# The information in the marginal law of a Markov chain

Mathieu Kessler * Anton Schick † Wolfgang Wefelmeyer* ‡

**Abstract**

If we have a parametric model for the invariant distribution of a Markov chain but cannot or do not want to use any information about the transition distribution (except, perhaps, that the chain is reversible) — what, then, is the best use we can make of the observations? We determine a lower bound for the asymptotic variance of regular estimators and show constructively that the bound is attainable. The results apply to discretely observed diffusions.

*AMS 1991 subject classifications.* Primary 62G20, 62M05; secondary 62F12.

*Key words and Phrases.* Efficient estimator, ergodic Markov chain, discretely observed diffusion.

## 1 Introduction

Let $X_0, \ldots, X_n$ be observations from a stationary Markov chain. Suppose we have a parametric model $\pi_\vartheta(dx)$, $\vartheta \in \Theta$, for the distribution of $X_i$, but no convincing or tractable model for the transition distribution, say $Q(x, dy)$, of $X_i$ given $X_{i-1} = x$. We want to estimate $\vartheta$.

It is clear that it is not optimal to proceed as if the observations were independent. The possible transition distributions are constrained by the condition that their invariant distribution must be in the parametric family $\pi_\vartheta$, $\vartheta \in \Theta$. Hence additional information about $\vartheta$ is likely to be obtainable through an estimator of $Q$. We pose the following questions. How much information about $\vartheta$ is contained in the observations? How can one exploit this information for estimating $\vartheta$? The answers are surprisingly involved.

*Departamento de Matemática Aplicada y Estadística, Universidad Politécnica de Cartagena, 30203 Cartagena, Spain. Mathieu.Kessler@upct.es. http://spongeman.upct.es/∼mathieu/.

†Department of Mathematical Sciences, Binghamton University, Binghamton, NY 13902-6000, USA. anton@math.binghamton.edu. http://math.binghamton.edu/anton/index.html.

‡Fachbereich 6 Mathematik, Universität-GH Siegen, Walter-Flex-Str. 3, 57068 Siegen, Germany. wefelmeyer@mathematik.uni-siegen.de. http://www.math.uni-siegen.de/statistik/wefelmeyer.html.

The paper is organized as follows. In Theorem 1, Section 3, we describe the information about $\vartheta$ by determining a lower bound for the asymptotic variance of regular estimators. In Theorem 2, Section 4, we show that reversibility of the chain carries no additional information about $\vartheta$. In Theorem 3, Section 5, we describe how to construct an efficient estimator if a $n^{1/2}$-consistent estimator of $\vartheta$ and an appropriate estimator of the efficient influence function are available. The construction utilizes the sample splitting techniques of Schick (1998). Theorem 4, Section 6, gives an explicit construction of an estimator of the efficient influence function with the desired properties. Section 7 compares our results with known results for *parametric* Markov chain models.

The results apply when we have a parametric model for a stationary continuous-time stochastic process and observe the process at $n+1$ equidistant time points. Then the marginal distribution of the observations usually follows a tractable parametric model, while the transition distribution is often intractable. In Section 8 we compare our estimator with certain estimators based on parametric diffusion models which have been suggested in the literature. Our estimator has the advantage of being robust against misspecification of the underlying continuous-time process.

# 2 Characterization of efficient estimators

In this section we introduce some notation and recall a characterization of least dispersed regular (i.e. *efficient*) estimators for real-valued functionals of Markov chain models. Let $X_0, \ldots, X_n$ be observations from a stationary Markov chain on an arbitrary state space $S$ with countably generated $\sigma$-field $\mathcal{S}$, with transition distribution $Q(x, dy)$ and invariant distribution $\pi(dx)$.

We will use the following notation. The joint law of two successive observations is

$$\pi \otimes Q\,(dx, dy) = \pi(dx)Q(x, dy).$$

For a suitably integrable function $f(x)$ write

$$(Qf)(x) = \int Q(x, dy)f(y), \quad \pi f = \int \pi(dx)f(x).$$

For a function $k(x, y)$ of two arguments we write

$$(Qk)(x) = \int Q(x, dy)k(x, y). \tag{2.1}$$

For $j \geq 2$, let $Q^j k = Q^{j-1}Qk$ so that $(Q^j k)(X_0) = \mathrm{E}[k(X_{j-1}, X_j)|X_0]$. This differs from the application of the $j$-step transition measure $Q^j$ to $k$ in the sense of (2.1), which would give $\mathrm{E}[k(X_0, X_j)|X_0]$.

It will later be convenient to write functions $f(x)$ of one argument as functions of two arguments,

$$(Lf)(x, y) = f(x), \quad (Rf)(x, y) = f(y).$$

Here $L$ and $R$ stand for 'left' and 'right'.

For a measure $\nu$, let $L_2(\nu)$ be the space of $\nu$-square integrable functions, and $L_{2,0}(\nu)$ the subspace of functions with $\nu$-integral 0. Let $\|f\| = (\pi f^2)^{1/2}$ denote the norm of a function $f$ in $L_2(\pi)$, and $\|K\| = \sup\{\|Kf\| : \|f\| = 1\}$ the corresponding operator norm of a kernel $K(x, dy)$. Write $J(x, dy) = \varepsilon_x(dy)$ for the *identity kernel*, and $\Pi(x, dy) = \pi(dy)$ for the *stationary projection*. We have

$$\Pi Q = Q\Pi = \Pi. \tag{2.2}$$

The following assumption will be in force throughout.

**Assumption 1.** The chain fulfills $\|Q - \Pi\| < 1$.

We introduce a *local model* around $Q$ by perturbing $Q$ as follows. As local parameter space we take

$$H = \{h \in L_2(\pi \otimes Q\,) : Qh = 0\}. \tag{2.3}$$

For $h \in H$ we set

$$Q_{nh}(x, dy) = Q(x, dy)[1 + n^{-1/2}h_n(x, y)] \tag{2.4}$$

with

$$h_n = \overline{h}_n - LQ\overline{h}_n \quad \text{and} \quad \overline{h}_n = h1_{(2|h|\le n^{1/8})}. \tag{2.5}$$

We have used the truncated and centered version $h_n$ of $h$ because $Q_{nh}(x, dy)$ must be a probability measure.

Write $P_n$ and $P_{nh}$ for the joint distribution of $(X_0, \ldots, X_n)$ under the transition distribution $Q$ and $Q_{nh}$, respectively. Under Assumption 1, we have a nonparametric version of *local asymptotic normality*,

$$\begin{aligned} \log \frac{dP_{nh}}{dP_n}(X_0, \ldots, X_n) &= \log \frac{d\pi_{nh}}{d\pi}(X_0) + \sum_{i=1}^{n} \log[1 + n^{-1/2}h_n(X_{i-1}, X_i)] \\ &= n^{-1/2} \sum_{i=1}^{n} h(X_{i-1}, X_i) - \frac{1}{2}\, \pi \otimes Q\, h^2 + o_{P_n}(1) \end{aligned} \tag{2.6}$$

and

$$n^{-1/2} \sum_{i=1}^{n} h(X_{i-1}, X_i) \Rightarrow (\pi \otimes Q\, h^2)^{1/2} \cdot N, \tag{2.7}$$

where $N$ is standard normal. A parametric version of local asymptotic normality for Markov chains was first given in Roussas (1965); a nonparametric version in Penev

(1991). Local asymptotic normality for Markov step processes and Hellinger differentiable $Q_{nh}$ in the sense of Höpfner, Jacod and Ladelli (1990), and hence for $Q_{nh}$ as in (2.4), is proved in Höpfner (1993a, 1993b). He starts the chain in a fixed value $X_0 = x_0$, so that $\log d\pi_{nh}/d\pi(X_0)$ vanishes. We consider a stationary chain, for which $\log d\pi_{nh}/d\pi(X_0)$ is negligible because the invariant distribution $\pi$ depends continuously on the transition distribution; see Kartashov (1996).

So far we have looked at the full nonparametric model of all (sufficiently regular) transition distributions. Consider now a submodel, described by a family $\mathcal{Q}$ of transition distributions on $\mathcal{S}$. Suppose $\mathcal{Q}$ contains the transition distribution $Q$ fixed above. The local model is now obtained by perturbing $Q$ *within* the family $\mathcal{Q}$. In regular cases, the local parameter space will then run through a linear subspace $H_0$ of $H$. For $Q_{nh}$ to lie *exactly* in $\mathcal{Q}$, the construction (2.4) and (2.5) will have to be modified slightly. For the models considered below, we will omit the (tedious) details.

Consider a real-valued functional $t$ on $\mathcal{Q}$. It is called *differentiable* at $Q$ with *gradient* $g$ if $g \in H$ and

$$n^{1/2}[t(Q_{nh}) - t(Q)] \to \pi \otimes Q\,(hg) \quad \text{for } h \in H_0. \tag{2.8}$$

The *canonical gradient* is the projection $g_0$ of $g$ onto $H_0$.

