_\vartheta, \qquad 0=\partial_\vartheta R_\vartheta\varrho_\vartheta = R_\vartheta\varrho_\vartheta\varrho_\vartheta^\top + R_\vartheta\dot\varrho_\vartheta.$$

In particular, the partial Fisher information matrix for Model R is $J_\vartheta=-P_3[\dot\varrho_\vartheta]=P_3[\varrho_\vartheta\varrho_\vartheta^\top]$. Hence, for the correctly specified model, the partial maximum likelihood estimator $\hat\vartheta_R$ has the stochastic expansion

$$n^{1/2}(\hat\vartheta_R-\vartheta)=mJ_\vartheta^{-1}n^{-1/2}\sum_{j=1}^N\varrho_\vartheta(X_{j-1},X_j,U_j)+o_p(1).$$

This means that $\hat\vartheta_R$ is asymptotically linear with influence function $mJ_\vartheta^{-1}(0,\varrho_\vartheta)$, and $n^{1/2}(\hat\vartheta_R-\vartheta)$ is asymptotically normal with covariance matrix mJ_ϑ^{-1}.

If the model is misspecified, then $\varrho_{K_R(P_3)}$ is not in W^d. We apply the martingale approximation (7) to (16) and see that $\hat\vartheta_R$ is asymptotically linear with influence function

$$-m(P_3[\dot\varrho_{K_R(P_3)}])^{-1}(AR\varrho_{K_R(P_3)},\varrho_{K_R(P_3)}-R\varrho_{K_R(P_3)}).$$

Hence $n^{1/2}(\hat\vartheta_R-K_R(P_3))$ is asymptotically normal with covariance matrix

$$m(P_3[\dot\varrho_{K_R(P_3)}])^{-1}\Sigma_R(P_3[\dot\varrho_{K_R(P_3)}])^{-1},$$

where

$$\Sigma_R=P_2[AR\varrho_{K_R(P_3)}A^\top R\varrho_{K_R(P_3)}]+P_3[(\varrho_{K_R(P_3)}-R\varrho_{K_R(P_3)})(\varrho_{K_R(P_3)}-R\varrho_{K_R(P_3)})^\top].$$

Let us now prove efficiency of $\hat\vartheta_R$, first for the correctly specified model. For $c\in\mathbb{R}^d$ set $\vartheta_{nc}=\vartheta+n^{-1/2}c$. Assume that $r_{nc}=r_{\vartheta_{nc}}$ is Hellinger differentiable at ϑ,

$$\int\int\Big(r_{nc}^{1/2}(x,y,u)-r_\vartheta^{1/2}(x,y,u)-\frac{1}{2}n^{-1/2}c^\top\varrho_\vartheta(x,y,u)r_\vartheta^{1/2}(x,y,u)\Big)^2\nu(du)P_2(d(x,y))\to 0. \tag{17}$$

Let $\mathcal{Q}$ denote the set of all transition distributions of the embedded Markov chain. For $v\in V$ choose a sequence Q_{nv} in $\mathcal{Q}$ that is Hellinger differentiable (13) at Q. Then the assumptions of Section 2 hold with $\Delta=\mathcal{Q}\times\Theta$, $K=V\times\mathbb{R}^d$, $D_Q(v,c)=v$, $D_R(v,c)=c^\top\varrho_\vartheta$. The functional to be estimated is $\varphi(Q,\vartheta)=\vartheta$.

By orthogonality of V and W, its canonical gradient is obtained from (6) as $(0, c_\vartheta^\top \varrho_\vartheta)$ with $d \times d$ matrix c_ϑ determined by

$$c = m^{-1} c_\vartheta^\top J_\vartheta c, \quad c \in \mathbb{R},$$

i.e. $c_\vartheta = m J_\vartheta^{-1}$. Hence the canonical gradient of ϑ is $m J_\vartheta^{-1}(0, \varrho_\vartheta)$ and equals the influence function of $\hat{\vartheta}$, which is therefore efficient for the correctly specified model.

