The Annals of Statistics
1987, Vol. 15, No. 1, 79–98

REGRESSION MODELS FOR NONSTATIONARY CATEGORICAL TIME SERIES: ASYMPTOTIC ESTIMATION THEORY

BY HEINZ KAUFMANN

University of Regensburg

For the analysis of nonstationary categorical time series, a parsimonious and flexible class of models is proposed. These models are generalizations of regression models for stochastically independent categorical observations. Consistency, asymptotic normality and efficiency of the maximum likelihood estimator are shown under weak and easily verifiable requirements. Some models for binary time series are discussed in detail. To demonstrate asymptotic properties, a theorem is given addressing maximum likelihood estimation for general stochastic processes. Then it is shown that the assumptions of this theorem are consequences of the requirements for categorical time series. For this proof some lemmas are used which may be of interest in similar cases.

1. Introduction. Until recently, categorical time series were mostly analyzed as time homogeneous Markov chains, i.e., Markov chains with stationary transition probabilities. This holds, in particular, if only a single time series is observed, or for panel data, if a model is fitted separately for each individual time series. The asymptotic theory for inference for time homogeneous Markov chains is given in Billingsley (1961). Assuming that the original time series $\{y_t\}$, say, is a homogeneous chain of first order is very restrictive. By considering the vectors $\{(y_t, \ldots, y_{t-l+1})\}$, higher-order Markov chains can be reduced to first-order Markov chains, but without further constraints the number of parameters increases exponentially with the order of the Markov chain. Moreover, in many applications, nonhomogeneous Markov chains are more appropriate, since exogenous variables possibly give rise to nonstationary transition probabilities.

Generalizing regression models for independent categorical observations, we obtain a model family admitting a flexible and parsimonious treatment of higher-order dependence as well as a form of nonstationarity (Section 2). It is assumed that the conditional distribution of y_t, given the whole past, is a function $h(\cdot)$ of a linear combination $Z_t'\beta$. The matrix Z_t depends on exogenous variables and past observations, and β is a vector of unknown parameters. Some particular models for binary time series are given in Section 3.

Statistical inference is based on the asymptotic properties of the maximum likelihood estimator (MLE) of β. For the general model, conditions assuring these properties are stated in Section 5. Checking of the conditions is illustrated in Section 6 with some corollaries for the binary models of Section 3. A theorem stating asymptotic properties of the MLE for general stochastic processes, not

Received May 1985; revised June 1986.

AMS 1980 *subject classifications.* Primary 62M10; secondary 62F12.

Key words and phrases. Time series, categorical data, nonstationary Markov chains, asymptotic estimation theory.

only for categorical time series, is given in Section 7. In Section 8, the assumptions of this theorem are verified from the assumptions made in Section 5, using some lemmas which may be of interest in similar cases.

The model family of Section 2, especially possibilities for the choice of $\{Z_t\}$, is considered in more detail in Fahrmeir and Kaufmann (1987). Moreover, under the assumptions of Section 5, a theorem on asymptotic χ^2-distributions of familiar test statistics is given there, and some tests of special interest in the time series situation are discussed.

In recent years, several other models for categorical, mainly binary, time series have been proposed. In the development of D(iscrete) ARMA processes, Jacobs and Lewis (1983) have been guided by the autocorrelation structure of ARMA processes for continuous variables. Their AR(l)-processes form a subclass of homogeneous Markov chains of higher order, parametrized in a particular way. The models are designed for a parsimonious treatment of stationary time series with discrete, interval-scaled observations. If the observations are categorical, the models are meaningful only for binary observations. Binary AR(l) processes fit into our setting with the identity mapping $h = \mathrm{id}$.

