Some Developments in Semiparametric Statistics

Anton Schick, Department of Mathematical Sciences, Binghamton University, Binghamton, NY 13902-6000, USA (anton@math.binghamton.edu)

Wolfgang Wefelmeyer, Mathematical Institute, University of Cologne, Weyertal 86–90, 50931 Cologne, Germany (wefelm@math.uni-koeln.de)

Abstract: In this paper we describe the historical development of some parts of semiparametric statistics. The emphasis is on efficient estimation. We understand semiparametric model in the general sense of a model that is neither parametric nor nonparametric. We restrict attention to models with independent and identically distributed observations and to time series.

Keywords: Efficiency, convolution theorem, Newton–Raphson procedure, linear constraint, marginal constraint, symmetric location model, copula model, semiparametric regression, partly linear regression, nonparametric regression, conditional constraint, moving average, autoregression, linear process, prediction, quasi-likelihood model, estimating equation, local U-statistic.

1 Introduction

To begin consider the case of independent and identically distributed observations, with distribution P in some *model* $\mathcal{P}$, a family of distributions. If no structural assumptions are made on the distributions in $\mathcal{P}$, the model is called nonparametric. If the distributions depend smoothly on some finite-dimensional parameter, say $\mathcal{P} = \{P_\vartheta : \vartheta \in \Theta\}$, the model is called parametric. We call a model *semiparametric* in a wide sense if it is neither parametric nor nonparametric. This covers models with infinite-dimensional parameters and models described by constraints on the distributions. It also covers semiparametric models in the strict sense, with distributions $P_{\vartheta\gamma}$ having a finite-dimensional parameter ϑ and an infinite-dimensional parameter γ. The simple linear regression model $Y = \vartheta X + \varepsilon$ illustrates these cases. If ε and X are independent with densities f and g, respectively, then an observation (X, Y) has density $g(x)f(y-\vartheta x)$ with one-dimensional parameter ϑ and infinite-dimensional parameter $\gamma = (f, g)$. If ε and X are not assumed independent, then the model is described by the conditional constraint $E(Y|X) = \vartheta X$ which itself depends on an unknown parameter ϑ. For time series, autoregression $X_t = \vartheta X_{t-1} + \varepsilon_t$ provides completely analogous examples. If the innovations ε_t are i.i.d. with density f, then the time series is a Markov chain with transition density $f(y - \vartheta x)$ and parameter (ϑ, f). (Throughout the paper, by *Markov chain* we mean a discrete-time Markov process with arbitrary state space. We note that in the literature the term Markov chain is often reserved for a continuous-time Markov process with discrete state space.) If, besides the Markov property, we assume only that $E(\varepsilon_t|X_{t-1}) = 0$, then the time series is a Markov chain with conditional constraint $E(X_t|X_{t-1}) = \vartheta X_{t-1}$. Another analogy exists between multivariate i.i.d. models with constraints on the marginal distributions and time series with constraints on the stationary distribution of a single realization. This survey article emphasizes such analogies between models with independent observations and time series models. The focus is on efficient estimation. For results on optimal testing we refer to Choi, Hall and Schick (1996).

In Section 2 we briefly describe the development of an efficiency concept for differentiable functionals on general parametric, nonparametric and semiparametric models. It is based on local asymptotic normality of likelihoods and on the convolution theorem, and is due to Le Cam and Hájek.

In Section 3 we sketch the historical development of efficient estimation for some semiparametric models with i.i.d. observations. We consider in particular models with linear constraints and with constraints on the marginal distributions, the symmetric location model, copula models, conditional constraints, in particular quasi-likelihood models, and regression models with errors independent of the covariates.

In Section 4 we consider efficient estimation for Markov chain models with parametric marginals and with conditional constraints, ARMA models, general invertible linear processes, and nonlinear autoregression. The emphasis is on estimators that exploit the semiparametric structure of the model; we say little about *nonparametric* estimators in semiparametric models, e.g. least squares estimators in linear regression, or kernel estimators for linear processes.

We have restricted attention to simple semiparametric models. There is a large literature on more involved models. For censored longitudinal data see van der Laan and Robins (2003); for missing data see Tsiatis (2006); for measurement error in nonlinear models see Carroll, Ruppert, Stefanski, and Crainiceanu (2006); for hidden Markov models see Cappé, Moulines and Rydén (2005). An interesting collection of articles on semiparametric inference is in Fan and Koul (2006). More on inference for time series is found in Taniguchi and Kakizawa (2000), Fan and Yao (2003) and Gao (2007).

We do not treat inference for continuous-time processes. Monographs on counting processes are Jacobsen (1982), Fleming and Harrington (1991), Andersen, Borgan, Gill and Keiding (1993), Kalbfleisch and Prentice (2002); on diffusion processes, Kutoyants (2004); on semimartingales, Prakasa Rao (1999).

2 Asymptotic Variance Bounds

Consider a sequence $P_{n\vartheta}$ of models having a k-dimensional parameter ϑ. Fix ϑ and set $\vartheta_{nu} = \vartheta + n^{-1/2}u$, where $u \in \mathbb{R}^k$. Le Cam (1960) calls the sequence of models *locally asymptotically normal* at ϑ if there are random vectors Δ_n and a positive definite matrix J such that

$$\log \frac{dP_{n\vartheta_{nu}}}{dP_{n\vartheta}} = u^\top \Delta_n - \frac{1}{2} u^\top J u + o_{P_{n\vartheta}}(1), \quad u \in \mathbb{R}^k, \tag{2.1}$$

$$\Delta_n \Rightarrow J^{1/2} N_k \quad \text{under } P_{n\vartheta}, \tag{2.2}$$

where N_k denotes a k-dimensional standard normal random vector. As shown below, models with i.i.d. observations are often locally asymptotically normal. This also holds when the observations come from a homogeneous time series or continuous-time process. Typically, Δ_n equals $n^{-1/2}$ times the derivative at ϑ of the log-likelihood and forms a martingale so that (2.2) can be established via martingale central limit theorems.

An estimator $\hat\vartheta$ is called *regular* at ϑ with *limit* L if L is a random vector such that

$$n^{1/2}(\hat\vartheta - \vartheta_{nu}) \Rightarrow L \quad \text{under } P_{n\vartheta_{nu}}, \quad u \in \mathbb{R}^k.$$

The convolution theorem says that for such an estimator,

$$\Big(J^{-1}\Delta_n, n^{1/2}(\hat\vartheta - \vartheta) - J^{-1}\Delta_n\Big) \Rightarrow (J^{-1/2}N_k, M) \quad \text{under } P_{n\vartheta}, \tag{2.3}$$

with M independent of N_k. Different proofs are in Hájek (1970), Roussas (1972, following an idea of Bickel), and Le Cam (1972). For forerunners see Le Cam (1953), Kaufman (1966) and Inagaki (1970). Introductions to the theory are Fabian and Hannan (1985) and Strasser (1985).

In particular, L is distributed as the convolution $J^{-1/2}N_k + M$. This justifies calling $\hat\vartheta$ *efficient* at ϑ if $n^{1/2}(\hat\vartheta - \vartheta) \Rightarrow J^{-1/2}N_k$ under $P_{n\vartheta}$. By the convolution theorem, an estimator $\hat\vartheta$ is regular and efficient at ϑ if and only if

$$n^{1/2}(\hat\vartheta - \vartheta) = J^{-1}\Delta_n + o_{P_{n\vartheta}}(1).$$

The asymptotic covariance matrix of $J^{-1}\Delta_n$ is J^{-1}; we call it a (lower) *variance bound.* A possibly efficient estimator for ϑ is the maximum likelihood estimator. As noted by Le Cam (1960, 1974), it is often better not to use such a global maximizer of the likelihood function but instead a *one-step* or *Newton–Raphson* improvement of a $n^{1/2}$-consistent estimator $\bar\vartheta$,

$$\hat\vartheta = \bar\vartheta + n^{-1/2} J_{\bar\vartheta}^{-1} \Delta_{n\bar\vartheta}.$$

Here we have written $\Delta_{n\vartheta}$ for Δ_n and J_ϑ for J. It is technically convenient to use a *discretized* estimator $\bar\vartheta$, taking values on a grid with length of order $n^{-1/2}$.