Suppose now that the model is misspecified, and let $\mathcal{R}$ be the set of all transition distributions of the embedded Markov chain. Let R denote the true transition distribution. For $w \in W$ choose a sequence R_{nw} in $\mathcal{R}$ that is Hellinger differentiable (12) at R. Then the assumptions of Section 2 hold with $\Delta = \mathcal{Q} \times \mathcal{R}$, $K = V \times W$, $D_Q(v, w) = v$, $D_R(v, w) = w$. The functional to be estimated is $\varphi(Q, R) = K_R(P_3)$. Heuristically,

$$0 = P_{3nvw}[\varrho_{K_R(P_{3nvw})}] = P_{3nvw}[\varrho_{K_R(P_3)}] + P_{3nvw}[\dot{\varrho}_{K_R(P_3)}](K_R(P_{3nvw}) - K_R(P_3)) + o_p(n^{-1/2}).$$

With $P_{3nvw}[\dot{\varrho}_{K_R(P_3)}] \to P_3[\dot{\varrho}_{K_R(P_3)}]$ we obtain

$$K_R(P_{3nvw}) - K_R(P_3) = -(P_3[\dot{\varrho}_{K_R(P_3)}])^{-1} P_{3nvw}[\varrho_{K_R(P_3)}] + o_p(n^{-1/2}).$$

Write $P_{3nvw} = P_{2nv} \otimes R_{nw}$ and apply the perturbation expansion (14) to obtain

$$\begin{aligned} n^{1/2}(K_R(P_{3nvw}) - K_R(P_3)) &\to -(P_3[\dot{\varrho}_{K_R(P_3)}])^{-1}\Big(P_2[vAR\varrho_{K_R(P_3)}] + P_3[w\varrho_{K_R(P_3)}]\Big) \\ &= -(P_3[\dot{\varrho}_{K_R(P_3)}])^{-1}\Big(P_2[vAR\varrho_{K_R(P_3)}] + P_3[w(\varrho_{K_R(P_3)} - R\varrho_{K_R(P_3)})]\Big), \end{aligned}$$

and the canonical gradient of K_R is obtained from (6) as

$$-m(P_3[\dot{\varrho}_{K_R(P_3)}])^{-1}(AR\varrho_{K_R(P_3)}, \varrho_{K_R(P_3)} - R\varrho_{K_R(P_3)})$$

and equals the influence function of $\hat{\vartheta}_R$, which is therefore efficient for the misspecified model.

5. Model S

While Models Q and R are semiparametric, Models S is parametric. In Model S we assume parametric models q_ϑ and r_ϑ, $\vartheta \in \Theta \subset \mathbb{R}^d$, for the μ-density of the transition distribution of the embedded Markov chain and for the ν-density of the conditional inter-arrival time. We have $s_\vartheta(x, y, u) = q_\vartheta(x, y) r_\vartheta(x, y, u)$. Hence the KL functional $K_S(P_3)$ maximizes $P_3[\log s_\vartheta] = P_2[\log q_\vartheta] + P_3[\log r_\vartheta]$, and the partial maximum likelihood estimator $\hat{\vartheta}_S$ maximizes $\mathbb{P}_3[\log s_\vartheta] = \mathbb{P}_2[\log q_\vartheta] + \mathbb{P}_3[\log r_\vartheta]$. Write

$$\sigma_\vartheta(x, y, u) = \partial_\vartheta \log s_\vartheta(x, y, u) = \chi_\vartheta(x, y) + \varrho_\vartheta(x, y, u)$$

for the d-dimensional vector of partial derivatives of $\log s_\vartheta(x, y, u)$. Then $K_S(P_3)$ solves $P_3[\sigma_\vartheta] = P_2[\chi_\vartheta] + P_3[\varrho_\vartheta] = 0$, and $\hat{\vartheta}_S$ solves $\mathbb{P}_3[\sigma_\vartheta] = \mathbb{P}_2[\chi_\vartheta] + \mathbb{P}_3[\varrho_\vartheta] = 0$. Taylor expansions analogous to (9) and (15) imply