In a different approach, the binary observable time series $\{y_t\}$ is assumed to be generated by truncation of a latent series $\{y_t^*\}$ [e.g., Gourieroux et al. (1983), Grether and Maddala (1982) and Heckman (1981)]. In the most general model considered by Heckman (1981), the latent variable y_t^* is a linear combination of exogenous variables, past values of $\{y_t\}$ and $\{y_t^*\}$, and an error term. Asymptotic estimation theory is given only by Gourieroux et al. (1983) for the special situation where past values of $\{y_t^*\}$ have no influence, and the error process is a Gaussian ARMA process. If past values of $\{y_t^*\}$ have no influence and the error variables are i.i.d., not necessarily Gaussian, then the observable time series is of the type discussed in Section 2, and the results of Section 5 become applicable.

A neurophysiological example is presented by Brillinger and Segundo (1979). This example involves the firing of a neuron subject to presynaptic currents. Recording by $\{y_t\}$ firing or not in successive time intervals, certain assumptions lead to a binary probit model similar to (3.2). Finally, we mention Hauser and Wisniewski (1982), who propose regression models for semi-Markov processes with continuous time.

2. Regression models for categorical time series. Let $\{y_t,\ t = 1, 2, \dots\}$ be a time series with m possible categories for each observation. Suppose the tth observation to be given as a vector $y_t = (y_{t1}, \dots, y_{tq})'$ of length $q = m - 1$, where the component y_{tj} is one, if the jth category is observed, and zero otherwise. Let $\pi_t = (\pi_{t1}, \dots, \pi_{tq})'$ denote the corresponding vector of conditional probabilities given the past observations, i.e., $\pi_{tj} = P(y_{tj} = 1 | y_{t-1}, \dots, y_1)$, $j = 1, \dots, q$. The probability of any event which is determined by a finite number of observations can be computed from the conditional probabilities $\{\pi_t\}$. In most applications however, these quantities are unknown and estimated from the first t observations.

A family of models admitting a flexible treatment of higher-order dependence as well as nonstationarity can be obtained by generalizing models for indepen-

dent categorical responses $\{y_t\}$. In a *regression model for categorical time series*, we assume that, after a certain time l, the conditional probabilities are of the form

$$\pi_t = h(Z_t'\beta), \qquad t > l. \tag{2.1}$$

Here β denotes a vector of unknown parameters, which lies in a p-dimensional open set B. The *link function* h maps a subset $D \subset \mathbb{R}^q$ bijectively onto $\{(\pi_1, \ldots, \pi_q)', \pi_j > 0, j = 1, \ldots, q, \Sigma_1^q \pi_j < 1\}$. The inverse of the q-dimensional logit function is an example for h, leading to the logit model logit $\pi_t = Z_t'\beta$, $t > l$. The assumptions on h imply that all conditional probabilities are strictly positive. In principle, it is possible to relax this assumption. However, this requires a more refined theory, which shall not be studied in this paper.

The $p \times q$-matrix Z_t is a known function of past observations and exogenous variables, which are assumed to be nonstochastic and known at time t. Thus, the matrix Z_t is *predetermined* in that its value is fixed before y_t is observed. If Z_t is a function of exogenous variables and the last l observations only, then $\{y_t\}$ is a Markov chain of order l. In general, this Markov chain is nonhomogeneous. It is homogeneous, if Z_t does not depend on exogenous time varying regressors. In particular, this means that the time t must not enter explicitly but only through $y_{t-1}, \ldots, y_{t-l}$. If Z_t depends only on exogenous variables, we obtain the well-known categorical regression model with independent observations.

3. Some models for binary time series. For observations with only two possible categories, the vector y_t is one dimensional, with $y_t = 1$ if the first and $y_t = 0$ if the second category is observed. The matrix Z_t is a vector, for instance $Z_t' = (1, y_{t-1}, \ldots, y_{t-l})$, leading to

$$\pi_t = h(\beta_0 + \beta_1 y_{t-1} + \cdots + \beta_1 y_{t-l}), \tag{3.1}$$

a homogeneous Markov chain of order l. A nonhomogeneous Markov chain of order l, with exogenous variables $x_{t1}, \ldots, x_{tk}$, is given by

$$\pi_t = h(\beta_0 + \beta_1 y_{t-1} + \cdots + \beta_l y_{t-l} + \alpha_1 x_{t1} + \cdots + \alpha_k x_{tk}). \tag{3.2}$$