Let $T_n$ be an estimator of $t(Q)$. We call $T_n$ *asymptotically linear* at $t(Q)$ with *influence function* $h$ if $h \in H$ and

$$n^{1/2}[T_n - t(Q)] = n^{-1/2} \sum_{i=1}^{n} h(X_{i-1}, X_i) + o_{P_n}(1).$$

We call $T_n$ *regular* at $Q$ with *limit* $L$ if

$$n^{1/2}[T_n - t(Q_{nh})] \Rightarrow L \quad \text{under } P_{nh} \text{ for } h \in H_0.$$

The convolution theorem of Hájek (1970) in the version of Pfanzagl and Wefelmeyer (1982, Theorem 9.3.1), see now Bickel, Klaassen, Ritov and Wellner (1993, p.63, Theorem 2), says that if $T_n$ is regular, then

$$\begin{gathered}\Big\{n^{-1/2} \textstyle\sum_{i=1}^n g_0(X_{i-1}, X_i),\ n^{1/2}[T_n - t(Q)] - n^{-1/2} \textstyle\sum_{i=1}^n g_0(X_{i-1}, X_i)\Big\} \\ \Rightarrow \big\{(\pi \otimes Q\, g_0^2)^{1/2} \cdot N, M\big\} \quad \text{under } P_n,\end{gathered}$$

with $N$ standard normal and $M$ independent of $N$. In particular,

$$L = (\pi \otimes Q\, g_0^2)^{1/2} \cdot N + M \quad \text{in distribution.}$$

The estimator $T_n$ is (asymptotically) least dispersed if

$$L = (\pi \otimes Q\, g_0^2)^{1/2} \cdot N \quad \text{in distribution.} \tag{2.9}$$

By the convolution theorem, $T_n$ is least dispersed among all regular estimators for $t(Q)$ if and only if it is asymptotically linear with influence function equal to the canonical gradient,

$$n^{1/2}[T_n - t(Q)] = n^{-1/2} \sum_{i=1}^{n} g_0(X_{i-1}, X_i) + o_{P_n}(1). \tag{2.10}$$

# 3 The information in the marginal law

As in Section 2, let $X_0, \ldots, X_n$ be observations from a stationary Markov chain on an arbitrary state space $S$ with countably generated $\sigma$-field $\mathcal{S}$. Suppose we have a parametric model $\{\pi_\tau : \tau \in \Theta\}$ for the invariant distribution, and that the transition distribution is unspecified otherwise.

We consider two submodels of the full nonparametric model. The first, $\mathcal{Q}_*$, consists of all transition distributions with invariant distribution in the family $\{\pi_\tau : \tau \in \Theta\}$. The second, $\mathcal{Q}_*^{\text{rev}}$, consists of all transition distributions which fulfill the additional restriction that the chain is reversible. The models are semiparametric, or rather nonparametric with a parametric family of restrictions. (In Section 7 we will also discuss models described by a parametric family of transition distributions.) We are interested in estimating $\tau$.

For simplicity we take $\Theta$ one-dimensional and open. We fix a parameter $\vartheta$. In the following, we will often suppress this parameter in the notation. In particular, we will write $\pi$ for $\pi_\vartheta$. In this and the next section, the following additional assumption will be in force. We need it to determine a lower bound for the asymptotic variance of estimators of $\vartheta$. It is the usual condition in the i.i.d. case.

**Assumption 2.** For $\tau \in \Theta$, the invariant distribution $\pi_\tau$ has a positive $\mu$-density $p_\tau$, and the map $\tau \mapsto p_\tau$ is *Hellinger differentiable* at $\vartheta$: There is a function $\ell \in L_{2,0}(\pi)$, the *Hellinger derivative*, such that

$$\mu \left[ p_\tau^{1/2} - p_\vartheta^{1/2} - \frac{1}{2}(\tau - \vartheta)\ell p_\vartheta^{1/2} \right]^2 = o[(\tau - \vartheta)^2]. \tag{3.1}$$

Also, $\pi\ell^2 > 0$.

Fix a transition distribution $Q$ with invariant distribution $\pi = \pi_\vartheta$. The local model around $Q$ is obtained by perturbing $Q$ as in (2.4), subject to the restriction that the invariant distributions are in the family $\{\pi_\tau : \tau \in \Theta\}$. The restriction entails a restriction on the local parameter $h$ of the perturbed transition distribution $Q_{nh}$. To determine the restriction, we consider the invariant distribution of $Q_{nh}$. By Kartashov (1985a, 1985b, 1996), the transition distribution $Q_{nh}$ has a unique invariant distribution $\pi_{nh}$ which admits the following perturbation expansion: For $h \in H$ and $f \in L_2(\pi)$,

$$n^{1/2}(\pi_{nh} f - \pi f) \to \pi \otimes Q\,(h \cdot RUf), \tag{3.2}$$

where $U$ is the kernel

$$U = \sum_{j=0}^{\infty} (Q^j - \Pi) \quad \text{on } L_2(\pi). \tag{3.3}$$

Since $Qh = 0$, we may center $RUf$,

$$\pi \otimes Q\,(h \cdot RUf) = \pi \otimes Q\,(h \cdot Af), \tag{3.4}$$

where

$$A = RU - LQU = \sum_{j=0}^{\infty} (RQ^j - LQ^{j+1}) \quad \text{on } L_2(\pi). \tag{3.5}$$

The operator $A$ maps $L_2(\pi)$ into $H$,

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

We will need the adjoint of $A$ in the inner product (3.4). It is expressed in terms of the *reversed* chain, with transition distribution $\overline{Q}(y, dx)$ defined by

$$\pi(dx)Q(x, dy) = \pi(dy)\overline{Q}(y, dx). \tag{3.6}$$

For a function $h(x, y)$ of two arguments we will follow the convention that the transition distribution of the reversed chain acts on $h$ from right to left, i.e. on the first argument of $h$,

$$(\overline{Q}h)(y) = \int \overline{Q}(y, dx)h(x, y).$$

For $j \geq 2$, let $\overline{Q}^j h = \overline{Q}^{j-1}\overline{Q}h$. Introduce

$$\overline{V} = \sum_{j=1}^{\infty} (\overline{Q}^j - \pi \otimes Q) \quad \text{on } L_2(\pi \otimes Q).$$

For $f \in L_2(\pi)$ and $h \in H$,

$$\pi \otimes Q\,(h \cdot Af) = \pi(\overline{V}h \cdot f). \tag{3.7}$$

This is Lemma 1 of Greenwood and Wefelmeyer (1999), specialized to functions of *one* argument. With (3.4) and (3.7), the perturbation expansion (3.2) is

$$n^{1/2}(\pi_{nh} f - \pi f) \to \pi(\overline{V}h \cdot f). \tag{3.8}$$

So far, we have not used the restriction that the invariant distributions are in the parametric family $\{\pi_\tau : \tau \in \Theta\}$. Hellinger differentiability (3.1) of the invariant distribution implies for all bounded functions $f$ and $u \in \mathbf{R}$,

$$n^{1/2}(\pi_{\vartheta+n^{-1/2}u}f - \pi f) \to u\pi(\ell f). \tag{3.9}$$

Comparing with (3.8), we obtain a restriction on the local parameter $h$, namely $\overline{V}h = u\ell$ if $Q_{nh}$ has invariant distribution $\pi_{\vartheta+n^{-1/2}u_n}$ with $u_n \to u$. Hence the local parameter space of $\mathcal{Q}_*$ at $Q$ is

$$H_* = \bigcup_{u\in\mathbf{R}} H_u$$

with

$$H_u = \{h \in H : \overline{V}h = u\ell\}.$$

We turn to the problem of determining a lower bound for the variance of estimators for the parameter $\tau$. According to Section 2, the bound is expressed in terms of the canonical gradient. Consider $\tau$ as a functional on $\mathcal{Q}_*$, defined by $t(Q) = \tau$ if $Q$ has invariant distribution $\pi_\tau$. Then

$$n^{1/2}[t(Q_{nh}) - t(Q)] = n^{1/2}(\vartheta + n^{-1/2}u - \vartheta) + o(1) \to u \quad \text{for } h \in H_u.$$

By definition (2.8), a gradient $g \in H$ is determined by

$$\pi \otimes Q\,(hg) = u \quad \text{for } h \in H_u. \tag{3.10}$$

The canonical gradient will turn out to be of the form $Af$ with $f \in L_2(\pi)$. The following simple characterization will be useful.

**Lemma 1.** *Let $f \in L_2(\pi)$. Then $Af$ is a gradient for $\vartheta$ if and only if $\pi(\ell f) = 1$.*

**Proof.** We have $\overline{V}h = u\ell$ for $h \in H_u$. By (3.7),

$$\pi \otimes Q\,(h \cdot Af) = \pi(\overline{V}h \cdot f) = u\pi(\ell f).$$

Hence (3.10) holds for $g = Af$ if and only if $\pi(\ell f) = 1$. □

The canonical gradient, say $g_*$, is the projection of an arbitrary gradient into $H_*$. In particular, $\overline{V}g_* = u\ell$ for some $u$. Does the class of gradients in Lemma 1 contain the *canonical* gradient? This is the case if we can find $f \in L_2(\pi)$ such that $\overline{V}Af = u\ell$, with $u$ determined by $\pi(\ell f) = 1$. A sufficient condition is invertibility of $\overline{V}A$. To calculate $\overline{V}A$, we introduce an operator $V$ analogous to $\overline{V}$,

$$V = \sum_{j=1}^{\infty}(Q^j - \pi \otimes Q) \quad \text{on } L_2(\pi \otimes Q).$$

In accordance with our convention, the restrictions of $V$ and $\overline{V}$ to functions of *one* variable are $VR$ and $\overline{V}L$, or

$$V = \sum_{j=1}^{\infty}(Q^j - \Pi), \quad \overline{V} = \sum_{j=1}^{\infty}(\overline{Q}^j - \overline{\Pi}) \quad \text{on } L_2(\pi).$$

We have

$$\overline{V}A = J - \Pi + V + \overline{V} \quad \text{on } L_2(\pi). \tag{3.11}$$

This is Lemma 2 of Greenwood and Wefelmeyer (1999), specialized to functions of one argument.