$$\begin{aligned} 0 = \mathbb{P}_3[\sigma_{\hat{\vartheta}_S}] &= \frac{1}{N}\sum_{j=1}^N \sigma_{\hat{\vartheta}_S}(X_{j-1}, X_j, U_j) \\ &= \frac{1}{N}\sum_{j=1}^N \sigma_{K_S(P_3)}(X_{j-1}, X_j, U_j) + \frac{1}{N}\sum_{j=1}^N \dot{\sigma}_{K_S(P_3)}(X_{j-1}, X_j, U_j)(\hat{\vartheta}_S - K_S(P_3)) + o_p(n^{-1/2}), \end{aligned}$$

where $\dot\sigma_\vartheta(x,y,u)=\dot\chi_\vartheta(x,y)+\dot\varrho_\vartheta(x,y,u)$ is the $d\times d$ matrix of partial derivatives of $\sigma_\vartheta(x,y,u)$. We obtain

$$n^{1/2}(\hat\vartheta_S-K_S(P_3))=-m(P_3[\dot\sigma_{K_S(P_3)}])^{-1}n^{-1/2}\sum_{j=1}^N\sigma_{K_S(P_3)}(X_{j-1},X_j,U_j)+o_p(1). \tag{18}$$

If Model S is correctly specified with $Q=Q_\vartheta$ and $R=R_\vartheta$, then $K_S(P_3)=\vartheta$. From Sections 3 and 4 we obtain the Fisher information matrix for Model S as $I_\vartheta+J_\vartheta$. Hence, for the correctly specified model, the maximum likelihood estimator $\hat\vartheta_S$ has the stochastic expansion

$$n^{1/2}(\hat\vartheta_S-\vartheta)=m(I_\vartheta+J_\vartheta)^{-1}n^{-1/2}\sum_{j=1}^N\sigma_\vartheta(X_{j-1},X_j,U_j)+o_p(1).$$

This means that $\hat\vartheta_S$ is asymptotically linear with influence function $m(I_\vartheta+J_\vartheta)^{-1}(\chi_\vartheta,\varrho_\vartheta)$, and $n^{1/2}(\hat\vartheta_S-\vartheta)$ is asymptotically normal with covariance matrix $m(I_\vartheta+J_\vartheta)^{-1}$.

If the model is misspecified, then $\chi_{K_S(P_3)}$ is not in V^d and $\varrho_{K_S(P_3)}$ is not in W^d. We apply the martingale approximation (7) to (18) and see that $\hat\vartheta_S$ is asymptotically linear with influence function

$$-m(P_3[\dot\sigma_{K_S(P_3)}])^{-1}(A\chi_{K_S(P_3)}+AR\varrho_{K_S(P_3)},\varrho_{K_S(P_3)}-R\varrho_{K_S(P_3)}).$$

Hence $n^{1/2}(\hat\vartheta_S-K_S(P_3))$ is asymptotically normal with covariance matrix

$$m(P_3[\dot\sigma_{K_S(P_3)}])^{-1}\Sigma_S(P_3[\dot\sigma_{K_S(P_3)}])^{-1},$$

where

$$\Sigma_S=P_2[A(\chi_{K_S(P_3)}+R\varrho_{K_S(P_3)})A^\top(\chi_{K_S(P_3)}+R\varrho_{K_S(P_3)})]+P_3[(\varrho_{K_S(P_3)}-R\varrho_{K_S(P_3)})(\varrho_{K_S(P_3)}-R\varrho_{K_S(P_3)})^\top].$$

Let us now prove efficiency of $\hat\vartheta_S$, first for the correctly specified model. For $c\in\mathbb{R}^d$ set $\vartheta_{nc}=\vartheta+n^{-1/2}c$. Assume that $q_{nc}=q_{\vartheta_{nc}}$ is Hellinger differentiable (11) at ϑ, and $r_{nc}=r_{\vartheta_{nc}}$ is Hellinger differentiable (17) at ϑ. Then the assumptions of Section 2 hold with $\Delta=\Theta$, $K=\mathbb{R}^d$, $D_Qc=c^\top\chi_\vartheta$, $D_Rc=c^\top\varrho_\vartheta$. The functional to be estimated is $\varphi(\vartheta)=\vartheta$. The canonical gradient is obtained from (6) as $m(I_\vartheta+J_\vartheta)^{-1}(\chi_\vartheta,\varrho_\vartheta)$. It equals the influence function of $\hat\vartheta_S$, which is therefore efficient in the correctly specified model.