For these *autoregressive processes*, the number of parameters increases only linearly with l. Quadratic terms $y_{t-1} \cdot y_{t-2}, \ldots, y_{t-1} \cdot x_{t1}$, etc., or higher-order interaction terms can also be included. If the model is homogeneous and all interactions up to $y_{t-1} \cdot \cdots \cdot y_{t-l}$ are included, then the model contains as many parameters as the general binary homogeneous Markov chain of order l, and within this class, the model is saturated. The only remaining restriction is that all transition probabilities are strictly positive, due to the assumptions on h.

To illustrate these features, consider the model

$$\begin{aligned}\pi_t = h(&\beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_3 x_t + \beta_4 y_{t-1} y_{t-2} + \beta_5 y_{t-1} x_t \\ &+ \beta_6 y_{t-2} x_t + \beta_7 y_{t-1} y_{t-2} x_t),\end{aligned} \tag{3.3}$$

with a nonrandom scalar sequence $\{x_t\}$ assuming at least two values. In general, this is a nonhomogeneous Markov chain of order two, including quadratic and

cubic interaction terms. Setting some of the parameters equal to zero, various submodels are obtained. For further reference, we mention

$$\pi_t = h(\beta_0 + \beta_1 y_{t-1} + \beta_3 x_t + \beta_5 y_{t-1} x_t), \tag{3.4}$$

nonhomogeneous Markov chain of order one, and

$$\pi_t = h(\beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \beta_4 y_{t-1} y_{t-2}), \tag{3.5}$$

the saturated homogeneous Markov chain of order two.

4. Maximum likelihood estimation. This section and the next one refer to the general model of Section 2. From the first t observations, the parameter β can be estimated by the maximum likelihood method. With $y_{sm} = 1 - \sum_{j=1}^{m-1} y_{sj}$, $\pi_{sm} = 1 - \sum_{j=1}^{m-1} \pi_{sj}$, the log-likelihood of the observations $y_{l+1}, \ldots, y_t$ is

$$l_t(\beta) = \sum_{s=l+1}^{t} \sum_{j=1}^{m} y_{sj} \ln \pi_{sj}, \qquad \pi_s = h(Z_s'\beta).$$

Strictly speaking, this is the conditional log-likelihood, given the starting values $y_1, \ldots, y_l$. If the starting distribution does not depend on β, then maximization of the unconditional likelihood would lead to the same estimates. If it does depend on β in a known way, then the unconditional likelihood can be used for estimation. In any case, the loss of information on β, due to maximizing the conditional likelihood only, vanishes asymptotically, under the conditions of Section 5.

The first derivative of the log-likelihood, the *score function*, is

$$s_t(\beta) = \sum_{l+1}^{t} Z_s U_s(\beta)(y_s - \pi_s(\beta)), \tag{4.1}$$

using $u = \text{logit} \circ h$ instead of h for convenience, and defining $U_s(\beta) = [\partial u(\gamma)/\partial \gamma]'$, evaluated at $\gamma = Z_s'\beta$. This matrix has dimension $q \times q$. Set $a_t(\beta) = s_t(\beta) - s_{t-1}(\beta)$ for the increments of the score function. The conditional information,

$$G_t(\beta) = \sum_{l+1}^{t} \text{cov}_\beta(a_s(\beta) | y_{s-1}, \ldots, y_1),$$

plays an important role in the asymptotic considerations. In our context, it is given by

$$G_t(\beta) = \sum_{l+1}^{t} Z_s V_s(\beta) Z_s', \tag{4.2}$$

with $V_s(\beta) = U_s(\beta)\Sigma_s(\beta)U_s'(\beta)$, $\Sigma_s(\beta) = \text{cov}_\beta(y_s | y_{s-1}, \ldots, y_1)$. Integrating out the observations $y_{l+1}, \ldots, y_t$, we get $E_t(\beta) = \text{cov}_\beta(s_t(\beta) | y_l, \ldots, y_1)$, the *information for given starting values*. By further integration, we obtain the *unconditional information* $F_t(\beta) = \text{cov}_\beta s_t(\beta)$. Since the starting distribution and the matrices $\{Z_t\}$ as functions of past observations and exogenous regressors are unspecified, we do not give explicit expressions for $E_t(\beta)$, $F_t(\beta)$. Finally, the