These concepts and results carry over to infinite-dimensional parameter spaces, and to differentiable functionals of the parameter. Consider a sequence $P_{n\beta}$ of models, with β running through some set B. As observed by Stein (1956), in order to determine an asymptotic variance bound for real-valued functionals of β, it suffices to consider one-dimensional submodels β_{nt} with t in some closed linear space T which are locally asymptotically normal

$$\begin{aligned} \log \frac{dP_{n\beta_{nt}}}{dP_{n\beta}} &= \Delta_n(At) - \frac{1}{2}\|At\|^2 + o_{P_{n\beta}}(1), \quad t \in T, && (2.4)\\ \Delta_n(h) &\Rightarrow \|h\|N \quad \text{under } P_{n\beta}, \quad h \in H, && (2.5)\end{aligned}$$

with N a standard normal random variable and A a bounded linear operator from T into a Hilbert space H. Call a functional $\varphi : B \to \mathbb{R}$ *differentiable* at β with *gradient* g if g belongs to the closure of $AT = \{At : t \in T\}$ and

$$n^{1/2}(\varphi(\beta_{nt}) - \varphi(\beta)) \to (g, At), \quad t \in T.$$

An estimator $\hat\varphi$ is called *regular* for φ at β with *limit* L if

$$n^{1/2}(\hat\varphi - \varphi(\beta_{nt})) \Rightarrow L \quad \text{under } P_{n\beta_{nt}}, \quad t \in T.$$

It follows from the convolution theorem (2.3) that for such an estimator,

$$(\Delta_n(g), n^{1/2}(\hat\varphi - \varphi(\beta)) - \Delta_n(g)) \Rightarrow (\|g\|N, M) \quad \text{under } P_{n\beta}, \tag{2.6}$$

with M independent of N. In particular, L is distributed as the convolution $\|g\|N + M$. This justifies calling $\hat\varphi$ *efficient* at β if $n^{1/2}(\hat\varphi - \varphi(\beta)) \Rightarrow \|g\|N$ under $P_{n\beta}$. Again, an estimator $\hat\varphi$ is regular and efficient at β if and only if

$$n^{1/2}(\hat\varphi - \varphi(\beta)) = \Delta_n(g) + o_{P_{n\beta}}(1). \tag{2.7}$$

The asymptotic variance of $\Delta_n(g)$ is $\|g\|^2$; we call it a (lower) *variance bound* for regular estimators of φ at β.

Generalizations of the convolution theorems (2.3) and (2.6) to multivariate functionals φ are straightforward. One simply applies the above to the components of φ. This results in a version of (2.7) in which $\Delta_n(g)$ is replaced by a vector whose i-th component is $\Delta_n(g_i)$ with g_i the gradient of the i-th component of φ. Generalizations to functionals with values in Banach spaces are in Beran (1977), Begun, Hall, Huang and Wellner (1983), Millar (1985), Schick and Susarla (1990), van der Vaart (1991), and van der Vaart and Wellner (1996). For the construction of efficient estimators of functionals with values in Banach spaces we refer to Klaassen and Putter (2005).

Le Cam (1966, 1969) gives sufficient conditions for local asymptotic normality of models with i.i.d. observations. Let $X_1, \dots, X_n$ be i.i.d. with distribution P_ϑ having a one-dimensional parameter ϑ. Assume that $P_{\vartheta_{nu}}$ is *Hellinger differentiable* at ϑ with *derivative* ℓ in the sense that $E[\ell(X)] = 0$ and

$$\int \Big(n^{1/2}\big(\sqrt{dP_{\vartheta_{nu}}} - \sqrt{dP_\vartheta}\big) - \frac{1}{2} u \ell \sqrt{P_\vartheta} \Big)^2 \to 0.$$

The latter is short for

$$\int \left(n^{1/2}\Big(\sqrt{\frac{dP_{\vartheta_{nu}}}{d\nu}} - \sqrt{\frac{dP_\vartheta}{d\nu}}\Big) - \frac{1}{2} u \ell \sqrt{\frac{dP_\vartheta}{d\nu}} \right)^2 d\nu \to 0$$

for some measure ν dominating P_ϑ and the sequence $P_{\vartheta_{nu}}$. A Taylor expansion then shows that the model is locally asymptotically normal in the sense of (2.1), (2.2), with $\Delta_n = n^{-1/2} \sum_{j=1}^n \ell(X_j)$ and $J = E[\ell^2(X)]$ the *Fisher information* at ϑ. Le Cam (1984) proves that the converse also holds.

Consider now independent observations from a semiparametric model $P_{\vartheta\gamma}$ in the strict sense, with ϑ finite-dimensional and γ infinite-dimensional. Let ϑ_{nu} and γ_{nv} be sequences such that

$$\int \Big(n^{1/2}\big(\sqrt{dP_{\vartheta_{nu}\gamma_{nv}}} - \sqrt{dP_{\vartheta\gamma}}\big) - \frac{1}{2} (u^\top \lambda + Dv) \sqrt{dP_{\vartheta\gamma}} \Big)^2 \to 0$$

for some bounded linear operator D into $L_2(P_{\vartheta\gamma})$, and u and v running through closed linear spaces U and V, respectively. Then local asymptotic normality in the sense of (2.4), (2.5) holds with $T = U \times V$, $A(u,v) = u^\top \lambda + Dv$, $\beta = (\vartheta, \gamma)$, $\beta_{nt} = (\vartheta_{nu}, \gamma_{nv})$, and with

$$\begin{aligned} \Delta_n(A(u,v)) &= n^{-1/2} \sum_{j=1}^n (u^\top \lambda(X_j) + Dv(X_j)), \\ \|A(u,v)\|^2 &= E[(u^\top \lambda(X) + Dv(X))^2]. \end{aligned}$$

A gradient of a finite-dimensional differentiable functional φ of (ϑ, γ) is of the form $g = M\lambda + w$ for some matrix M and some vector w with components in the closure W of $DV = \{Dv : v \in V\}$. Let λ_W denote the vector whose i-th component is the projection of the i-th component of λ onto W. Then $\lambda_* = \lambda - \lambda_W$ is called the *efficient score function* for ϑ at (ϑ, γ), and $J_* = E[\lambda_*(X)\lambda_*^\top(X)]$ the *efficient information matrix* for ϑ. If this matrix is invertible, then the functional $\varphi(\vartheta, \gamma) = \vartheta$ is differentiable with gradient $J_*^{-1}\lambda_*$, and an efficient estimator $\hat\vartheta$ of ϑ is characterized by

$$\hat\vartheta = \vartheta + \frac{1}{n} \sum_{j=1}^n J_*^{-1} \lambda_*(X_j) + o_{P^n_{\vartheta\gamma}}(n^{-1/2}).$$

We call ϑ and γ *adaptive* if $u^\top \lambda$ and Dv are orthogonal for all u and v. Then the gradient for ϑ is $I^{-1}\lambda$, where $I = E[\lambda(X)\lambda^\top(X)]$ is the information for the model with γ known. This means that we should be able to estimate ϑ as well not knowing γ as knowing γ. Analogously, we should be able to estimate differentiable functionals of γ as well not knowing ϑ as knowing ϑ. In the literature, an efficient estimator in an adaptive model is also called *adaptive*. If we find an estimator for ϑ that does not depend on γ but attains the asymptotic variance bound in each model with γ known, then the model must be adaptive and the estimator must be efficient. Then it suffices to determine local asymptotic normality in each model with γ known.