Suppose now that the model is misspecified. Let $\mathcal{Q}$ be the set of all transition distributions of the embedded Markov chain, and let $\mathcal{R}$ be the set of all transition distributions of the embedded Markov chain. For $v\in V$ choose a sequence Q_{nv} in $\mathcal{Q}$ that is Hellinger differentiable (13) at Q. For $w\in W$ choose a sequence R_{nw} in $\mathcal{R}$ that is Hellinger differentiable (12) at R. Then the assumptions of Section 2 hold with $\Delta=\mathcal{Q}\times\mathcal{R}$, $K=V\times W$, $D_Q(v,w)=v$, $D_R(v,w)=w$. The functional to be estimated is $\varphi(Q,R)=K_S(P_3)$. Similarly as in Section 4,

$$0=P_{3nvw}[\varrho_{K_S(P_{3nvw})}]=P_{3nvw}[\varrho_{K_S(P_3)}]+P_{3nvw}[\dot\varrho_{K_S(P_3)}](K_S(P_{3nvw})-K_S(P_3))+o_p(n^{-1/2}),$$
$$K_S(P_{3nvw})-K_S(P_3)=-(P_3[\dot\sigma_{K_S(P_3)}])^{-1}P_{3nvw}[\sigma_{K_S(P_3)}]+o_p(n^{-1/2}),$$

and therefore

$$n^{1/2}(K_S(P_{3nvw})-K_R(P_3))\to-(P_3[\dot\varrho_{K_S(P_3)}])^{-1}\Big(P_2[v(A\chi_{K_S(P_3)}+AR\varrho_{K_S(P_3)})]+P_3[w(\varrho_{K_S(P_3)}-R\varrho_{K_S(P_3)})]\Big).$$

Hence by (6) the canonical gradient of K_S is obtained as

$$-m(P_3[\dot\sigma_{K_S(P_3)}])^{-1}(A\chi_{K_S(P_3)}+AR\varrho_{K_S(P_3)},\varrho_{K_S(P_3)}-R\varrho_{K_S(P_3)})$$

and equals the influence function of $\hat{\vartheta}_S$, which is therefore efficient for the misspecified model.

6. Remarks

In this section we comment on examples and possible extensions of our results.

1. If the distribution of the inter-arrival times charges only 1, so that $R(x, y, du) = \delta_1(du)$, then the semi-Markov process reduces to a Markov chain with transition distribution Q, and for Model Q we recover the results of Greenwood and Wefelmeyer [15].

2. Our results carry over to observations $(X_0, T_0), \ldots, (X_n, T_n)$ of the embedded Markov renewal process. Just replace N by n. In particular, instead of the central limit theorem (3) with random summation index N, use

$$n^{-1/2} \sum_{j=1}^{n} f(X_{j-1}, X_j, U_j) \Rightarrow (P_3[f^2])^{1/2} Y,$$

and replace m by 1 everywhere.

In some examples we can describe the KL functional more explicitly.

3. Suppose the embedded Markov chain is a linear autoregressive model of order 1, i.e. $X_j = \vartheta X_{j-1} + \varepsilon_j$, where $\vartheta \in \mathbb{R}$ and the innovations ε_j are i.i.d. with mean 0, finite variance, and known density f. Then Model Q holds with $Q(x, dy) = f(y - \vartheta x)dy$, and $\chi_\vartheta(x, y) = x\ell(y - \vartheta x)$ with $\ell = -f'/f$. Hence the KL functional solves $E[X_0 \ell(X_1 - \vartheta X_0)] = 0$. If f is the density of τY for some $\tau > 0$, then $\ell(x) = \tau^{-2} x$ and $E[X_0 \ell(X_1 - \vartheta X_0)] = \tau^{-2}(E[X_0 X_1] - \vartheta E[X_0^2])$. Hence the KL functional is $K_Q(P_2) = E[X_0 X_1]/E[X_0^2]$, and the partial maximum likelihood estimator for ϑ is the least squares estimator

$$\hat{\vartheta}_Q = K_Q(\mathbb{P}_2) = \frac{\sum_{j=1}^{N} X_{j-1} X_j}{\sum_{j=1}^{N} X_{j-1}^2},$$

a ratio of two empirical estimators.