second derivative of the log-likelihood, multiplied by -1, is $H_t(\beta) = G_t(\beta) - R_t(\beta)$, with the remainder term

$$R_t(\beta) = \sum_{s=l+1}^{t} \sum_{j=1}^{q} Z_s W_{sj}(\beta) Z_s' \big(y_{sj} - \pi_{sj}(\beta)\big), \tag{4.3}$$

where $W_{sj}(\beta) = \partial^2 u_j(Z_s'\beta)/\partial\gamma\,\partial\gamma'$. For the logit model, $h = \text{logit}^{-1}$, these expressions simplify considerably:

$$s_t(\beta) = \sum_{l+1}^{t} Z_s(y_s - \pi_s(\beta)), \qquad H_t(\beta) = G_t(\beta) = \sum_{l+1}^{t} Z_s \Sigma_s(\beta) Z_s'.$$

Without concavity assumptions, the assertions of the following sections refer to local maxima. Hence we consider as MLE any measurable function $\hat{\beta}_t$ of $y_1, \ldots, y_t$ maximizing the log-likelihood locally on B. Since B is open, we may equivalently solve

$$s_t(\hat{\beta}_t) = 0, \qquad H_t(\hat{\beta}_t) \text{ positive semidefinite.} \tag{4.4}$$

A MLE exists if and only if (4.4) has a solution $\hat{\beta}_t$ inside of B. If $H_t(\hat{\beta}_t)$ is positive definite, then the MLE is locally unique. Outside of the (measurable) set where local maxima exist, $\hat{\beta}_t$ may be defined as an arbitrary constant, to obtain a random variable defined throughout the sample space.

If the admissible set B is open and convex and the log-likelihood is (weakly) concave, more distinct assertions can be obtained. This is due to the fact that in this case the zeros of the score function form a convex set maximizing the log-likelihood globally. Hence, a random vector $\hat{\beta}_t$ is a (global) MLE if and only if it is a function of $y_1, \ldots, y_t$, and $s_t(\hat{\beta}_t) = 0$ on the set where s_t has a zero at all. (Global) uniqueness of $\hat{\beta}_t$ is equivalent to full rank of $H_t(\hat{\beta}_t)$. In fact, concavity of the log-likelihood holds for most of the common link functions, see Section 6 of this paper and Wedderburn (1976) for dichotomous, and Kaufmann (1987) for polytomous observations. Both papers also give uniqueness conditions which can be checked without computing a MLE; Wedderburn (1976) also discusses its existence for a finite sample.

5. Asymptotic estimation theory. At first, we note the following simple remark. Varying $y_{t-1}, \ldots, y_1$ and keeping the exogenous variables fixed, we obtain for any $t > l$ a finite number $n(t)$ of possible values of Z_t,

$$Z_{t,1}, \ldots, Z_{t,n(t)}, \tag{5.1}$$

where each $Z_{t,j}$ is nonrandom. This list is used in formulating the following conditions assuring asymptotic properties of the MLE. For any symmetric matrix A, its smallest (largest) eigenvalue is denoted by $\lambda_{\min}(A)$ $[\lambda_{\max}(A)]$.

ASSUMPTION A. (i) The time series $\{y_t\}$ is a Markov chain of order l, for all parameter vectors β out of the admissible set B.

(ii) The possible values $\{Z_{t,j}\}$ lie in a compact set C such that $Z'\beta$ lies within the domain D of h, for all $Z \in C$, $\beta \in B$.

(iii) $\lambda_{\min}(\sum_{s=l+1}^{t}\sum_{j=1}^{n(s)}Z_{s,j}Z'_{s,j}) \to \infty$.