**Lemma 2.** *The operator $\overline{V}A$ is invertible on $L_{2,0}(\pi)$, and*

$$\begin{aligned}(\overline{V}A)^{-1} &= (J-Q)(J-\overline{Q}Q)^{-1}(J-\overline{Q}) \\ &= \sum_{j=0}^{\infty}(J-Q)(\overline{Q}Q)^j(J-\overline{Q}) \quad \textit{on } L_{2,0}(\pi).\end{aligned}$$

**Proof.** On $L_{2,0}(\pi)$,

$$V = \sum_{j=1}^{\infty} Q^j = Q(J-Q)^{-1}, \quad \overline{V} = \sum_{j=1}^{\infty}\overline{Q}^j = (J-\overline{Q})^{-1}\overline{Q}.$$

With relation (3.11), we find that on $L_{2,0}(\pi)$

$$\begin{aligned}\overline{V}A &= J + V + \overline{V} \\ &= J + Q(J-Q)^{-1} + (J-\overline{Q})^{-1}\overline{Q} \\ &= (J-\overline{Q})^{-1}[(J-\overline{Q})(J-Q) + \overline{Q}(J-Q) + (J-\overline{Q})Q](J-Q)^{-1} \\ &= (J-\overline{Q})^{-1}(J-\overline{Q}Q)(J-Q)^{-1}.\end{aligned}$$

Now use the fact that $Q$ and $\overline{Q}$ viewed as operators on $L_{2,0}(\pi)$ have norms less than 1 in view of Assumption 1. Actually, both norms equal $\|Q - \Pi\|$. Hence $\overline{V}A$ is invertible on $L_{2,0}(\pi)$, and the inverse has the asserted form. □

**Theorem 1.** *The canonical gradient for $\vartheta$ is*

$$g_* = [\pi(\ell e_*)]^{-1}Ae_* \quad \textit{with} \quad e_* = (\overline{V}A)^{-1}\ell.$$

*We have*

$$\pi \otimes Q\,(Ae_*)^2 = \pi(\ell e_*), \tag{3.12}$$

$$\pi \otimes Q\,g_*^2 = [\pi(\ell e_*)]^{-1}. \tag{3.13}$$

**Proof.** By Lemma 2, the operator $\overline{V}A$ is invertible on $L_{2,0}(\pi)$. The function $e_*$ fulfills $\overline{V}Ae_* = \ell$, hence $Ae_* \in H_1 \subset H_*$. Furthermore, $g_* = [\pi(\ell e_*)]^{-1}Ae_*$ is a gradient by Lemma 1. Finally, (3.12) follows from the fact that $\overline{V}$ is the adjoint of $A$, and implies (3.13). □

By (2.9), a least dispersed regular estimator for $\vartheta$ in $\mathcal{Q}_*$ has asymptotic variance $\pi \otimes Q\, g_*^2$. The inverse, $\pi(\ell e_*)$, may therefore be called the *information* about $\vartheta$ contained in the marginal laws of the Markov chain.

**Remark 1.** Suppose the observations $X_0, \ldots, X_n$ happen to be i.i.d. Then the best estimator is the maximum likelihood estimator. It solves $\sum_{i=1}^n \ell_\tau(X_i) = 0$. Theorem 1 implies an infinitesimal robustness property of the maximum likelihood estimator against Markovian departures from independence: We have $Q = \Pi$, $\overline{Q} = \overline{\Pi}$ and $V = \overline{V} = 0$ on $L_2(\pi)$, so that, by (3.11), $\overline{V}A = J$ and $(\overline{V}A)^{-1} = J$ on $L_{2,0}(\pi)$. By Theorem 1, the canonical gradient for $\vartheta$ is $g_* = (\pi\ell^2)^{-1}R\ell$. By the characterization (2.10), an estimator $\hat{\vartheta}_n$ is least dispersed and regular for $\vartheta$ in the model $\mathcal{Q}_*$ if and only if

$$n^{1/2}(\hat{\vartheta}_n - \vartheta) = (\pi\ell^2)^{-1}n^{-1/2}\sum_{i=1}^n \ell(X_i) + o_{P_n}(1).$$

Under appropriate regularity conditions, the maximum likelihood estimator has this stochastic approximation. For a related robustness result in fully nonparametric Markov chain models see Penev (1993).

# 4 The information for reversible chains

In this section we show that reversibility of the Markov chain carries no additional information about the parameter of the invariant distribution. (A related result is proved in Greenwood and Wefelmeyer (1999): In a *nonparametric* Markov chain model, reversibility carries no information about functionals of the invariant distribution.) Nevertheless, the canonical gradient simplifies for reversible chains.

Consider the model $\mathcal{Q}_*^{\mathrm{rev}}$ of all reversible transition distributions in $\mathcal{Q}_*$,

$$\mathcal{Q}_*^{\mathrm{rev}} = \{Q \in \mathcal{Q}_* : \overline{Q}(x, dy) = Q(x, dy)\}.$$

Then $\pi \otimes Q$ is symmetric in the two components. To translate this property into a property of local parameters, we extend some results of Section 3; see also Greenwood and Wefelmeyer (1999, Section 3).

The perturbation expansion (3.2) generalizes immediately to functions of two arguments: For $h \in H$ and $k \in L_2(\pi \otimes Q)$,

$$n^{1/2}(\pi_{nh} \otimes Q_{nh}\, k - \pi \otimes Q\, k) \to \pi \otimes Q\,(h \cdot Ak), \tag{4.1}$$

where

$$A = I_2 - LQ + AQ \quad \text{on } L_2(\pi \otimes Q), \tag{4.2}$$

with $I_2$ the identity on $L_2(\pi \otimes Q)$. Note that $A$ maps $L_2(\pi \otimes Q)$ *onto* $H$. The adjoint of this extended operator is obtained from (3.7): For $h \in H$ and $k \in L_2(\pi \otimes Q)$,

$$\pi \otimes Q\,(h \cdot Ak) = \pi \otimes Q\,(Bh \cdot k), \tag{4.3}$$

where

$$B = I_2 + L\overline{V} \quad \text{on } H. \tag{4.4}$$

If $Q$ and $Q_{nh}$ are reversible, then $\pi \otimes Q$ and $\pi_{nh} \otimes Q_{nh}$ are symmetric. The perturbation expansion (4.1) and relation (4.3) imply that $Bh$ is symmetric. Hence the local parameter space of $\mathcal{Q}_*^{\text{rev}}$ is

$$H_*^{\text{rev}} = \{h \in H_* : Bh \text{ symmetric}\}.$$

The canonical gradient $g_* = [\pi(\ell e_*)]^{-1} A e_*$ is of the form $Af$ with $f \in L_2(\pi)$. We show that such functions fulfill the additional property of $H_*^{\text{rev}}$, namely, $BAf$ is symmetric. We have

$$BA = I_2 - \pi \otimes Q + RV + L\overline{V} \quad \text{on } L_2(\pi \otimes Q).$$

This is Lemma 2 of Greenwood and Wefelmeyer (1999). We rewrite $BA$ for functions of one argument. We have $QR = Q$ and $QL = J$. Similarly, $\overline{Q}R = J$ and $\overline{Q}L = \overline{Q}$. Hence $VL = U$ and $\overline{V}R = \overline{U}$, where

$$\overline{U} = \sum_{j=0}^{\infty} (\overline{Q}^j - \overline{\Pi}) \quad \text{on } L_2(\pi) \tag{4.5}$$

is defined in analogy to (3.3). We obtain $BAR = BAL = RU + L\overline{U}$ on $L_2(\pi)$ and may write, without ambiguity,

$$BA = RU + L\overline{U} \quad \text{on } L_2(\pi). \tag{4.6}$$

In particular, if the chain is reversible, $Q = \overline{Q}$, then $BAf$ is symmetric for $f \in L_2(\pi)$.

**Theorem 2.** *Let $Q = \overline{Q}$. Then the canonical gradient for $\vartheta$ in model $\mathcal{Q}_*^{\text{rev}}$ equals the canonical gradient $g_* = [\pi(\ell e_*)]^{-1} A e_*$ for $\vartheta$ in model $\mathcal{Q}_*$. We have $\pi \otimes Q\, g_*^2 = [\pi(\ell e_*)]^{-1}$ and*

$$\begin{aligned} e_* &= \ell + 2\sum_{j=1}^{\infty} (-1)^j Q^j \ell, \\ \pi(\ell e_*) &= \pi\ell^2 + 2\sum_{j=1}^{\infty} (-1)^j \pi(\ell \cdot Q^j \ell), \\ Ae_* &= R\ell + (R+L)\sum_{j=1}^{\infty} (-1)^j Q^j \ell. \end{aligned}$$

**Proof.** By (4.6),

$$Bg_* = [\pi(\ell e_*)]^{-1} BAe_* = [\pi(\ell e_*)]^{-1}(RUe_* + L\overline{U}e_*).$$

Since $Q = \overline{Q}$, we have $U = \overline{U}$, and $Bg_*$ is symmetric. Hence $g_* \in H_*^{\mathrm{rev}}$. This shows that $g_*$ is canonical in the model $\mathcal{Q}_*^{\mathrm{rev}}$. Finally, on $L_{2,0}(\pi)$ we have $(\overline{V}A)^{-1} = (J-Q)(J+Q)^{-1}$, $A = (R - LQ)(J - Q)^{-1}$ and $A(\overline{V}A)^{-1} = (R - LQ)(J + Q)^{-1}$. Now expand $(J + Q)^{-1}$ as a series to get the desired formulas. □

# 5 Construction of efficient estimators

In this section we construct least dispersed regular estimators for $\vartheta$. We need a stronger version of Assumption 2, namely *continuous* Hellinger differentiability of $\tau \to p_\tau$ at $\vartheta$.