4. Suppose the inter-arrival time U_j given $X_{j-1} = x$ and $X_j = y$ is exponentially distributed with mean $1/\lambda(x)$ not depending on y,

$$R(x, y, du) = \lambda(x) \exp(-u\lambda(x))du.$$

Then the semi-Markov process is a Markov step process. If the mean is constant, $\lambda(x) = \vartheta$, $\vartheta > 0$, then Model R holds with $R_\vartheta(x, y, du) = \vartheta \exp(\vartheta u)$, and $\varrho_\vartheta(x, y, u) = \vartheta^{-1} - u$. Hence the KL functional solves $E[\varrho_\vartheta(X_0, X_1, U_1)] = \vartheta^{-1} - E[U_1] = 0$, and we obtain $K_R(P_3) = 1/E[U_1]$. The partial maximum likelihood estimator for ϑ is

$$\hat{\vartheta}_R = 1 \Big/ \frac{1}{N} \sum_{j=1}^{N} U_j,$$

a function of an empirical estimator. Efficiency of empirical estimators in Markov step processes is studied in Greenwood and Wefelmeyer [12].

The models Q, R and S are described in terms of the *conditional* distributions $Q(x, dy)$ and $R(x, y, du)$. It is occasionally reasonable to model instead the *marginal* distributions P_1, P_2 or P_3. Results for these three models differ considerably among each other and from Models Q, R and S.

5. Suppose we have a parametric model for the μ-density $p_{1\vartheta}$ of P_1. The *marginal maximum likelihood estimator* $\hat{\vartheta}_1$ maximizes

$$\mathbb{P}_1[\log p_{1\vartheta}] = \frac{1}{N}\sum_{j=1}^{N} \log p_{1\vartheta}(X_{j-1}).$$

It estimates the *KL functional* $K(P_1)$, the parameter that maximizes $P_1[\log p_{1\vartheta}]$. Note that the marginal maximum likelihood estimator is an empirical version of the KL functional, $\hat{\vartheta}_1 = K(\mathbb{P}_1)$.

However, $\hat{\vartheta}_1$ is *not* efficient for ϑ when the marginal model is correctly specified. The reason is that the specification $p_{1\vartheta}$ of the marginal density implies a constraint on the conditional distribution Q of the embedded Markov chain, but the marginal maximum likelihood estimator does not use this information. An efficient estimator for ϑ is difficult to construct. See Kessler, Schick and Wefelmeyer [24] for an efficient estimator of ϑ in a Markov chain model with a (correctly specified) parametric model for the (one-dimensional) marginal density. On the other hand, $\hat{\vartheta}_1$ is efficient for $K(P_1)$ in a nonparametric sense when the marginal model is misspecified.

We note that, in this respect, semi-Markov processes and Markov chains are different from the i.i.d. case. Suppose we have i.i.d. observations (X_j, Y_j) with joint distribution $p_{1\vartheta}(x)dx\,Q(x,dy)$, where Q is unknown. Then Q is not constrained by the marginal model $p_{1\vartheta}$, and the marginal maximum likelihood estimator is efficient for ϑ if the marginal model is correctly specified, and also efficient for $K(P_1)$ if the marginal model is misspecified.

6. Suppose we have a parametric model for the μ^2-density $p_{2\vartheta}$ of P_2. The *marginal maximum likelihood estimator* $\hat{\vartheta}_2$ maximizes

$$\mathbb{P}_2[\log p_{2\vartheta}] = \frac{1}{N}\sum_{j=1}^{N} \log p_{2\vartheta}(X_{j-1}, X_j).$$

It estimates the *KL functional* $K(P_2)$, the parameter that maximizes $P_2[\log p_{2\vartheta}]$, and $\hat{\vartheta}_2 = K(\mathbb{P}_2)$. The perturbation expansion (8) suggests that maximizing $\mathbb{P}_2[\log p_{2\vartheta}]$ is asymptotically equivalent to solving $\mathbb{P}_2[A\chi_\vartheta] = 0$, and the martingale approximation (7) suggests that this is asymptotically equivalent to solving $\mathbb{P}_2[\chi_\vartheta] = 0$. Hence the marginal maximum likelihood estimator $\hat{\vartheta}_2$ is asymptotically equivalent to the conditional maximum likelihood estimator $\hat{\vartheta}_Q$ and therefore efficient in the correctly specified model $p_{2\vartheta}$. The reason is that $p_{2\vartheta}(x,y) = p_{1\vartheta}(x)q_\vartheta(x,y)$, and $q_\vartheta(x,y)$ determines $p_{1\vartheta}$, which therefore does not contain additional information about ϑ.