(iv) The link function h is two times continuously differentiable, $\det(\partial h(\gamma)/\partial\gamma) \neq 0$.

Assumption A(i) holds if and only if the matrix Z_t, $t > l$, does not depend on observations more than l lags back. Apart from $Z'\beta \in D$ for all $Z \in C$, $\beta \in B$, assuring that the argument of h is always well defined, A(ii) and (iii) depend only on the asymptotic behaviour of the possible values $\{Z_{t,j}\}$ of the predetermined matrices. In the classical linear regression model with i.i.d. errors and the sequence $\{z_t\}$ of regressors, $\lambda_{\min}(\sum_1^t z_s z_s') \to \infty$ is necessary and sufficient for weak [Drygas (1976)] and strong [Lai, Robbins and Wei (1979)] consistency. In A(iii), one has additionally to sum over the possible values of Z_s, at each s.

As norming quantities we use square roots of positive definite matrices. By $A^{1/2}(A^{T/2})$ we denote a left (the corresponding right) square root of a positive definite matrix A, i.e., $A^{1/2}A^{T/2} = A$. In addition, we set $A^{-1/2} = (A^{1/2})^{-1}$, $A^{-T/2} = (A^{T/2})^{-1}$. Left (right) square roots are unique up to an orthogonal transformation from the right (from the left). If A is nonrandom or measurable with respect to some σ-field, then $A^{1/2}$ is assumed to share this property. Unique, continuous "versions" of the square root are the symmetric, positive definite square root and the Cholesky square root. The left Cholesky square root is defined by the condition to be a lower triangular matrix with positive diagonal entries.

THEOREM 1. *Under assumption* A, *the probability that a locally unique MLE exists converges to one. Moreover, there exists a sequence* $\{\hat{\beta}_t\}$ *of MLE's which is consistent and asymptotically normal,*

$$G_t^{T/2}(\beta)\big(\hat{\beta}_t - \beta\big) \to_d N(0, I), \tag{5.2}$$

with an appropriate square root $G_t^{T/2}(\beta)$, *for instance the Cholesky square root.*

REMARK 1. Appropriate square roots, in particular the Cholesky square roots, of $G_t(\hat{\beta}_t)$, $H_t(\beta)$ or $H_t(\hat{\beta}_t)$ can also be used as norming quantities. This holds also for arbitrary square roots of $E_t(\beta)$ or $F_t(\beta)$, provided they do not depend on observations y_s, $s > l$, resp. are nonrandom, according to our conventions on square roots.

Concavity of the log-likelihood is implied by concavity of $\log h_j(\gamma)$, $j = 1, \ldots, m$, where $h_j(\gamma)$ is the jth component of the link function, $j = 1, \ldots, q$, and $h_m(\gamma) = 1 - \sum_{j=1}^{q} h_j(\gamma)$. Following the discussion at the end of Section 4, we have the following corollary.

COROLLARY 1. *If condition* A *holds with a convex admissible set* B, *and if* $\log h_j(\gamma)$ *is concave,* $j = 1, \ldots, m$, *then the probability that a unique MLE exists converges to one. Any sequence* $\{\hat{\beta}_t\}$ *of MLE's is consistent and asymptotically normal as in Theorem* 1.

In conjunction with Theorem 1, several efficiency results can be derived from the LAN (locally asymptotically normal) condition, see Ibragimov and Has'minskii (1981, Chapter 2), Jeganathan (1982) or Basawa and Scott (1983, Chapter 2). Recall that, due to our conventions on square roots, $E_t^{T/2}(\beta)$ denotes a square root of the conditional information nonrandom for given starting values.