**Assumption 3.** For $\tau \in \Theta$, the invariant distribution $\pi_\tau$ has a positive $\mu$-density $p_\tau$. The function $\tau \to p_\tau$ is Hellinger differentiable with derivative $\ell_\tau$ in a neighborhood of $\vartheta$, and

$$\mu(\ell_\tau p_\tau^{1/2} - \ell_\vartheta p_\vartheta^{1/2})^2 \to 0 \quad \text{for } \tau \to \vartheta. \tag{5.1}$$

Also, $\pi_\vartheta \ell_\vartheta^2 > 0$.

By Theorem 1 and the characterization (2.10), an estimator $\hat{\vartheta}_n$ is least dispersed and regular for $\vartheta$ if and only if

$$n^{1/2}(\hat{\vartheta}_n - \vartheta) = [\pi(\ell e_*)]^{-1} n^{-1/2} \sum_{i=1}^{n} (Ae_*)(X_{i-1}, X_i) + o_{P_n}(1). \tag{5.2}$$

We will construct such an estimator as a one-step estimator, improving an initial estimator. As usual, the initial estimator will be a discretized and $n^{1/2}$-consistent estimator $\tilde{\vartheta}_n$, see Bickel et al. (1993). Such a discretized estimator can be treated as a deterministic sequence in the proof.

From Meyn and Tweedie (1993, Section 17.4) we obtain the following martingale approximation. It goes back to Gordin (1969); see also Gordin and Lifšic (1978). For $f \in L_2(\pi)$,

$$\sum_{i=1}^{n} [f(X_i) - \pi f] = \sum_{i=1}^{n} (Af)(X_{i-1}, X_i) + (Vf)(X_0) - (Vf)(X_n). \tag{5.3}$$

In particular,

$$n^{-1/2} \sum_{i=1}^{n} e_*(X_i) = n^{-1/2} \sum_{i=1}^{n} (Ae_*)(X_{i-1}, X_i) + o_{P_n}(1). \tag{5.4}$$

Our construction of the efficient estimator will therefore involve an estimator for $e_*$, and not for $Ae_*$.

We also rely on the sample splitting techniques of Schick (1998). For simplicity we use his *two-split*, which picks two blocks $\mathbf{X}_1 = (X_0, \ldots, X_{m_n})$ and $\mathbf{X}_2 = (X_{n-m_n}, \ldots, X_n)$. We need that

$$n - 2m_n \to \infty \quad \text{and } n^{-1/2}(n - 2m_n) \to 0.$$

With $e_n(x, \tilde{\vartheta}_n, X_0, \ldots X_n)$ denoting an estimator of $e_*(x)$, our estimator has the form

$$\begin{aligned} \hat{\vartheta}_n &= \frac{1}{2}\left(\tilde{\vartheta}_n + \frac{\frac{1}{m_n}\sum_{i=1}^{m_n} e_{m_n}(X_i, \tilde{\vartheta}_n, \mathbf{X}_2) - \pi_{\tilde{\vartheta}_n} e_{m_n}(\cdot, \tilde{\vartheta}_n, \mathbf{X}_2)}{\pi_{\tilde{\vartheta}_n}[e_{m_n}(\cdot, \tilde{\vartheta}_n, \mathbf{X}_2)\ell_{\tilde{\vartheta}_n}]}\right) \\ &\quad + \frac{1}{2}\left(\tilde{\vartheta}_n + \frac{\frac{1}{m_n}\sum_{i=n-m_n+1}^{n} e_{m_n}(X_i, \tilde{\vartheta}_n, \mathbf{X}_1) - \pi_{\tilde{\vartheta}_n} e_{m_n}(\cdot, \tilde{\vartheta}_n, \mathbf{X}_1)}{\pi_{\tilde{\vartheta}_n}[e_{m_n}(\cdot, \tilde{\vartheta}_n, \mathbf{X}_1)\ell_{\tilde{\vartheta}_n}]}\right). \end{aligned} \tag{5.5}$$

Call a sequence $\vartheta_n$ in $\Theta$ *local* if $n^{1/2}(\vartheta_n - \vartheta)$ is bounded.

**Theorem 3.** *Let Assumptions 1 and 3 hold. Suppose that for every local sequence $\vartheta_n$,*

$$\sup_x |e_n(x, \vartheta_n, X_0, \ldots, X_n)| = o_{P_n}(n^{1/2}), \tag{5.6}$$

$$\int \pi(dx)[e_n(x, \vartheta_n, X_0, \ldots, X_n) - e_*(x)]^2 = o_{P_n}(1). \tag{5.7}$$

*Then the one-step estimator $\hat{\vartheta}_n$ defined in (5.5) satisfies the stochastic expansion (5.2) and is therefore a least dispersed regular estimator for $\vartheta$.*

**Proof.** Since the initial estimator $\tilde{\vartheta}_n$ is discretized, it suffices to prove the stochastic expansion (5.4) with local sequences $\vartheta_n$ replacing $\tilde{\vartheta}_n$ in the definition (5.5) of $\hat{\vartheta}_n$. Fix a local sequence $\vartheta_n$. Because of the sample splitting, we only need to show expansion (5.4) for the 'estimator'

$$\vartheta_n + \frac{\frac{1}{n}\sum_{i=1}^{n} \tilde{e}_n(X_i) - \pi_{\vartheta_n}(\tilde{e}_n)}{\pi_{\vartheta_n}(\tilde{e}_n \ell_{\vartheta_n})}$$

with $\tilde{e}_n(x) = e_n(x, \vartheta_n, \tilde{\mathbf{X}})$ and $\tilde{\mathbf{X}}$ an independent copy of $(X_0, \ldots, X_n)$; see Schick (1998). It suffices to show that

$$\pi_{\vartheta_n}(\tilde{e}_n \ell_{\vartheta_n}) = \pi(e_* \ell_\vartheta) + o_{P_n}(1), \tag{5.8}$$

$$\pi_{\vartheta_n}(\tilde{e}_n) - \pi(\tilde{e}_n) = (\vartheta_n - \vartheta)\pi(e_* \ell_\vartheta) + o_{P_n}(n^{-1/2}), \tag{5.9}$$

$$n^{-1/2}\sum_{i=1}^{n}[\tilde{e}_n(X_i) - \pi(\tilde{e}_n)] = n^{-1/2}\sum_{i=1}^{n}(A\tilde{e}_n)(X_{i-1}, X_i) + o_{P_n}(1), \tag{5.10}$$

$$n^{-1/2}\sum_{i=1}^{n}(A\tilde{e}_n)(X_{i-1}, X_i) = n^{-1/2}\sum_{i=1}^{n}(Ae_*)(X_{i-1}, X_i) + o_{P_n}(1). \tag{5.11}$$

It follows from (5.6) and Hellinger differentiability at $\vartheta$ that

$$\mu[\tilde{e}_n(p_{\vartheta_n}^{1/2} - p_\vartheta^{1/2})]^2 = o_{P_n}(1). \tag{5.12}$$

We conclude from (5.7) and (5.12) that

$$\mu(\tilde{e}_n p_{\vartheta_n}^{1/2} - e_* p_\vartheta^{1/2})^2 \le 2\mu[\tilde{e}_n(p_{\vartheta_n}^{1/2} - p_\vartheta^{1/2})]^2 + 2\pi(\tilde{e}_n - e_*)^2 = o_{P_n}(1).$$

It follows from this and (5.1) that

$$\pi_{\vartheta_n}(\tilde{e}_n \ell_{\vartheta_n}) = \mu(\tilde{e}_n p_{\vartheta_n}^{1/2} \ell_{\vartheta_n} p_{\vartheta_n}^{1/2}) = \mu(e_* p_\vartheta^{1/2} \ell_\vartheta p_\vartheta^{1/2}) + o_{P_n}(1),$$

which yields (5.8). Similarly, one verifies

$$\mu(\tilde{e}_n \ell_\vartheta p_\vartheta) = \mu(e_* \ell_\vartheta p_\vartheta) + o_{P_n}(1).$$

Thus (5.9) follows if we show that

$$\mu\{\tilde{e}_n[p_{\vartheta_n} - p_\vartheta - (\vartheta_n - \vartheta)\ell_\vartheta p_\vartheta]\} = o_{P_n}(n^{-1/2}).$$

To see this write its left hand side as

$$\mu\{\tilde{e}_n(p_{\vartheta_n}^{1/2} + p_\vartheta^{1/2})[p_{\vartheta_n}^{1/2} - p_\vartheta^{1/2} - \frac{1}{2}(\vartheta_n - \vartheta)\ell_\vartheta p_\vartheta^{1/2}]\}$$
$$+\mu[\tilde{e}_n(p_{\vartheta_n}^{1/2} - p_\vartheta^{1/2})(\vartheta_n - \vartheta)\frac{1}{2}\ell_\vartheta p_\vartheta^{1/2}].$$

Now apply the Cauchy–Schwarz inequality to both terms and then use (5.12) and Hellinger differentiability at $\vartheta$ to conclude the desired result.