Efficient estimators for ϑ can be constructed by an appropriate version of the one-step or Newton–Raphson procedure. More generally, call $\hat\vartheta$ *asymptotically linear* with *influence function* $f_{\vartheta\gamma}$ if $E[|f_{\vartheta\gamma}|^2(X)]$ is finite and $E[f_{\vartheta\gamma}(X)] = 0$, and

$$n^{1/2}(\hat\vartheta - \vartheta) = n^{-1/2}\sum_{j=1}^{n} f_{\vartheta\gamma}(X_j) + o_{P^n_{\vartheta\gamma}}(1).$$

If $\bar\vartheta$ is $n^{1/2}$-consistent (and discretized) and $\hat\gamma$ is an appropriate estimator of γ, we expect the one-step improved estimator

$$\hat\vartheta = \bar\vartheta + \frac{1}{n}\sum_{j=1}^{n} f_{\bar\vartheta\hat\gamma}(X_j)$$

to have influence function $f_{\vartheta\gamma}$. For adaptive models, Bickel (1982) proves this, splitting the sample in a large part for estimating ϑ and a small part for estimating γ. Schick (1986) and Klaassen (1987) give necessary and sufficient conditions in the general case, using a symmetrized sample splitting technique that works with parts of equal sizes. Schick (1987) gives necessary conditions for a construction that avoids sample splitting. See also Forrester, Hooper, Peng and Schick (2003) for an overview and simplifications. Achievability of the asymptotic variance bound in semiparametric models is discussed in Bickel and Ritov (1990).

Nonparametric models with i.i.d. observations are treated by parametrizing them with the underlying distribution P itself, and by introducing sequences P_{nw} which are Hellinger differentiable in the sense that

$$\int \Big(\sqrt{n}\Big(\sqrt{dP_{nw}} - \sqrt{dP}\Big) - \frac{1}{2}w\sqrt{P}\Big)^2 \to 0.$$

Then local asymptotic normality in the sense of (2.4), (2.5) holds with $\Delta_n(w) = n^{-1/2}\sum_{j=1}^{n} w(X_j)$ and $\|w\|^2 = E[w^2(X)]$. Constraints on P then translate into constraints on w.

Conditions for local asymptotic normality of time series have been given in many different cases. For Markov chains, a version of Hellinger differentiability of the transition distribution suffices. Parametric models are treated in Roussas (1965, 1970). For nonparametric models see Penev (1991). Markov step processes are considered in Höpfner, Jacod and Ladelli (1990), and Höpfner (1993). General sufficient conditions for local asymptotic normality for models with dependent data are in Jeganathan (1982) and Fabian and Hannan (1987). For Markov chains, one-step improvement of $n^{1/2}$-consistent estimators is studied in Schick (2001).

Markov chain models described by constraints on the *transition* distribution are best parametrized by the latter. On the other hand, for Markov chain models defined through constraints on the *stationary* distribution, it is more convenient to parametrize by the stationary distribution (of several observations), as shown by Bickel (1993) and Bickel and Kwon (2001). This leads, for example, to a simpler proof of the result of Greenwood and Wefelmeyer (1999) that the symmetrized empirical estimator $(2n)^{-1}\sum_{j=1}^n(f(X_{j-1},X_j)+f(X_{j-1},X_j))$ is efficient for $E[f(X_0,X_1)]$.

General introductions to efficient estimation in semiparametric and nonparametric models are Ibragimov and Has'minskii (1981), Pfanzagl and Wefelmeyer (1982), Le Cam (1986), Le Cam and Yang (1990), Pfanzagl (1990), Bickel, Klaassen, Ritov and Wellner (1998), and van der Vaart (1998, 2002).

3 Models With i.i.d. Observations

Linear constraints. Let $X_1,\dots,X_n$ be independent with distribution fulfilling the linear constraint $E[a(X)]=0$ for some vector-valued function a. Then the empirical estimator for the expectation $E[f(X)]$ of a function f can be modified as

$$\frac{1}{n}\sum_{j=1}^n f(X_j)-c^\top\frac{1}{n}\sum_{j=1}^n a(X_j).$$

By the Cauchy–Schwarz inequality, the asymptotic variance $E[(f(X)-c^\top a(X))^2]$ is minimized by $c=c_f=(E[a(X)a^\top(X)])^{-1}E[a(X)f(X)]$. Estimating c_f empirically, we arrive at the estimator

$$\frac{1}{n}\sum_{j=1}^n f(X_j)-\sum_{j=1}^n f(X_j)a^\top(X_j)\Big(\sum_{j=1}^n a(X_j)a^\top(X_j)\Big)^{-1}\frac{1}{n}\sum_{j=1}^n a(X_j). \tag{3.1}$$

It is efficient; see Koshevnik and Levit (1976).

An efficient estimator of $E[f(X)]$ is also obtained by weighting the empirical estimator, following the empirical likelihood approach of Owen (1988, 2001). This leads to the estimator $(1/n)\sum_{j=1}^n w_j f(X_j)$ with positive weights $w_j=1/(1+\lambda^\top a(X_j))$, where the vector λ is chosen such that $\sum_{j=1}^n w_j a(X_j)=0$. A computational disadvantage of empirical likelihood is that the weights w_j must be determined by the method of Lagrange multipliers, while the estimator (3.1) is given explicitly. On the other hand, for the weighted empirical distribution $\mathbb{P}_n=(1/n)\sum_{j=1}^n w_j\delta_{X_j}$, the linear constraint $\int a\,d\mathbb{P}_n=(1/n)\sum_{j=1}^n w_j a(X_j)=0$ holds exactly. This may be advantageous for small sample size. Empirical likelihood with infinitely many constraints is studied in Hjort, McKeague and Van Keilegom (2008).

These results extend to constraints $E[a_\vartheta(X)] = 0$ involving an unknown parameter ϑ. By the parametric plug-in principle, efficiency continues to hold if we use an efficient estimator for ϑ; see Müller and Wefelmeyer (2002a). As shown in Qin and Lawless (1994), the method of maximum empirical likelihood estimation provides efficient estimators for ϑ.

Constraints on marginals. Suppose we observe i.i.d. copies $(X_1, Y_1), \dots, (X_n, Y_n)$ of a random vector (X, Y). A natural estimator of an expectation $E[\psi(X, Y)]$ is the empirical estimator $(1/n)\sum_{j=1}^n \psi(X_j, Y_j)$. It is efficient if no structural information on the distribution of (X, Y) is available.

The empirical estimator can be improved if the marginal distributions are *known*. This is equivalent to saying that $E[a(X)]$ and $E[b(Y)]$ are known for all functions a and b. Hence the model is described by infinitely many linear constraints. If $E[a(X)] = E[b(Y)] = 0$, a new estimator for $E[\psi(X, Y)]$ is

$$\frac{1}{n}\sum_{j=1}^{n}(\psi(X_j, Y_j) - a(X_j) - b(Y_j)).$$

Bickel, Ritov and Wellner (1991) show that an efficient estimator is equivalent to the best estimator in the above class, which corresponds to the choices of $a = a_*$ and $b = b_*$ that minimize the variance. For contingency tables, Deming and Stephan (1940) use a modified chi-square method to improve the empirical estimator. To construct an efficient estimator Bickel, Ritov and Wellner (1991) adapt the method to general X and Y. Their construction relies on an appropriate partition of the state space of (X, Y). Peng and Schick (2002) use orthonormal bases and a least squares approach to estimate a_* and b_* directly.

Peng and Schick (2004a) assume parametric models for the marginals. Peng and Schick (2004b, 2005) construct efficient estimators for linear functionals in bivariate models with equal, but unknown marginals.