This is again different from the i.i.d. case. Suppose we have i.i.d. observations (X_j, Y_j) with joint density $p_{1\vartheta}(x)q_\vartheta(x,y)$. Then $p_{1\vartheta}$ contains, in general, additional information about ϑ.

7. Suppose we have a parametric model for the $\mu^2\otimes\nu$-density $p_{3\vartheta}$ of P_3. The *marginal maximum likelihood estimator* $\hat{\vartheta}_3$ maximizes

$$\mathbb{P}_3[\log p_{3\vartheta}] = \frac{1}{N}\sum_{j=1}^{N} \log p_{3\vartheta}(X_{j-1}, X_j, U_j).$$

It estimates the *KL functional* $K(P_3)$, the parameter that maximizes $P_3[\log p_{3\vartheta}]$, and $\hat{\vartheta}_3 = K(\mathbb{P}_3)$. We can write $p_{3\vartheta}(x,y,u) = p_{2\vartheta}(x,y)r_\vartheta(x,y,u)$. Now $r_\vartheta(x,y,u)$ carries additional information about ϑ, similarly as in the i.i.d. case.

8. Remark 5 tells us in particular the following, rather obvious, fact. If a parametric estimator is efficient in a nonparametric sense, then the reason is not that it is efficient in a parametric model. Rather, an estimator usually is nonparametrically efficient because it is a smooth function of an empirical estimator. We can illustrate this also with Model S. Suppose we have parametric models q_ϑ and r_ϑ for the densities of Q and R. Let $\hat{\vartheta}_Q = K_Q(\mathbb{P}_2)$ be the conditional maximum likelihood estimator based on the model q_ϑ alone.

In general, $\hat{\vartheta}_Q$ will not be efficient for ϑ when model S is correctly specified, because $\hat{\vartheta}_Q$ does not use the information about ϑ in the model r_ϑ. But if both q_ϑ and r_ϑ are misspecified, $\hat{\vartheta}_Q$ will be nonparametrically efficient for $K_Q(P_2)$, which is the KL functional for Model Q but not for Model S.