To prove relation (5.10), note first that by (5.7),

$$\pi[(V\tilde{e}_n)^2] \to \pi[(Ve_*)^2].$$

Hence, for $\varepsilon > 0$, the conditional Markov inequality yields

$$\max_i P\left(|(V\tilde{e}_n)(X_i)| > \varepsilon n^{1/2}\right) \le \mathrm{E}\left(\frac{\pi[(V\tilde{e}_n)^2]}{n\varepsilon^2} \wedge 1\right) \to 0.$$

Relation (5.10) now follows from the martingale approximation (5.3).

We verify relation (5.11) with the aid of Schick (1998, Theorem 3.3). We have

$$\int Q(x, dy)[(A\tilde{e}_n)(x, y) - (Ae_*)(x, y)] = 0$$

and

$$\pi \otimes Q\,(A\tilde{e}_n - Ae_*)^2 \le \|A\|^2 \cdot \pi(\tilde{e}_n - e_*)^2 \to 0$$

by (5.7) and since $A$ is a bounded operator. Then by Schick (1998, Remark 3.4) the conditions of his Theorem 3.3 hold, and (5.11) follows. □

**Remark 2.** Let $e_n$ be an estimator that satisfies condition (5.7) of Theorem 3. Then the estimator $\overline{e}_n = (-B_n) \vee e_n \wedge B_n$ satisfies (5.7) for every sequence of positive numbers $B_n$ tending to infinity. This truncated estimator also satisfies (5.6) if $B_n = o(n^{1/2})$. Consequently, only condition (5.7) poses any difficulties.

# 6 Estimation of $e_*$

The results of the previous section show that we can construct an efficient estimator of $\vartheta$ if one can construct an estimate $e_n$ of $e_*$ which satisfies (5.7). We shall now construct such an estimator $e_n$ under the assumption that we can choose appropriate orthonormal bases for the spaces $L_{2,0}(\pi_\tau)$. More precisely, for each $\tau \in \Theta$ let $\{\psi_{j,\tau} : j \geq 1\}$ be an orthonormal basis for $L_{2,0}(\pi_\tau)$. We require the following additional properties of these functions.

(A1) For every $j \geq 1$,

$$\mu(\psi_{j,\tau} p_\tau^{1/2} - \psi_{j,\vartheta} p_\vartheta^{1/2})^2 \to 0 \quad \text{for } \tau \to \vartheta.$$

(A2) There are positive numbers $\alpha$, $\beta$ and $C_1$ such that for all positive integers $k$ and all $\tau$ close to $\vartheta$,

$$\sum_{j=1}^{k} \|\psi_{j,\tau} - \psi_{j,\vartheta}\|^2 \leq C_1 k^\alpha |\tau - \vartheta|^{2\beta}.$$

(A3) There are positive numbers $\gamma$ and $C_2$ such that for all positive integers $k$ and all $\tau$ close to $\vartheta$,

$$\sum_{j=1}^{k} \pi_\tau(\psi_{j,\tau}^4) \leq C_2 k^\gamma.$$

**Remark 3.** Let us mention that such functions $\psi_{j,\tau}$ can easily be constructed if the state space is the real line $\mathbf{R}$ and the dominating measure $\mu$ is the Lebesgue measure. In this case, each $\pi_\tau$ possesses a continuous distribution function $F_\tau$. This allows us to choose $\psi_{j,\tau} = \phi_j \circ F_\tau$, where $\{\phi_j : j \geq 1\}$ is an orthonormal basis for $L_{2,0}(\lambda)$, with $\lambda$ the Lebesgue measure on $[0,1]$. We may choose the trigonometric basis

$$\begin{aligned} \phi_{2k-1}(x) &= \sqrt{2}\sin[2k\pi(x - 1/2)], \\ \phi_{2k}(x) &= \sqrt{2}\cos[2k\pi(x - 1/2)], \quad \text{for } 0 \leq x \leq 1 \text{ and } k \geq 1. \end{aligned}$$

Since $|\phi_j| \leq \sqrt{2}$ for all $j \geq 1$, condition (A3) holds with $C_2 = 4$ and $\gamma = 1$, while condition (A1) follows from Hellinger differentiability of the map $\tau \mapsto p_\tau$ at $\tau = \vartheta$. It follows from the Cauchy–Schwarz inequality and from Hellinger differentiability at $\vartheta$ that

$$|F_\tau(x) - F_\vartheta(x)| \leq \mu(|p_\tau - p_\vartheta|) \leq 2[\mu(|p_\tau^{1/2} - p_\vartheta^{1/2}|^2)]^{1/2} = O(|\tau - \vartheta|) \quad \text{as } \tau \to \vartheta.$$

Since $|\phi_j'| \leq \sqrt{8}\pi j$ for $j \geq 1$, condition (A2) holds with $\alpha = 3$ and $\beta = 1$.

Let $I_k$ denote the $k \times k$ identity matrix, and introduce vectors

$$\Psi_{k,\tau} = (\psi_{1,\tau}, \ldots, \psi_{k,\tau})^T \quad \text{and} \quad b_{k,\tau} = \pi_\tau(\ell_\tau \Psi_{k,\tau}).$$

**Theorem 4.** *Suppose that Assumptions 1 and 3 hold, and that conditions (A1) to (A3) are satisfied for certain $\alpha, \beta, \gamma$. Let $k_n$ be a sequence of positive integers such that*

$$k_n \to \infty, \quad k_n^{\alpha} n^{-\beta} \to 0, \quad k_n^{1+\gamma} n^{-1} \to 0.$$

*Then condition (5.7) holds for the sequence of estimators*

$$e_n(x, \vartheta_n, X_0, \ldots, X_n) = b_{k_n,\vartheta_n}^T (I_{k_n} - \hat{A}_n^T)(I_{k_n} - \hat{A}_n \hat{A}_n^T)^{-1}(I_{k_n} - \hat{A}_n)\Psi_{k,\vartheta_n}$$

*with*

$$\hat{A}_n = \frac{1}{n}\sum_{i=1}^{n} \Psi_{k_n,\vartheta_n}(X_j)\Psi_{k_n,\vartheta_n}^T(X_{j-1}).$$

To prove this theorem, we shall rely on the following two approximation results.

**Proposition 1.** *Let $\{\psi_j : j \geq 1\}$ be an orthonormal basis of $L_{2,0}(\pi)$. Let $\Gamma_k$ denote the projection in $L_{2,0}(\pi)$ onto the linear span of $\{\psi_1, \ldots, \psi_k\}$, and set $Q_k = \Gamma_k Q$, $\overline{Q}_k = \Gamma_k \overline{Q}$,*

$$e_{*,k} = \sum_{j=0}^{\infty} \Gamma_k (J - Q)(\overline{Q}_k Q_k)^j \Gamma_k (J - \overline{Q})\Gamma_k \ell_\vartheta, \quad k \geq 1.$$

*Then*

$$\|e_{*,k} - e_*\| \to 0 \quad \text{as } k \to \infty. \tag{6.1}$$

**Proof.** Keep in mind that $Q$ and $\overline{Q}$, viewed as operators on $L_{2,0}(\pi)$, have norms equal to $\|Q - \Pi\|$ which is less than 1 by Assumption 1, and that $\|a - \Gamma_k a\| \to 0$ as $k \to \infty$, for every $a \in L_{2,0}(\pi)$. Thus the desired result follows from the identity

$$\begin{aligned} e_* - e_{*,k} = (J - \Gamma_k)e_* \quad &+ \quad \sum_{j=1}^{\infty} \Gamma_k (J - Q)[(\overline{Q}Q)^j - (\overline{Q}_k Q_k)^j](J - \overline{Q})\ell_\vartheta \\ &+ \quad \sum_{j=0}^{\infty} \Gamma_k (J - Q)(\overline{Q}_k Q_k)^j (J - \Gamma_k)(J - \overline{Q})\ell_\vartheta \\ &+ \quad \sum_{j=0}^{\infty} \Gamma_k (J - Q)(\overline{Q}_k Q_k)^j \Gamma_k (J - \overline{Q})(J - \Gamma_k)\ell_\vartheta \end{aligned}$$

and the expansion

$$\sum_{j=1}^{\infty}[(\overline{Q}Q)^j - (\overline{Q}_k Q_k)^j] = \sum_{i=0}^{\infty}(\overline{Q}_k Q_k)^i[(\overline{Q} - \overline{Q}_k)Q + \overline{Q}_k(Q - Q_k)]\sum_{j=0}^{\infty}(\overline{Q}Q)^j,$$

valid on $L_{2,0}(\pi)$. Here we have used $C^j - D^j = \sum_{i=0}^{j-1} D^i(C-D)C^{j-i-1}$ for operators $C$ and $D$ and positive integers $j$. □

**Proposition 2.** *Suppose that conditions (A1) and (A2) hold for certain $\alpha, \beta$. For $k \geq 1$ and $\tau \in \Theta$ define*

$$e_{*,k,\tau} = b_{k,\tau}^T(I_k - A_{k,\tau}^T)(I_k - A_{k,\tau}A_{k,\tau}^T)^{-1}(I_k - A_{k,\tau})\Psi_{k,\tau}$$