Symmetric location model. Let $X_1, \dots, X_n$ be independent and real-valued with density $f(\cdot - \vartheta)$, where f is symmetric about zero. The parameters f and ϑ are adaptive. Assume that f is absolutely continuous with finite Fisher information for location $I = \int \ell^2(x) f(x)\, dx$, where $\ell = -f'/f$. The efficient influence function for ϑ is $I^{-1}\ell(x - \vartheta)$. Efficient estimators for ϑ are constructed in particular by van Eeden (1970), Fabian (1974), Beran (1974, 1978), Sacks (1975), Stone (1975), Bickel (1982), Schick (1987), Faraway (1992) and Jin (1992).

Copula models. Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be i.i.d. copies of a random vector (X, Y) with joint distribution function H and marginal distribution functions F and G. Sklar (1959) proves that there exists a *copula* C such that $H(x, y) = C(F(x), G(y))$. If H has a density h, then C is uniquely determined, and the density h of H is of the form

$$h(x, y) = \varphi(F(x), G(y)) f(x) g(y),$$

where f and g are the densities of F and G, respectively, and $\varphi(u,v) = \partial_u \partial_v C(u,v)$. For an introduction see Nelsen (2006).

A *copula model* is given by a parametric family C_ϑ of copulas. This is a constraint on the joint distribution of (X,Y) that involves an unknown parameter ϑ. Estimators of ϑ are considered in Genest, Ghoudi and Rivest (1995) and Tsukahara (2005). Efficiency questions are studied by Klaassen and Wellner (1997) and Genest and Werker (2002). Semiparametric density estimators for copula models are introduced in Biau and Wegkamp (2005) and Liebscher (2005).

Conditional constraints. Suppose we observe independent copies of a vector (X,Y) satisfying the conditional constraint $E(a_\vartheta(X,Y)|X) = 0$, where a_ϑ is a vector of functions depending on an unknown parameter vector ϑ. This covers *quasi-likelihood models*, with real-valued Y and constraints for the conditional means and variances,

$$E(Y|X) = r_\vartheta(X), \quad E((Y - r_\vartheta(X))^2|X) = s_\vartheta^2(X).$$

A quasi-likelihood model can be written as a nonlinear and heteroscedastic regression model $Y = r_\vartheta(X) + s_\vartheta(X)\varepsilon$ with $E(\varepsilon|X) = 0$ and $E(\varepsilon^2|X) = 1$. Most of the literature, in particular in econometrics, refers to the autoregressive versions of these models, which we discuss in Section 3, in the subsection on Markov chains with conditional constraints. Versions of these results for regression are in Chamberlain (1987, 1992).

Regression with independent error and covariate. Suppose we observe independent copies $(X_1, Y_1), \dots, (X_n, Y_n)$ of a vector (X,Y), where the real-valued response Y depends on the covariate X through

$$Y = r_{\vartheta\gamma}(X) + \varepsilon,$$

where ε and X are independent and ε has mean zero, finite variance, and density f. Here ϑ is finite-dimensional and γ is arbitrary. The simplest such model is the *linear* regression model $Y = \mu + \vartheta^\top X + \varepsilon$; the classical estimators for μ and ϑ are the least squares estimators. Efficient estimators of ϑ have been constructed in Bickel (1982) and of (μ, ϑ) by Schick (1987). For symmetric errors, efficient estimators of (μ, ϑ) are obtained by Bickel (1982), Koul and Susarla (1983), and Schick (1987). In the *nonlinear* regression model $Y = r_\vartheta(X) + \varepsilon$, the parameter ϑ can again be estimated by a least squares estimator. Empirical processes of residuals $\hat\varepsilon_j = Y_j - r_{\hat\vartheta}(X_j)$ are studied by Koul (1970) and Loynes (1980), among others; see also Koul (2002).

The *partly linear* regression model is $Y = \vartheta^\top U + r(V) + \varepsilon$ with covariate $X = (U,V)$. Estimators of ϑ are studied by Engle, Granger, Rice and Weiss (1986), Chen (1988), Robinson (1988) and Cuzick (1992a), among others. Cuzick (1992b), Schick (1993, 1996), and Forrester, Hooper, Peng and Schick (2003) construct efficient estimators of ϑ; Bhattacharya and Zhao (1997) do so for symmetric errors. The empirical distribution function based on residuals $\hat\varepsilon_j = Y_j - \hat\vartheta^\top U_j - \hat r(V_j)$ is shown to be efficient in Müller, Schick and Wefelmeyer (2007) if $\hat{\vartheta}$ is an efficient estimator of ϑ and $\hat{r}$ is an appropriately chosen linear smoother.

The *nonparametric* regression model $Y = r(X) + \varepsilon$ with unknown smooth regression function r is semiparametric in our sense because ε and X are assumed independent. The regression function can be estimated only nonparametrically, for example by a Nadaraya–Watson estimator $\hat{r}$. Empirical estimators based on residuals $\hat{\varepsilon}_j = Y_j - \hat{r}(X_j)$ are studied by Akritas and Van Keilegom (2001) for heteroscedastic nonparametric regression and by Müller, Schick and Wefelmeyer (2007) for homoscedastic nonparametric regression.

General procedures for constructing efficient estimators in semiparametric regression models are described by Schick (1993, 1994) and by Forrester, Hooper, Peng and Schick (2003). We refer to Müller, Schick and Wefelmeyer (2004) for a comparison with regression models defined by conditional constraints, i.e. with ε and X not necessarily independent, that were considered in the previous subsection.

4 Time Series

Markov chains with parametric marginals. Let $X_0, \dots, X_n$ be observations from a geometrically ergodic first-order Markov chain with unknown transition distribution $Q(x, dy)$. Suppose we have a parametric model $\pi_\vartheta(dx)$ for the stationary distribution. Kessler, Schick and Wefelmeyer (2001) give efficient estimators for ϑ. Penev, Peng, Schick and Wefelmeyer (2004) construct efficient estimators for linear functionals $E[\psi(X_0, X_1)]$ of the joint stationary distribution. They are obtained similarly as in the i.i.d. case, Peng and Schick (2004a).

Markov chains with conditional constraints. Let $X_{1-p}, \dots, X_n$ be observations of a Markov chain of order p satisfying the linear constraint $E(a_\vartheta(\mathbf{X}_{t-1}, X_t)|\mathbf{X}_{t-1}) = 0$, where a_ϑ is a m-dimensional vector of known functions depending on an unknown k-dimensional parameter ϑ. This is analogous to the i.i.d. models with conditional constraints considered in Section 3 and covers *quasi-likelihood models* with real-valued state space and constraints for the conditional means and variances,

$$E(X_t|\mathbf{X}_{t-1}) = r_\vartheta(\mathbf{X}_{t-1}), \quad E((X_t - r_\vartheta(\mathbf{X}_{t-1}))^2|\mathbf{X}_{t-1}) = s_\vartheta^2(\mathbf{X}_{t-1}).$$

A quasi-likelihood model can be written as a nonlinear and heteroscedastic autoregression model $X_t = r_\vartheta(\mathbf{X}_{t-1}) + s_\vartheta(\mathbf{X}_{t-1})\varepsilon$ with $E(\varepsilon_t|\mathbf{X}_{t-1}) = 0$ and $E(\varepsilon_t^2|\mathbf{X}_{t-1}) = 1$.