References

[1] 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.
[2] Bickel, P. J., 1993, Estimation in semiparametric models. In: (C. R. Rao, Ed.) *Multivariate Analysis: Future Directions* (Amsterdam: North-Holland), pp. 55–73
[3] Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A., 1998, *Efficient and Adaptive Estimation for Semiparametric Models* (New York: Springer).
[4] Bickel, P. J. and Kwon, J., 2001, Inference for semiparametric models: Some questions and an answer (with discussion). *Statist. Sinica* **11**, 863–960.
[5] Dahlhaus, R. and Wefelmeyer, W., 1996, Asymptotically optimal estimation in misspecified time series models. *Ann. Statist.* **24**, 952–974.
[6] Daniels, H. E., 1961, The asymptotic efficiency of a maximum likelihood estimator. *Proc. Fourth Berkeley Sympos. Math. Statist. and Probability* **1**, 151–163.
[7] Doksum, K., Ozeki, A., Kim, J. and Neto, E. C., 2007, Thinking outside the box: Statistical inference based on Kullback-Leibler empirical projections. *Statist. Probab. Lett.* **77**, 1201-1213
[8] Dürr, D. and Goldstein, S., 1986, Remarks on the central limit theorem for weakly dependent random variables. In: (S. Albeverio, P. Blanchard and L. Streit, Eds.) *Stochastic Processes — Mathematics and Physics*, Lecture Notes in Mathematics **1158** (Berlin: Springer), pp. 104–118
[9] Gordin, M. I., 1969, The central limit theorem for stationary processes. *Soviet Math. Dokl.* **10**, 1174–1176.
[10] Gordin, M. I. and Lifšic, B. A. 1978, The central limit theorem for stationary Markov processes. *Soviet Math. Dokl.* **19**, 392–394.
[11] Greenwood, P. E., Müller, U. U. and Wefelmeyer, W., 2004, Efficient estimation for semiparametric semi-Markov processes. *Comm. Statist. Theory Methods* **33**, 419-435.
[12] Greenwood, P. E. and Wefelmeyer, W., 1994, Nonparametric estimators for Markov step processes. *Stochastic Process. Appl.* **52**, 1-16.
[13] Greenwood, P. E. and Wefelmeyer, W., 1995, Efficiency of empirical estimators for Markov chains, *Ann. Statist.*, **23**, 132–143.
[14] Greenwood, P. E. and Wefelmeyer, W., 1996, Empirical estimators for semi-Markov processes. *Math. Meth. Statist.* **5**, 299-315.
[15] Greenwood, P. E. and Wefelmeyer, W., 1997, Maximum likelihood estimator and Kullback–Leibler information in misspecified Markov chain models. *Theory Probab. Appl.* **42**, 103–111.
[16] Höpfner, R., 1993a, On statistics of Markov step processes: representation of log-likelihood ratio processes in filtered local models. *Probab. Theory Related Fields* **94**, 375–398.
[17] Höpfner, R., 1993b, Asymptotic inference for Markov step processes: observation up to a random time. *Stochastic Process. Appl.* **48**, 295–310.
[18] Höpfner, R., Jacod, J. and Ladelli, L., 1990, Local asymptotic normality and mixed normality for Markov statistical models. *Probab. Theory Related Fields* **86**, 105–129.
[19] Hosoya, Y., 1989, The bracketing condition for limit theorems on stationary linear processes. *Ann. Statist.* **17**, 401–418.
[20] Huber, P. J., 1967, The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* **1**, 221–233.
[21] 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.
[22] 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.
[23] Kartashov, N. V., 1996, *Strong Stable Markov Chains* (Utrecht: VSP).
[24] Kessler, M., Schick, A. and Wefelmeyer, W., 2001, The information in the marginal law of a Markov chain. *Bernoulli* **7**, 243-266.
[25] Kutoyants, Yu. A., 1988, On an identification problem for dynamical systems with small noise. *Izv. Akad. Nauk Armyan. SSR* **23**, 270–285.
[26] Kutoyants, Yu. A., 2004, *Statistical Inference for Ergodic Diffusion Processes*, Springer Series in Statistics (London: Springer).
[27] Maigret, N., 1978, Théorème de limite centrale fonctionnel pour une chaîne de Markov récurrente au sens de Harris et positive. *Ann. Inst. H. Poincaré Probab. Statist.* **14**, 425–440.
[28] McKeague, I. W., 1984, Estimation for diffusion processes under misspecified models. *J. Appl. Probab.* **21**, 511–520.
[29] Meyn, S. P. and Tweedie, R. L., 1993, *Markov Chains and Stochastic Stability* (London: Springer).
[30] Müller, U. U., 2007, Weighted least squares estimators in possibly misspecified nonlinear regression. *Metrika* **66**, 39–59.
[31] Ogata, Y., 1980, Maximum likelihood estimates of incorrect Markov models for time series and the derivation of AIC. *J. Appl. Probab.* **17**, 59–72.
[32] Penev, S., 1991, Efficient estimation of the stationary distribution for exponentially ergodic Markov chains. *J. Statist. Plann. Inference* **27**, 105–123.
[33] Pollard, D., 1985, New ways to prove central limit theorems. *Econometric Theory* **1**, 295–314.
[34] Sin, C.-Y. and White, H., 1996, Information criteria for selecting possibly misspecified parametric models. *J. Econometrics* **71**, 207–225.
[35] White, H., 1982, Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.
[36] White, H., 1984, Maximum likelihood estimation of misspecified dynamic models. In: T. K. Dijkstra (Ed) *Misspecification Analysis*, Lecture Notes in Economics and Mathematical Systems **237** (Berlin: Springer), pp. 1–19.
[37] White, H., 1994, *Estimation, Inference and Specification Analysis*, Econometric Society Monographs **22** (Cambridge: Cambridge University Press).