*with $A_{k,\tau} = \pi(Q\Psi_{k,\tau} \cdot \Psi_{k,\tau}^T)$ and $Q\Psi_{k,\tau} = (Q\psi_{1,\tau}, \ldots, Q\psi_{k,\tau})^T$. Then*

$$\|e_{*,k_n,\vartheta_n} - e_*\| \to 0$$

*for all local sequences $\vartheta_n$, and for every sequence $k_n$ of positive integers satisfying*

$$k_n \to \infty, \quad k_n^\alpha n^{-\beta} \to 0. \tag{6.2}$$

**Proof.** Fix a sequence $k_n$ satisfying (6.2), and a local sequence $\vartheta_n$. We shall first show that

$$\|e_{*,k_n,\vartheta_n} - e_{*,k_n,\vartheta}\| \to 0. \tag{6.3}$$

Write $|v|_2$ for the Euclidean norm of a vector $v$, and $|M|_*$ for the spectral norm of a matrix $M$. Then $|M|_*^2$ is the largest eigenvalue of $M^TM$. It is now easy to see that (6.3) follows if we show that

$$\sup_k |b_{k,\vartheta}|_2 < \infty, \quad |b_{k_n,\vartheta_n} - b_{k_n,\vartheta}|_2 \to 0, \tag{6.4}$$

$$\sup_k |A_{k,\vartheta}|_* < 1, \quad |A_{k_n,\vartheta_n} - A_{k_n,\vartheta}|_* \to 0, \tag{6.5}$$

$$\pi(|\Psi_{k_n,\vartheta_n} - \Psi_{k_n,\vartheta}|_2^2) \to 0. \tag{6.6}$$

Of course, relation (6.6) follows from condition (A2) and assumption (6.2). From assumption (5.1) we obtain that $\pi_{\vartheta_n}(\ell_{\vartheta_n}^2) \to \pi_\vartheta(\ell_\vartheta^2)$. This is equivalent to

$$\sum_{j=1}^\infty [\pi_{\vartheta_n}(\ell_{\vartheta_n}\psi_{j,\vartheta_n})]^2 \to \sum_{j=1}^\infty [\pi(\ell_\vartheta\psi_{j,\vartheta})]^2.$$

Moreover, by condition (A1) and assumption (5.1) we have $\pi_{\vartheta_n}(\ell_{\vartheta_n}\psi_{j,\vartheta_n}) \to \pi_\vartheta(\ell_\vartheta\psi_{j,\vartheta})$ for all $j \geq 1$. Hence

$$\sum_{j=1}^\infty [\pi_{\vartheta_n}(\ell_{\vartheta_n}\psi_{j,\vartheta_n}) - \pi_\vartheta(\ell_\vartheta\psi_{j,\vartheta})]^2 \to 0.$$

Relation (6.4) follows from this and $|b_{k,\vartheta}|_2^2 \leq \pi(\ell_\vartheta^2)$. The spectral norm of a $k \times k$ matrix $M$ can be expressed as

$$|M|_* = \sup\{|u^TMv| : u, v \in \mathbf{R}^k, |u|_2 = |v|_2 = 1\}.$$

This representation is particularly helpful when dealing with a matrix of the form $M = \pi(Q\Phi \cdot \Psi^T)$ with $\Phi$ and $\Psi$ in $L^k_{2,0}(\pi)$. In this case one finds with the aid of the Cauchy–Schwarz inequality that

$$|u^T\pi(Q\Phi \cdot \Psi^T)v| = |\pi(Q(u^T\Phi) \cdot v^T\Psi)| \le \|Q - \Pi\| \, \|u^T\Phi\| \, \|v^T\Psi\|^2,$$
$$|u^T \left(\pi(Q\Phi \cdot \Phi^T) - \pi(Q\Psi \cdot \Psi^T)\right) v| \le \|u^T(\Phi - \Psi)\| \, \|v^T\Phi\| + \|u^T\Psi\| \, \|v^T(\Phi - \Psi)\|.$$

From the first inequality we obtain $|A_{k,\vartheta}|_* \le \|Q - \Pi| < 1$ and hence the first part of relation (6.5). The second inequality and relation (6.6) imply the second part of relation (6.5). This concludes the proof of (6.3).

Now we show that $e_{*,k,\vartheta}$ coincides with $e_{*,k}$ of the Proposition 1 if we take $\psi_j = \psi_{j,\vartheta}$. Indeed, for this choice of orthonormal basis, we find that $\Gamma_k \ell_\vartheta = b^T_{k,\vartheta}\Psi_{k,\vartheta}$, and that for each $a \in \mathbf{R}^k$ we have $Q_k(a^T\Psi_{k,\vartheta}) = a^T A_{k,\vartheta}\Psi_{k,\vartheta}$ and $\overline{Q}_k(a^T\Psi_{k,\vartheta}) = a^T A^T_{k,\vartheta}\Psi_{k,\vartheta}$. In the last step we have used the fact that $\pi(\overline{Q}\Psi_{k,\vartheta} \cdot \Psi^T_{k,\vartheta}) = \pi[\Psi_{k,\vartheta}(Q\Psi_{k,\vartheta})^T] = A^T_{k,\vartheta}$. Using the above and the fact that $|A_{k,\vartheta}|_* < 1$, it is now easy to see that

$$e_{*,k} = \sum_{j=0}^{\infty} b^T_{k,\vartheta}(I_k - A^T_{k,\vartheta})(A_{k,\vartheta}A^T_{k,\vartheta})^j(I_k - A_{k,\vartheta})\Psi_{k,\vartheta}.$$

This simplifies to $e_{*,k,\vartheta}$. Thus the desired result follows from (6.3) and Proposition 1.

□

**Proof of Theorem 4.** Note that $E(\hat{A}_n) = A_{k_n,\vartheta_n}$. It follows from the arguments in Proposition 2 that it suffices to show that $|\hat{A}_n - E(\hat{A}_n)|_* = o_{P_n}(1)$. We shall prove the stronger property $E|\hat{A}_n - E(\hat{A}_n)|_2^2 \to 0$.

By Assumption 1, there exists a finite constant $c$ such that for all $h \in L_2(\pi \otimes Q)$,

$$E\left\{\frac{1}{n}\sum_{i=1}^{n} h(X_{i-1}, X_i) - E[h(X_0, X_1)]\right\}^2 \le \frac{c}{n}E[h^2(X_0, X_1)].$$

From this, the Cauchy–Schwarz inequality, condition (A3) and the properties of $k_n$,

$$E|\hat{A}_n - E(\hat{A}_n)|_2^2 \le \frac{c}{n}\sum_{i=1}^{k_n}\sum_{j=1}^{k_n} E[\psi^2_{i,\vartheta_n}(X_1)\psi^2_{j,\vartheta_n}(X_0)] \le \frac{ck_n}{n}\sum_{i=1}^{k_n}\pi(\psi^4_{i,\vartheta_n}) \to 0.$$

□

**Remark 4.** In the reversible case we have $A_k = A^T_k$. This allows us to replace $\hat{A}_n$ by the symmetrized estimate $\frac{1}{2}(\hat{A}_n + \hat{A}^T_n)$.

# 7 Comparison with parametric results

Our results apply in particular to the situation in which we have a parametric model for the transition distributions, say $\{Q_\tau : \tau \in \Theta\}$. The estimator in Section 5 does not use the model except through the associated family of invariant distributions $\{\pi_\tau : \tau \in \Theta\}$. In this section, we compare known results for such parametric models with our results. For the sake of brevity, we keep the discussion heuristic and do not reproduce the regularity conditions given in the literature.

As before, we assume that $\Theta$ is one-dimensional and that the $\pi_\tau$ have positive $\mu$-densities $p_\tau$. Then $Q_\tau(x, dy)$ has as $\mu(dy)$-density, say $q_\tau(x, y)$, for ($\pi$-almost) all $x$ in the state space $S$.

Write $m_\tau(x, y) = \dot{q}_\tau(x, y)/q_\tau(x, y)$ for the logarithmic derivative, with respect to the parameter $\tau$, of $q_\tau(x, y)$. As before, we will omit the parameter if it equals the true parameter $\vartheta$. A perturbation of the transition distribution at $\tau = \vartheta$ is of the form

$$Q_{\vartheta+n^{-1/2}u}(x, dy) \doteq Q(x, dy)[1 + n^{-1/2}um(x, y)]. \tag{7.1}$$

Hence the local parameter space at $\vartheta$ is the linear span, say $H_{\rm par}$, of $m$. Here par stands for 'parametric'.

The perturbation expansion (3.8), applied to $Q_{nh} = Q_{\vartheta+n^{-1/2}u}$, with approximation (7.1), gives

$$n^{1/2}(\pi_{\vartheta+n^{-1/2}u}f - \pi f) \to u\pi(\overline{V}m \cdot f).$$

Comparing with (3.9), we obtain

$$\ell = \overline{V}m. \tag{7.2}$$

The canonical gradient for the parameter, viewed as a functional $t(Q_\tau) = \tau$ of the transition distribution, is of the form $g_{\rm par} = u_{\rm par}m$, with $u$ determined by (2.8),

$$n^{1/2}[t(Q_{\vartheta+n^{-1/2}u}) - t(Q)] = n^{1/2}(\vartheta + n^{-1/2}u - \vartheta) = u \stackrel{!}{=} u_{\rm par}u \cdot \pi \otimes Q\, m^2.$$

Hence

$$g_{\rm par} = (\pi \otimes Q\, m^2)^{-1}m,$$

and the lower bound for the asymptotic variance of regular estimators is

$$\pi \otimes Q\, g_{\rm par}^2 = (\pi \otimes Q\, m^2)^{-1}.$$

An efficient estimator for the parameter is the maximum likelihood estimator. It is a solution in $\tau$ of the estimating equation

$$\sum_{i=1}^{n} m_\tau(X_{i-1}, X_i) = 0.$$

Of course, the canonical gradient $g_{\mathrm{par}}$ is also obtained as projection onto $H_{\mathrm{par}}$ of the gradient $g_* = [\pi(\ell e_*)]^{-1} A e_*$, with $e_* = (\overline{V}A)^{-1}\ell$, which is canonical for the larger model $\mathcal{Q}_*$ of *all* transition distributions with invariant distribution in $\{\pi_\tau : \tau \in \Theta\}$; see Theorem 1. To show that $g_*$ projects to $g_{\mathrm{par}}$, we note first that (3.7) and (7.2) imply

$$\pi \otimes Q\,(m \cdot Ae_*) = \pi(\overline{V}m \cdot e_*) = \pi(\ell e_*),$$

so that $\pi \otimes Q\,(mg_*) = 1$. We also have $\pi \otimes Q\,(mg_{\mathrm{par}}) = 1$ and therefore $\pi \otimes Q\,[m(g_* - g_{\mathrm{par}})] = 0$, i.e. $g_{\mathrm{par}}$ is the projection of $g_*$ onto $H_{\mathrm{par}}$.