Hansen (1982, 1985) suggests estimating ϑ by the *generalized method of moments*, a minimizer $\hat{\vartheta}$ of

$$\sum_{j=1}^n a_\vartheta^\top(\mathbf{X}_{j-1}, X_j) W_\vartheta(\mathbf{X}_{j-1}) M_n \sum_{j=1}^n W_\vartheta^\top(\mathbf{X}_{j-1}) a_\vartheta(\mathbf{X}_{j-1}, X_j),$$

where M_n is a random symmetric $k \times k$ matrix converging to a fixed deterministic matrix and W_ϑ is a $m \times k$ matrix of weight functions. The optimal weights W_ϑ^* are determined by minimizing the asymptotic covariance matrix of $\hat{\vartheta}$ and of the form

$$W_\vartheta^*(\mathbf{X}_{j-1}) = E\big(a_\vartheta(\mathbf{X}_{j-1}, X_j)a_\vartheta^\top(\mathbf{X}_{j-1}, X_j)|\mathbf{X}_{t-1}\big)^{-1}E\big(\dot{a}_\vartheta(\mathbf{X}_{j-1}, X_j)|\mathbf{X}_{t-1}\big),$$

where $\dot{a}_\vartheta(\mathbf{x}, y)$ is the $m \times k$ matrix of partial derivatives of $a_\vartheta(\mathbf{x}, y)$ with respect to ϑ. Another estimator of ϑ is obtained as a solution of the estimating equation

$$\sum_{j=1}^n W_\vartheta^\top(\mathbf{X}_{j-1})a_\vartheta(\mathbf{X}_{j-1}, X_j) = 0.$$

The optimal weights are again W_ϑ^*. The weights depend on the unknown transition distribution of the Markov chain and must be estimated, say by Nadaraya–Watson estimators and some initial estimator for ϑ. The resulting estimators are efficient; see Müller and Wefelmeyer (2002b). Efficient estimation for quasi-likelihood models is treated in Wefelmeyer (1996). Reviews of the generalized method of moments are Newey and McFadden (1994). Estimating equations for general models are studied in Heyde (1997).

ARMA models. Let $X_{1-p}, \dots, X_n$ be observations of an ergodic *ARMA(p,q) process* satisfying

$$X_t - \varrho_1 X_{t-1} + \dots + \varrho_p X_{t-p} = \varepsilon_t + \varphi_1\varepsilon_{t-1} + \dots + \varphi_q\varepsilon_{t-q},$$

where ε_t are i.i.d. innovations with mean zero, finite variance, and density f. For $q = 0$ this is an *AR(p) process*, a Markov chain of order p. For $p = 0$ we have an *MA(q) process*, which is not Markov. Least squares estimators for the autoregressive parameters $\boldsymbol{\varrho} = (\varrho_1, \dots, \varrho_p)$ and for the moving average parameters $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_q)$ are not efficient, in general. For symmetric f, the ARMA(p,q) model is adaptive for $(\boldsymbol{\varrho}, \boldsymbol{\varphi})$. Kreiss (1987a) proves local asymptotic normality for fixed f and constructs efficient one-step estimators for $(\boldsymbol{\varrho}, \boldsymbol{\varphi})$. More general results on local asymptotic normality and generalizations are in Jeganathan (1995).

For mean zero innovations, the AR(p) model is adaptive for $\boldsymbol{\varrho}$. Akritas and Johnson (1982) and Kreiss (1987b) prove local asymptotic normality for fixed f and for unknown f, respectively; Kreiss (1987b) constructs efficient one-step estimators for $\boldsymbol{\varrho}$.

Nonparametric (kernel) estimators for the stationary density of time series are well-studied. The semiparametric structure of autoregressive time series with independent innovations can be exploited to obtain better estimators. For the MA(1) process $X_t = \varepsilon_t + \varrho\varepsilon_{t-1}$, the stationary density g has the convolution representation $g(x) = \int f(x - \varrho y)f(y)\,dy$. Saavedra and Cao (1999) estimate f by a kernel estimator based on residuals $\hat{\varepsilon}_j$ and show that the plug-in estimator $\hat{g}(x) = \int \hat{f}(x - \hat{\varrho}y)\hat{f}(y)\,dy$

has rate $n^{-1/2}$. Schick and Wefelmeyer (2004a) prove this for the closely related residual-based local U-statistic

$$\hat{g}(x) = \frac{1}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} k_b(x - \hat{\varepsilon}_i - \hat{\varrho}\hat{\varepsilon}_j),$$

where k is a kernel, $k_b(x) = k(x/b)/b$, and b is a bandwidth. The estimator is motivated by density estimators for functions of at least two independent innovations, introduced by Frees (1994); for a general recent result see Giné and Mason (2007). Functional central limit theorems for residual-based local U-statistics estimating the stationary density in MA(p) models are in Schick and Wefelmeyer (2004a).

Linear processes. Consider a *linear process* described by an infinite-order moving average representation

$$X_t = \varepsilon_t + \sum_{s=1}^{\infty} \varphi_s \varepsilon_{t-s}$$

with summable coefficients φ_s and i.i.d. innovations ε_t with mean zero, finite variance, and density f. Suppose the linear process is *invertible.* This means that the observations have an infinite-order autoregressive representation,

$$\varepsilon_t = X_t - \sum_{s=1}^{\infty} \varrho_s X_{t-s}.$$

For the case that the coefficients $\varphi = \varphi(\vartheta)$ depend on an infinite-dimensional parameter ϑ, Kreiss (1990) proves local asymptotic normality for fixed f. Schick and Wefelmeyer (2002b) construct efficient estimators for a finite-dimensional parameter ϑ. Boldin (1982) and Kreiss (1991) estimate linear functionals of the innovation distribution by empirical estimators based on residuals $\hat{\varepsilon}_j$ obtained from the autoregressive representation; Schick and Wefelmeyer (2002b) describe efficient versions that use the linear constraint $E[\varepsilon] = 0$ on the innovation distribution. Schick and Wefelmeyer (2004b) estimate linear functionals of the stationary distribution by residual-based U-statistics. Robinson (1987) studies residual-based estimators for the innovation density.

Kernel density estimators for the stationary density of linear processes are well-studied. Similarly as for first-order moving average processes, one can obtain $n^{1/2}$-consistent estimators for the stationary density h of X_t through the convolution representation $X_t = \varepsilon_t + Y_t$ with $Y_t = \sum_{s=1}^{\infty} \varphi_s \varepsilon_{t-s}$. Estimators $\hat{\varepsilon}_j$ for the innovations can again be constructed using the autoregressive representation of the process; the innovation density f can be estimated by a kernel estimator based on these residuals. The density g of Y_t can be estimated by a kernel estimator based on $\hat{Y}_j = X_j - \hat{\varepsilon}_j$. Schick and Wefelmeyer (2007b) prove that the convolution estimator $\hat{h}(x) = \int \hat{f}(x-y)\hat{g}(y)\,dy$ is uniformly $n^{1/2}$-consistent and that $n^{1/2}(\hat{h} - h)$ converges weakly in C_0 to a centered

Gaussian process; Schick and Wefelmeyer (2008b) show analogous results in weighted L_1 spaces.

Nonlinear autoregression. A *nonlinear autoregressive process* of order p is given by

$$X_t = r_\vartheta(\mathbf{X}_{t-1}) + \varepsilon_t,$$

where ϑ is a finite-dimensional parameter, $\mathbf{X}_{t-1} = (X_{t-p}, \dots, X_{t-1})$, and ε_t are i.i.d. innovations with mean zero, finite variance, and density f. The parameter ϑ can be estimated by least squares estimators; efficient estimators are obtained by one-step improvement, similarly as in nonlinear regression; see Koul and Schick (1997).

The innovations can be estimated by residuals $\hat\vartheta = X_t - r_{\hat\vartheta}(\mathbf{X}_{t-1})$. Liebscher (1999) studies residual-based kernel estimators of the innovation density; weighted versions are treated in Müller, Schick and Wefelmeyer (2005). Efficient weighted residual-based empirical estimators for linear functionals of the innovation distribution are in Schick and Wefelmeyer (2002a).