The last orthogonality property implies that, as expected, the asymptotic variance of the maximum likelihood estimator is never larger than that of the efficient estimator in the larger model $\mathcal{Q}_*$. The variance reduction can be considerable. An extreme case would be that the transition distributions $Q_\tau$ all have the same invariant distribution. Then the invariant distribution contains no information at all about the parameter. We had to exclude this case in Sections 3 to 6, through the assumption that $\pi\ell^2 > 0$.

The maximum likelihood estimator is only feasible if the transition distributions $Q_\tau$ are tractable. Kessler (2000) restricts attention to estimators which are solutions $\tau = \vartheta_n^f$ of estimating equations of the form

$$\sum_{i=1}^{n} f_\tau(X_i) = 0, \tag{7.3}$$

with $f_\tau \in L_{2,0}(\pi_\tau)$. A Taylor expansion shows that $\vartheta_n^f$ admits a stochastic expansion

$$n^{1/2}(\vartheta_n^f - \vartheta) = -(\pi\dot{f})^{-1} n^{-1/2} \sum_{i=1}^{n} f(X_i) + o_{P_n}(1). \tag{7.4}$$

For regularity conditions we refer to Sørensen (1998). Here $\dot{f}$ is the derivative, with respect to the parameter, of $f_\tau$ at $\tau = \vartheta$; we suppress the index $\vartheta$. Differentiating $\pi_\tau f_\tau = 0$ under the integral, we obtain $-\pi\dot{f} = \pi(\ell f)$. Together with the martingale approximation (5.3), we can write the stochastic expansion as

$$n^{1/2}(\vartheta_n^f - \vartheta) = [\pi(\ell f)]^{-1} n^{-1/2} \sum_{i=1}^{n} (Af)(X_{i-1}, X_i) + o_{P_n}(1). \tag{7.5}$$

Hence $\vartheta_n^f$ has asymptotic variance

$$[\pi(\ell f)]^{-2} \cdot \pi \otimes Q\,(Af)^2. \tag{7.6}$$

The asymptotic variance is minimized for the estimator $\vartheta_n^{e_*}$ obtained from the estimating equation (7.3) with

$$f = e_* = (\overline{V}A)^{-1}\ell.$$

To prove this, we note first that by (7.6) and (3.12), the asymptotic variance of $\vartheta_n^{e_*}$ is

$$[\pi(\ell e_*)]^{-2} \cdot \pi \otimes Q\,(Ae_*)^2 = [\pi(\ell e_*)]^{-1}.$$

Now write

$$\pi(\ell f) = \pi(\overline{V} Ae_* \cdot f) = \pi \otimes Q\,(Ae_* \cdot Af). \tag{7.7}$$

The Schwarz inequality and (3.12) give

$$[\pi(\ell f)]^2 \le \pi \otimes Q\,(Ae_*)^2 \cdot \pi \otimes Q\,(Af)^2 = \pi(\ell e_*) \cdot \pi \otimes Q\,(Af)^2.$$

We arrive at the inequality between the asymptotic variances of $\vartheta_n^f$ and $\vartheta_n^{e_*}$:

$$[\pi(\ell f)]^{-2} \cdot \pi \otimes Q\,(Af)^2 \ge [\pi(\ell e_*)]^{-1}.$$

A different characterization of the optimal influence function is given in Kessler (2000): The corresponding influence function is closest to the influence function $(\pi \otimes Q\,m^2)^{-1}m = g_{\text{par}}$ of the maximum likelihood estimator among all influence functions $[\pi(\ell f)]^{-1}Af$ of estimators $\vartheta_n^f$ with $f_\tau \in L_{2,0}(\pi_\tau)$. We have just shown that the optimal influence function is $Ae_*$. Indeed, $Ae_*$ is the projection of $m$ into the space $\{Af : f \in L_2(\pi)\}$. This follows from (7.7) and

$$\pi \otimes Q\,(m \cdot Af) = \pi(\overline{V} m \cdot f) = \pi(\ell f).$$

The estimating equations (7.3) contain the estimator which would be the maximum likelihood estimator if the observations were independent, the solution $\vartheta_n^\ell$ of

$$\sum_{i=1}^n \ell_\tau(X_i) = 0.$$

The asymptotic variance of $\vartheta_n^\ell$ is, using (3.7),

$$(\pi\ell^2)^{-2} \cdot \pi \otimes Q\,(A\ell)^2 = (\pi\ell^2)^{-2} \cdot \pi(\ell \cdot \overline{V} A\ell).$$

This is larger than the asymptotic variance of $\vartheta_n^{e_*}$ since by the Schwarz inequality and (3.12),

$$(\pi\ell^2)^2 = [\pi(\ell \cdot \overline{V} Ae_*)]^2 \le \pi \otimes Q\,(A\ell)^2 \cdot \pi \otimes Q\,(Ae_*)^2 = \pi \otimes Q\,(A\ell)^2 \cdot \pi(\ell e_*).$$

To calculate the maximum likelihood estimator, the logarithmic derivative $m_\tau$ of the transition distribution $Q_\tau$ must be tractable. To calculate the estimator $\vartheta_n^{e_*}$, the function $e_* = (\overline{V} A)^{-1}\ell$ must be tractable. The estimator $\vartheta_n^\ell$ requires only the logarithmic derivative $\ell_\tau$ of the invariant distribution $\pi_\tau$.

The estimator $\hat{\vartheta}_n$ introduced in Theorem 3 has the same asymptotic variance as $\vartheta_n^{e*}$. It does, however, not require knowledge of $Q_\tau$. Hence it is adaptive in the sense that whatever the model for the transition distributions, it is asymptotically as good as the estimator $\vartheta_n^{e*}$, which, in turn, is optimal among solutions of estimating equations $\sum_{i=1}^n f_\tau(X_i) = 0$ in the model $\{Q_\tau : \tau \in \Theta\}$. To put it differently: Even though $\vartheta_n^{e*}$ requires knowledge of $Q_\tau$, it does not exploit any of the information about $\tau$ in the model $\{Q_\tau : \tau \in \Theta\}$.

(Analogous results hold for quasi-likelihood models, which are defined by parametric models for the conditional mean and variance of a Markov chain. The maximum quasi-likelihood estimator requires knowledge of the conditional variance but does not extract any information from it. Furthermore, one can construct an estimator which is asymptotically as good but does not use the model for the conditional variance; see Wefelmeyer (1996). This estimator has thus an adaptivity property analogous to $\hat{\vartheta}_n$.)

# 8 Discretely observed diffusions

Consider a stationary version of the diffusion process $X$ defined by the stochastic differential equation

$$dX_t = b_\vartheta(X_t)dt + \sigma_\vartheta(X_t)dB_t, \tag{8.1}$$

where $B$ is Brownian motion. For simplicity, we assume again that $\vartheta$ is one-dimensional. Suppose we observe the process at $n$ equidistant time points $t_0 = 0, \ldots, t_n = n\Delta$. The observations $X_{t_1}, \ldots, X_{t_n}$ form a stationary and reversible Markov chain. Its transition distribution $Q_\vartheta(x, dy)$ is difficult to calculate, in general, but its invariant distribution $\pi_\vartheta(dx)$ is that of the diffusion process and can be given explicitly: The Lebesgue density of $\pi_\vartheta$ is

$$p_\vartheta(x) = [C_\vartheta s_\vartheta(x)\sigma_\vartheta(x)^2]^{-1}$$

with

$$s_\vartheta(x) = \exp\Big[-2\int_0^x \frac{b_\vartheta(y)}{\sigma_\vartheta(y)^2}dy\Big]$$

and norming constant

$$C_\vartheta = \int [s_\vartheta(x)\sigma_\vartheta(x)^2]^{-1}dx.$$

Estimation of $\vartheta$ was first studied for the case when $\Delta$ tends to zero with $n$ tending to infinity; see Le Breton (1976), Florens-Zmirou (1989), Jacod and Genon-Catalot (1993) and Kessler (1997). For comparison with our results we must assume that $\Delta$ is fixed. For fixed $\Delta$, a computer-intensive approximate maximum likelihood estimator based on numerical approximation of the transition density was developed in Pedersen (1995a), (1995b). By now there is a considerable literature on simpler, inefficient, estimators.

Here we restrict attention to estimators for which the asymptotic variance can be calculated explicitly. They are based on two types of estimating equations:

$$\sum_{i=1}^{n} f_\vartheta(X_{t_i}) = 0 \tag{8.2}$$

with $\pi_\vartheta f_\vartheta = 0$, see (7.3); and martingale estimating equations

$$\sum_{i=1}^{n} f_\vartheta(X_{t_{i-1}}, X_{t_i}) = 0 \tag{8.3}$$

with $Q_\vartheta f_\vartheta = 0$.