Conditional expectations $E(q(X_{n+1})|\mathbf{X}_n = \mathbf{x})$, with $\mathbf{x} = (x_1, \dots, x_p)$, are usually estimated by kernel estimators. In a nonlinear autoregressive model, such a conditional expectation can be written as an *unconditional* expectation $E[q(\varepsilon - r_\vartheta(\mathbf{x}))]$ and can be estimated by the residual-based empirical estimator

$$\frac{1}{n}\sum_{j=1}^n q(\hat\varepsilon_j - r_{\hat\vartheta}(\mathbf{x})).$$

Conditional expectations with higher-order lags can be estimated by residual-based von Mises statistics. For example, a conditional expectation with lag two can be written

$$E(q(X_{n+2}) \mid \mathbf{X}_n = \mathbf{x}) = E[q(\varepsilon_2 + r_\vartheta(\varepsilon_1 + r_\vartheta(\mathbf{x})))]$$

and can be estimated by the residual-based von Mises statistic

$$\frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n q(\hat\varepsilon_j - r_{\hat\vartheta}(\hat\varepsilon_i - r_{\hat\vartheta}(\mathbf{x}))).$$

Müller, Schick and Wefelmeyer (2006) construct smoothed and weighted versions of such estimators that are efficient. Analogous results for moving average processes and invertible linear processes are in Schick and Wefelmeyer (2008a) and (2007a).

Other autoregressive models. Every regression model has an autoregressive counterpart, with analogous results, but the proofs are more involved for the autoregressive versions since now the observations are dependent. Efficient estimators for ϑ in a *partly linear autoregressive process* $X_t = \vartheta X_{t-1} + r(X_{t-2}) + \varepsilon_t$ are constructed in Schick (1999).

The *nonparametric autoregressive process* $X_t = r(\mathbf{X}_{t-1}) + \varepsilon_t$ was introduced by Jones (1978). Grama and Neumann (2006) show that the nonparametric autoregressive model $X_t = r(X_{t-1}) + \varepsilon_t$ is (locally) asymptotically equivalent, in the sense of Le Cam's deficiency distance, to certain nonparametric regression models. For Nadaraya–Watson estimators of the autoregression function r we refer to Robinson (1983), Tjøstheim (1994) and Masry (2005); for local polynomial smoothers see Masry (1996) and Kreiss and Neumann (1998). Functionals of the innovation distribution can be estimated by empirical estimators based on residuals $\hat\varepsilon_j = X_j - \hat r(X_{j-1})$. Cheng and Tong (1993) estimate the innovation variance; Müller, Schick and Wefelmeyer (2008) estimate the innovation distribution function.

References

Akritas, M.G. and Johnson, R.A. (1982). Efficiencies of tests and estimators for p-order autoregressive processes when the error distribution is nonnormal. *Ann. Inst. Statist. Math.*, 34, 579–589.

Akritas, M.G. and Van Keilegom, I. (2001). Non-parametric estimation of the residual distribution. *Scand. J. Statist.*, 28, 549–567.

Andersen, P.K., Borgan, Ø., Gill, R.D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes.* Springer Series in Statistics, Springer, New York.

Begun, J.M., Hall, W.J., Huang, W.-M. and Wellner, J.A. (1983). Information and asymptotic efficiency in parametric–nonparametric models. *Ann. Statist.*, 11, 432–452.

Beran, R. (1974). Asymptotically efficient adaptive rank estimates in location models. *Ann. Statist.*, 2, 63–74.

Beran, R. (1977). Estimating a distribution function. *Ann. Statist.*, 5, 400–404.

Beran, R. (1978). An efficient and robust adaptive estimator of location. *Ann. Statist.*, 6, 292–313.

Bhattacharya, P.K. and Zhao, P.-L. (1997). Semiparametric inference in a partial linear model. *Ann. Statist.*, 25, 244–262.

Biau, G. and Wegkamp, M. (2005). A note on minimum distance estimation of copula densities. *Statist. Probab. Lett.*, 73, 105–114.

Bickel, P.J. (1982). On adaptive estimation. *Ann. Statist.*, 10, 647–671.

Bickel, P.J. (1993). Estimation in semiparametric models. In: *Multivariate Analysis: Future Directions* (C.R. Rao, ed.), 55–73, North-Holland, Amsterdam.

Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models.* Springer, New York.

Bickel, P.J. and Kwon, J. (2001). Inference for semiparametric models: Some questions and an answer (with discussion). *Statist. Sinica*, 11, 863–960.

Bickel, P.J., Ritov, Y. and Wellner, J.A. (1991). Efficient estimation of linear functionals of a probability measure P with known marginal distributions. *Ann. Statist.*, 19, 1316–1346.

Boldin, M.V. (1982). Estimation of the distribution of noise in an autoregressive scheme. *Theory Probab. Appl.*, 27, 866–871.

Cappé, O., Moulines, E. and Rydén, T. (2005). *Inference in Hidden Markov Models.* Springer Series in Statistics, Springer, New York.

Carroll, R.J., Ruppert, D., Stefanski, L.A. and Crainiceanu, C.M. (2006). *Measurement Error in Nonlinear Models. A Modern Perspective.* 2nd ed. Monographs on Statistics and Applied Probability 105, Chapman & Hall/CRC, Boca Raton.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *J. Econometrics*, 34, 305–334.

Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica*, 60, 567–596.

Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *Ann. Statist.*, 16, 136–146.

Cheng, B. and Tong, H. (1993). On residual sums of squares in nonparametric autoregression. *Stochastic Process. Appl.*, 48, 157–174.

Choi, S., Hall, W.J. and Schick, A. (1996). Asymptotically uniformly most powerful tests in parametric and semiparametric models. *Ann. Statist.*, 24, 841–861.

Cuzick, J. (1992a). Semiparametric additive regression. *J. Roy. Statist. Soc. Ser. B*, 54, 831–843.

Cuzick, J. (1992b). Efficient estimates in semiparametric additive regression models with unknown error distribution. *Ann. Statist.*, 20, 1129–1136.

Deming, W.E. and Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statistics*, 11, 427–444.

Engle, R.F., Granger, C.W.J., Rice, J. and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity. *J. Amer. Statist. Assoc.*, 81, 310–320.

Fabian, V. (1974). Asymptotically efficient stochastic approximation; the RM case. *Ann. Statist.* 1, 486–495.

Fabian, V. and Hannan, J. (1985). *Introduction to Probability and Mathematical Statistics.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, Wiley, New York.

Fabian, V. and Hannan, J. (1987). Local asymptotic behavior of densities. *Statist. Decisions*, 5, 105–138. Correction 6 (1988), 195.

Fan, J. and Koul, H., eds. (2006). *Frontiers in Statistics.* Imperial College Press, London.

Fan, J. and Yao, Q. (2003). *Nonlinear Time Series. Nonparametric and Parametric Methods.* Springer Series in Statistics, Springer, New York.

Faraway, J.J. (1992). Smoothing in adaptive estimation. *Ann. Statist.*, 20, 414–427.

Fleming, T.R. and Harrington, D.P. (1991). *Counting Processes and Survival Analysis.* Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, Wiley, New York.

Forrester, J., Hooper, W., Peng, H. and Schick, A. (2003). On the construction of efficient estimators in semiparametric models. *Statist. Decisions*, 21, 109–137.

Frees, E.W. (1994). Estimating densities of functions of observations. *J. Amer. Statist. Assoc.*, 89, 517–525.

Gao, J. (2007). *Nonlinear Time Series. Semiparametric and Nonparametric Methods.* Monographs on Statistics and Applied Probability 108, Chapman & Hall/CRC, Boca Raton.

Genest, C., Ghoudi, K. and Rivest, L.-P. (1995). A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82, 543–552.

Genest, C. and Werker, B.J.M. (2002). Conditions for the asymptotic semiparametric efficiency of an omnibus estimator of dependence parameters in copula models. In: *Distributions With Given Marginals and Statistical Modelling* (C.M. Cuadras, J. Fortiana and J.A. Rodríguez Lallena, eds.), 103–112, Kluwer, Dordrecht.

Giné, E. and Mason, D.M. (2007). On local U-statistic processes and the estimation of densities of functions of several sample variables. *Ann. Statist.*, 35, 1105–1145.

Grama, I.G. and Neumann, M.H. (2006). Asymptotic equivalence of nonparametric autoregression and nonparametric regression. *Ann. Statist.*, 34, 1701–1732.

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.

Hansen, L.P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50, 1029–1054.

Hansen, L.P. (1985). A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators. *J. Econometrics*, 30, 203–238.

Heyde, C.C. (1997). *Quasi-Likelihood And Its Application. A General Approach to Optimal Parameter Estimation.* Springer Series in Statistics, Springer, New York.

Hjort, N.L., McKeague, I.W. and Van Keilegom, I. (2008). Extending the scope of empirical likelihood. To appear in: *Ann. Statist.*

Höpfner, R. (1993). 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., Jacod, J. and Ladelli, L. (1990). Local asymptotic normality and mixed normality for Markov statistical models. *Probab. Theory Related Fields*, 86, 105–129.

Ibragimov, I.A. and Has'minskii, R.Z. (1981). *Statistical Estimation. Asymptotic Theory.* Applications of Mathematics 16, Springer, New York.

Inagaki, N. (1970). On the limiting distribution of a sequence of estimators with uniformity property. *Ann. Inst. Statist. Math.*, 22, 1–13.

Jacobsen, M. (1982). *Statistical Analysis of Counting Processes.* Lecture Notes in Statistics 12, Springer, New York.

Jeganathan, P. (1982). On the asymptotic theory of estimation when the limit of the log-likelihood ratios is mixed normal. Sankhya Ser. A 44, 173–212.

Jeganathan, P. (1995) Some aspects of asymptotic theory with applications to time series models. *Econometric Theory*, 11, 818–887.

Jin, K. (1992). Empirical smoothing parameter selection in adaptive estimation. *Ann. Statist.*, 20, 1844–1874.

Jones, D.A. (1978). Nonlinear autoregressive processes. *Proc. Roy. Soc. London Ser. A*, 360, 71–95.

Kalbfleisch, J.D. and Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data.* 2nd ed. Wiley Series in Probability and Statistics, Wiley, Hoboken.

Kaufman, S. (1966). Asymptotic efficiency of the maximum likelihood estimator. *Ann. Inst. Statist. Math.*, 18, 155–178.

Kessler, M., Schick, A. and Wefelmeyer, W. (2001). The information in the marginal law of a Markov chain. *Bernoulli*, 7, 243–266.

Klaassen, C.A.J. (1987). Consistent estimation of the influence function of locally asymptotically linear estimators. *Ann. Statist.*, 15, 1548–1562.

Klaassen, C.A.J. and Putter, H. (2005). Efficient estimation of Banach parameters in semiparametric models. *Ann. Statist.*, 33, 307–346.

Klaassen, C.A.J. and Wellner, J.A. (1997). Efficient estimation in the bivariate normal copula model: normal margins are least favourable. *Bernoulli*, 3, 55–77.

Koshevnik, Y.A. and Levit, B.Y. (1976). On a non-parametric analogue of the information matrix. *Theory Probab. Appl.*, 21, 738–753.

Koul, H.L. (1970). Some convergence theorems for ranks and weighted empirical cumulatives. *Ann. Math. Statist.*, 41, 1768–1773.

Koul, H.L. (2002). *Weighted Empiricals and Linear Models.* Lecture Notes in Statistics 166, Springer, New York.

Koul, H.L. and Schick, A. (1997). Efficient estimation in nonlinear autoregressive time-series models. *Bernoulli*, 3, 247–277.

Koul, H.L. and Susarla, V. (1983). Adaptive estimation in linear regression. *Statist. Decisions*, 1, 379–400.

Kreiss, J.-P. (1987a). On adaptive estimation in stationary ARMA processes. *Ann. Statist.*, 15, 112–133.

Kreiss, J.-P. (1987b). On adaptive estimation in autoregressive models when there are nuisance functions. *Statist. Decisions*, 5, 59–76.

Kreiss, J.-P. (1990). Local asymptotic normality for autoregression with infinite order. *J. Statist. Plann. Inference*, 26, 185–219.

Kreiss, J.-P. (1991). Estimation of the distribution function of noise in stationary processes. *Metrika*, 38, 285–297.

Kutoyants, Y.A. (2004). *Statistical Inference for Ergodic Diffusion Processes.* Springer Series in Statistics, Springer, London.

LeCam, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. California Publ. Statist.*, 1, 277–329.

Le Cam, L. (1960). Locally asymptotically normal families of distributions. *Univ. California Publ. Statist.*, 3, 37–98.

Le Cam, L. (1966). Likelihood functions for large numbers of independent observations. In: *Research Papers in Statistics. Festschrift for J. Neyman.* David, F.N., ed., 167–187.

Le Cam, L. (1969). *Théorie Asymptotique de la Décision Statistique.* Séminaire de Mathématiques Supérieures 33, Les Presses de l'Université de Montréal.

Le Cam, L. (1972). Limits of experiments. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.*, 1, 245–261.

Le Cam, L. (1974). *Notes on Asymptotic Methods in Statistical Decision Theory.* Centre de Recherches Mathématiques, Université de Montréal.

Le Cam, L. (1984). Differentiability, tangent spaces and Gaussian auras. Unpublished manuscript.

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory.* Springer Series in Statistics, Springer, New York.

Le Cam, L. and Yang, G.L. (1990). *Asymptotics in Statistics. Some Basic Concepts.* Springer Series in Statistics, Springer, New York.

Liebscher, E. (1999). Estimating the density of the residuals in autoregressive models. *Stat. Inference Stoch. Process.*, 2, 105–117.

Liebscher, E. (2005). Semiparametric density estimators using copulas. *Comm. Statist. Theory Methods*, 34, 59–71.

Loynes, R.M. (1980). The empirical distribution function of residuals from generalised regression. *Ann. Statist.*, 8, 285–299.

Masry, E. (1996). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *J. Time Ser. Anal.*, 17, 571–599.

Masry, E. (2005). Nonparametric regression estimation for dependent functional data: asymptotic normality. *Stochastic Process. Appl.*, 115, 155–177.

Millar, P.W. (1985). Nonparametric applications of an infinite-dimensional convolution theorem. *Z. Wahrsch. Verw. Gebiete*, 68, 545–556.

Müller, U.U. and Wefelmeyer, W. (2002a). Estimators for models with constraints involving unknown parameters. *Math. Methods Statist.*, 11, 221–235.

Müller, U.U. and Wefelmeyer, W. (2002b). Autoregression, estimating functions, and optimality criteria. In: *Advances in Statistics, Combinatorics and Related Areas* (C. Gulati, Y.-X. Lin, J. Rayner and S. Mishra, eds.), 180–195, World Scientific, Singapore 2002.

Müller, U.U., Schick, A. and Wefelmeyer, W. (2004). Estimating functionals of the error distribution in parametric and nonparametric regression. *J. Nonparametr. Statist.*, 16, 525–548.

Müller, U.U., Schick, A. and Wefelmeyer, W. (2005). Weighted residual-based density estimators for nonlinear autoregressive models. *Statist. Sinica*, 15, 177-195.

Müller, U.U., Schick, A. and Wefelmeyer, W. (2006). Efficient prediction for linear and nonlinear autoregressive models. *Ann. Statist.*, 34, 2496–2533.

Müller, U.U., Schick, A. and Wefelmeyer, W. (2007). Estimating the error distribution function in semiparametric regression. *Statist. Decisions*, 25. 1–18.

Müller, U.U., Schick, A. and Wefelmeyer, W. (2008). Estimating the innovation distribution in nonparametric autoregression. To appear in: *Probab. Theory Related Fields.*

Nelsen, R.B. (2006). *An Introduction to Copulas.* 2nd ed. Springer Series in Statistics, Springer, New York.