The first type was already discussed in Section 7 in the context of general parametric Markov chain models. For $f_\vartheta(x) = \ell_\vartheta(x) = \dot{p}_\vartheta(x)/p_\vartheta(x)$ we obtain what would be the maximum likelihood estimator if the observations were i.i.d. This estimator is not efficient. Kessler (2000) shows that the estimator $\vartheta_n^{e_*}$ based on $f_\vartheta = e_*$ defined in Theorem 2 is optimal among solutions of (8.2). If the diffusion model is correct, the estimator $\hat{\vartheta}_n$ introduced in Theorem 3 is asymptotically as good as $\vartheta_n^{e_*}$. By Theorem 2, the asymptotic variance of $\hat{\vartheta}_n$ (and hence of $\vartheta_n^{e_*}$) is $[\pi_\vartheta(\ell_\vartheta e_*)]^{-1}$ with

$$\pi_\vartheta(\ell_\vartheta e_*) = \pi_\vartheta \ell_\vartheta^2 + 2\sum_{j=1}^{\infty}(-1)^j \pi_\vartheta(\ell_\vartheta \cdot Q_\vartheta^j \ell_\vartheta).$$

However, $\vartheta_n^{e_*}$ depends on $Q_\vartheta$ through $\overline{V}A$. Hence, if the diffusion model is misspecified, then $\vartheta_n^{e_*}$ will, in general, be inconsistent, while our estimator remains consistent as long as the model for $\pi_\vartheta$ is correct.

For solutions $\vartheta_n^f$ of the second type of estimating equations, (8.3), we obtain a stochastic approximation similar to (7.4),

$$n^{1/2}(\vartheta_n^f - \vartheta) = (\pi_\vartheta \otimes Q_\vartheta \dot{f}_\vartheta)^{-1} n^{-1/2} \sum_{i=1}^{n} f_\vartheta(X_{t_{i-1}}, X_{t_i}) + o_{P_{n\vartheta}}(1). \tag{8.4}$$

Hence $\vartheta_n^f$ is asymptotically normal with variance

$$(\pi_\vartheta \otimes Q_\vartheta \dot{f}_\vartheta)^{-2} \pi_\vartheta \otimes Q_\vartheta f_\vartheta^2. \tag{8.5}$$

If we take $f_\vartheta(x, y) = m_\vartheta(x, y)$, the logarithmic derivative of the transition density, then the estimating equation (8.3) gives the maximum likelihood estimator, which is efficient if the diffusion model is correct, and in general better than our estimator $\hat{\vartheta}_n$. But as noted, $m_\vartheta$ is often not tractable.

The maximum likelihood estimator exploits the parametric diffusion model fully. Other choices of $f_\vartheta$ use less information about the model. A simple class of estimating

equations are the *quasi-likelihood* estimating equations, based on parametric models for certain conditional moments. They are also called *polynomial* estimating equations. The simplest is the *linear* estimating equation, with

$$f_\vartheta(x,y) = w_\vartheta(x)[y - a_\vartheta(x)], \tag{8.6}$$

where $w_\vartheta(x)$ is some weight function, and $a_\vartheta(x) = \int Q_\vartheta(x, dy)y$ is the conditional mean of $X_{t_i}$ given $X_{t_{i-1}} = x$. In many cases the conditional mean cannot be written explicitly and must be calculated numerically. This is, however, easier than calculating the maximum likelihood estimator. The asymptotic variance (8.5) with $f_\vartheta$ as in (8.6) is minimized for $w_\vartheta(x) = \dot{a}_\vartheta(x)/v_\vartheta(x)$, where $v_\vartheta(x) = \int Q_\vartheta(x, dy)[y - a_\vartheta(x)]^2$ is the conditional variance of $X_{t_i}$ given $X_{t_{i-1}} = x$. For this choice of $w_\vartheta$, the asymptotic variance is $[\pi_\vartheta(\dot{a}_\vartheta^2/v_\vartheta)]^{-1}$. We refer to Bibby and Sørensen (1995), and for generalizations to polynomial estimating equations to Kessler (1995) and Bibby and Sørensen (1996), and to the reviews of Bibby and Sørensen (1997) and Sørensen (1997). For quasi-likelihood models see also Wefelmeyer (1996).

Another class of martingale estimating equations is introduced by Kessler and Sørensen (1999). The generator of the diffusion process (8.1) is

$$L_\vartheta = \frac{1}{2}\sigma_\vartheta(x)^2\frac{d^2}{dx^2} + b_\vartheta(x)\frac{d}{dx}.$$

An eigenfunction $\varphi_\vartheta(x)$ with eigenvalue $\lambda_\vartheta$ solves $L_\vartheta\varphi_\vartheta(x) = -\lambda_\vartheta\varphi_\vartheta(x)$. We have $(Q_\vartheta\varphi_\vartheta)(x) = e^{-\lambda_\vartheta\Delta}\varphi_\vartheta(x)$ and obtain martingale estimating equations (8.3) with

$$f_\vartheta(x,y) = w_\vartheta(x)[\varphi_\vartheta(y) - e^{-\lambda_\vartheta\Delta}\varphi_\vartheta(x)]. \tag{8.7}$$

Kessler and Sørensen (1999) obtain optimal linear combinations of a finite number of such estimating equations.

From the form of the asymptotic variances we see that none of the estimators in this section, excepting the (intractable) maximum likelihood estimator, is always superior to our estimator. Some are simpler to calculate than ours, though none is straightforward. Unlike our estimator, they break down if certain features of the diffusion model are misspecified.

**Acknowledgment.** Mathieu Kessler and Wolfgang Wefelmeyer were supported in part by the Research Training Network "Statistical Methods for Dynamical Stochastic Models" under the program *Improving Human Potential*, financed by the The Fifth Framework Programme of the European Commission. Anton Schick was supported in part by NSF grant DMS-0072174. We thank the referees for suggestions which led in particular to Section 8 on discretely observed diffusions.

# References


Bibby, B. M. and Sørensen, M. (1995). Martingale estimating functions for discretely observed diffusion processes. *Bernoulli*, **1**, 17–39.

Bibby, B. M. and Sørensen, M. (1996). On estimation for discretely observed diffusions: a review. *Theory Stoch. Process.* **2 (18)**, 49–56.

Bibby, B. M. and Sørensen, M. (1997). A hyperbolic diffusion model for stock prices. *Finance Stoch.*, **1**, 25–41.

Bickel, P. J. , Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore.

Florens-Zmirou, D. (1989). Approximate discrete-time schemes for statistics of diffusion processes. *Statistics*, **20**, 547–557.

Genon-Catalot, V. and Jacod, J. (1993). On the estimation of the diffusion coefficient for multi-dimensional diffusion processes. *Ann. Inst. H. Poincare Probab. Statist.*, **29**, 119–151.

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. (1999). Reversible Markov chains and optimality of symmetrized empirical estimators. *Bernoulli*, **5**, 109–123.

Hájek, J. (1970). A characterization of limiting distributions of regular estimates. *Z. Wahrsch. Verw. Gebiete*, **14**, 323–330.

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.

Höpfner, R. (1993b). Asymptotic inference for Markov step processes: observation up to a random time. *Stochastic Process. Appl.*, **48**, 295–310.

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.

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.

Kartashov, N. V. (1996). *Strong Stable Markov Chains*. VSP, Utrecht.

Kessler, M. (1995). Martingale estimating functions for a Markov chain. Preprint.

Kessler, M. (1997). Estimation of an ergodic diffusion from discrete observations. *Scand. J. Statist.*, **24**, 211–229.

Kessler, M. (2000). Simple and explicit estimating functions for a discretely observed diffusion process. *Scand. J. Statist.*, **27**, 65–82.

Kessler, M. and Sørensen, M. (1999). Estimating equations based on eigenfunctions for a discretely observed diffusion process. *Bernoulli*, **5**, 299–314.

Le Breton, A. (1976). On continuous and discrete sampling for parameter estimation in diffusion type processes. *Math. Programming Stud.*, **5**, 124–144.

Meyn, S. P. and Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability.* Springer, London.

Pedersen, A. R. (1995a). A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scand. J. Statist.*, **22**, 55–71.

Pedersen, A. R. (1995b). Consistency and asymptotic normality of an approximate maximum likelihood estimator for discretely observed diffusion processes. *Bernoulli*, **1**, 257–279.

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

Penev, S. (1993). Stability of nonparametric procedures against dependence. *Theory Probab. Appl.*, **37**, 353–355.

Pfanzagl, J. and Wefelmeyer, W. (1982). *Contributions to a General Asymptotic Statistical Theory*, Lecture Notes in Statistics 13, Springer, New York.

Roussas, G. G. (1965). Asymptotic inference in Markov processes. *Ann. Math. Statist.*, **36**, 987–992.

Schick, A. (1998). Sample splitting with Markov chains. To appear in: *Bernoulli.*

Sørensen, M. (1997). Estimating functions for discretely observed diffusions: a review. In: *Selected Proceedings of the Symposium on Estimating Functions* (I. V. Basawa, V. P. Godambe and R. L. Taylor, eds.), 305–325, Lecture Notes — Monograph Series 32, Institute of Mathematical Statistics, Hayward, California.

Sørensen, M. (1998). On asymptotics of estimating functions. To appear in: *Brazil. J. Probab. Statist.*

Wefelmeyer, W. (1996). Quasi-likelihood models and optimal inference. *Ann. Statist.*, **24**, 405-422.