Neumann, M.H. and Kreiss, J.-P. (1998). Regression-type inference in nonparametric autoregression. *Ann. Statist.*, 26, 1570–1613.

Newey, W.K. (1993). Efficient estimation of models with conditional moment restrictions. In: *Handbook of Statistics 11. Econometrics* (G.S. Maddala, C.R. Rao and H.D. Vinod, eds.), 419–454, North-Holland, Amsterdam.

Newey, W.K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. In: *Handbook of Econometrics*, 4, 2111–2245, North-Holland, Amsterdam.

Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237–249.

Owen, A.B. (2001). *Empirical Likelihood.* Monographs on Statistics and Applied Probability 92, Chapman & Hall/CRC, London.

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

Penev, S., Peng, H., Schick, A. and Wefelmeyer, W. (2004). Efficient estimators for functionals of Markov chains with parametric marginals. *Statist. Probab. Lett.*, 66, 335–345.

Peng, H. and Schick, A. (2002). On efficient estimation of linear functionals of a bivariate distribution with known marginals. *Statist. Probab. Lett.*, 59, 83–91.

Peng, H. and Schick, A. (2004a). Estimation of linear functionals of bivariate distributions with parametric marginals. *Statist. Decisions*, 22, 61–77.

Peng, H. and Schick, A. (2004b). Efficient estimation of a linear functional of a bivariate distribution with equal, but unknown, marginals: the minimum chi-square approach. *Statist. Decisions*, 22, 301–318.

Peng, H. and Schick, A. (2005). Efficient estimation of linear functionals of a bivariate distribution with equal, but unknown marginals: the least-squares approach. *J. Multivariate Anal.*, 95, 385–409.

Pfanzagl, J. (1990). *Estimation in Semiparametric Models. Some Recent Developments.* Lecture Notes in Statistics 63, Springer-Verlag, New York.

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

Prakasa Rao, B.L.S. (1999). *Semimartingales and Their Statistical Inference.* Monographs on Statistics and Applied Probability 83, Chapman & Hall/CRC, Boca Raton.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.*, 22, 300–325.

Ritov, Y. and Bickel, P.J. (1990). Achieving information bounds in non and semiparametric models. *Ann. Statist.*, 18, 925–938.

Robinson, P.M. (1983). Nonparametric estimators for time series. *J. Time Ser. Anal.*, 4, 185–207.

Robinson, P.M. (1987). Time series residuals with application to probability density estimation. *J. Time Ser. Anal.*, 8, 329–344.

Robinson, P.M. (1988). Root-N-consistent semiparametric regression. *Econometrica*, 56, 931–954.

Roussas, G.G. (1965). Asymptotic inference in Markov processes. *Ann. Math. Statist.*, 36, 978–992.

Roussas, G.G. (1972). *Contiguity of Probability Measures: Some Applications in Statistics.* Cambridge Tracts in Mathematics and Mathematical Physics 63, Cambridge University Press, London.

Saavedra, A. and Cao, R. (1999). Rate of convergence of a convolution-type estimator of the marginal density of an MA(1) process. *Stochastic Process. Appl.*, 80, 129–155.

Sacks, J. (1975). An asymptotically efficient sequence of estimators of a location parameter. *Ann. Statist.*, 3, 285–298.

Schick, A. (1986). On asymptotically efficient estimation in semiparametric models. *Ann. Statist.*, 14, 1139–1151.

Schick, A. (1987). A note on the construction of asymptotically linear estimators. *J. Statist. Plann. Inference*, 16, 89–105. Correction 22 (1989), 269–270.

Schick, A. (1993). On efficient estimation in regression models. *Ann. Statist.*, 21, 1486–1521. Correction 23 (1995), 1862–1863.

Schick, A. (1994). On efficient estimation in regression models with unknown scale functions. *Math. Methods Statist.*, 3, 171–212.

Schick, A. (1996). Root-n-consistent and efficient estimation in semiparametric additive regression models. *Statist. Probab. Lett.*, 30, 45–51.

Schick, A. (1999). Efficient estimation in a semiparametric autoregressive model. *Stat. Inference Stoch. Process.*, 2, 69–98

Schick, A. (2001). Sample splitting with Markov chains. *Bernoulli*, 7, 33–61.

Schick, A. and Susarla, V. (1990). An infinite-dimensional convolution theorem with applications to random censoring and missing data models. *J. Statist. Plann. Inference*, 24, 13–23.

Schick, A. and Wefelmeyer, W. (2002a). Estimating the innovation distribution in nonlinear autoregressive models. *Ann. Inst. Statist. Math.*, 54, 245–260.

Schick, A. and Wefelmeyer, W. (2002b). Efficient estimation in invertible linear processes. *Math. Methods Statist.*, 11, 358–379.

Schick, A. and Wefelmeyer, W. (2004a). Root n consistent and optimal density estimators for moving average processes. *Scand. J. Statist.*, 31, 63–78.

Schick, A. and Wefelmeyer, W. (2004b). Estimating invariant laws of linear processes by U-statistics. *Ann. Statist.*, 32, 603–632.

Schick, A. and Wefelmeyer, W. (2004c). Functional convergence and optimality of plug-in estimators for stationary densities of moving average processes. *Bernoulli*, 10, 889–917.

Schick, A. and Wefelmeyer, W. (2007a). Prediction in invertible linear processes. *Statist. Probab. Lett.*, 77, 1322–1331.

Schick, A. and Wefelmeyer, W. (2007b). Uniformly root-n consistent density estimators for weakly dependent invertible linear processes. *Ann. Statist.*, 35, 815–843.

Schick, A. and Wefelmeyer, W. (2008a). Prediction in moving average processes. *J. Statist. Plann. Inference*, 138, 694–707.

Schick, A. and Wefelmeyer, W. (2008b). Root-n consistency in weighted L_1-spaces for density estimators of invertible linear processes. To appear in: *Stat. Inference Stoch. Process.*

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8, 229–231.

Stein, C. (1956). Efficient nonparametric testing and estimation. *Proc. Third Berkeley Symp. Math. Statist. Probab.*, 1, 187–195.

Stone, C.J. (1975). Adaptive maximum likelihood estimators of a location parameter. *Ann. Statist.*, 3, 267–284.

Strasser, H. (1985). *Mathematical Theory of Statistics. Statistical Experiments and Asymptotic Decision Theory.* De Gruyter Studies in Mathematics 7, de Gruyter, Berlin.

Taniguchi, M. and Kakizawa, Y. (2000). *Asymptotic Theory of Statistical Inference for Time Series.* Springer Series in Statistics, Springer, New York.

Tjøstheim, D. (1994). Non-linear time series: a selective review. *Scand. J. Statist.*, 21, 97–130.

Tsiatis, A.A. (2006). *Semiparametric Theory and Missing Data.* Springer Series in Statistics, Springer, New York.

Tsukahara, H. (2005). Semiparametric estimation in copula models. *Canad. J. Statist.*, 33, 357–375.

van der Laan, M.J. and Robins, J.M. (2003). *Unified Methods for Censored Longitudinal Data and Causality.* Springer Series in Statistics, Springer, New York.

van der Vaart, A.W. (1991). On differentiable functionals. *Ann. Statist.*, 19, 178–204.

van der Vaart, A.W. (1998). *Asymptotic Statistics.* Cambridge Series in Statistical and Probabilistic Mathematics 3, Cambridge University Press, Cambridge.

van der Vaart, A.W. (2002). Semiparametric statistics. In: *Lectures on Probability Theory and Statistics*, 331–457, Ecole d'Eté de Probabilités de Saint-Flour XXIX, Lecture Notes in Mathematics 1781, Springer, Berlin.

van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes. With Applications to Statistics.* Springer Series in Statistics, Springer, New York.

van Eeden, C. (1970). Efficiency-robust estimation of location. *Ann. Math. Statist.*, 41, 172–181.

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