LEMMA 1. *Under assumption* A, *the LAN condition is satisfied for any* $\beta \in B$ *and any norming sequence* $\{E_t^{-T/2}(\beta)\}$, *for given starting values.*

Theorem 2 below picks out two examples of the possible efficiency results. The MLE $\hat{\beta}_t$ is compared with estimators $\tilde{\beta}_t$, which are functions of $y_1, \ldots, y_t$ and, for some sequence $\{E_t^{T/2}(\beta)\}$, regular in the following sense: for $\lambda \in \mathbb{R}^p$, set $\beta(t) = \beta + E_t^{-T/2}(\beta)\lambda$. The estimator $\tilde{\beta}_t$, $t > l$, is regular, if $\{E_t^{T/2}(\beta)(\tilde{\beta}_t - \beta(t))\}$ converges, under $\{P_{\beta(t)}(\cdot|y_1, \ldots, y_l)\}$, in distribution to some random variable $Z(\beta)$, for any fixed λ.

THEOREM 2. *Under assumption* A, *the following statements hold for any norming sequence* $\{E_t^{T/2}(\beta)\}$: *the MLE* $\{\hat{\beta}_t\}$ *of Theorem* 1 *is regular. Within the class of regular estimators* $\{\tilde{\beta}_t\}$, *the asymptotic probability of concentration,*

$$\lim_{t\to\infty} P\big(E_t^{T/2}(\beta)\big(\tilde{\beta}_t - \beta\big) \in C\big),$$

attains its maximum if $\{\tilde{\beta}_t\} = \{\hat{\beta}_t\}$, *for any symmetric convex set* C. *Within the class of regular estimators* $\{\tilde{\beta}_t\}$ *with a normally distributed limit vector* $Z(\beta) \sim N(0, \Sigma_\beta)$ *say, the covariance matrix* I *of* $\{\hat{\beta}_t\}$ *is minimal in that the difference* $\Sigma_\beta - I$ *is always positive semidefinite.*

REMARKS. (i) Under the assumptions of Corollary 1, Theorem 2 holds for any sequence of MLE's.

(ii) For certain loss functions, $\{\hat{\beta}_t\}$ asymptotically has minimum expected loss within the class of regular estimators. If the MLE is compared with estimators where convergence in distribution is required for $\lambda = 0$, $\beta(t) = \beta$ only, then the assertions above continue to hold for almost all points of the parameter space B. See the references above for these and further results.

For many Markov models of order l, e.g., (3.1), (3.2) or (3.5) and its submodels, there is no interaction between the past observations and the exogenous variables. Equivalently, after proper reparameterization, the matrix Z_t can be partitioned into two parts, $Z_t' = (Y(y_{t-1}, \ldots, y_{t-l}), X_t)'$. The first matrix has dimension $p_1 \times q$. It must always be the same function of $y_{t-1}, \ldots, y_{t-l}$, independent of t. The $(p - p_1) \times q$-matrix X_t is the matrix of exogenous variables. In terms of the list (5.1), the matrices $Z_{t,j}$ can be partitioned,

$$Z_{t,j}' = \big(Y_j', X_t'\big), \qquad j = 1, \ldots, n,\ t > l, \tag{5.3}$$

and $n(t) = n$ is independent of t. The matrices $Y_1, \ldots, Y_n$ are the images of the

mapping Y. If such a partitioning is possible, we can always achieve, by proper reparameterization, that there are no constants left in the Y-part. Equivalently, the system $\mu = Y_j'\lambda$, $j = 1, \ldots, n$, only has solutions with $\mu = 0$.

If there is no interaction between the past observations and the exogenous variables, we can split the divergence condition A(iii) into two parts. For the Y-part, we need only require full rank. For the X-part, a simpler divergence condition can be given.

COROLLARY 2. *Assume that, for a Markov chain of order l, the matrices $Z_{t,j}$ can be partitioned as in* (5.3). *Let the system $\mu = Y_j'\lambda$, $j = 1, \ldots, n$, only have solutions with $\mu = 0$. Then a necessary and sufficient condition for* A(iii) *is*

$$\lambda_{\min}\left(\sum_{j=1}^{n} Y_j Y_j'\right) > 0, \qquad \lambda_{\min}\left(\sum_{s=l+1}^{t} X_s X_s'\right) \to \infty. \tag{5.4}$$

Hence, the conclusions of Theorems 1 *and* 2 *hold under assumption* A, *with* A(iii) *replaced by* (5.4).

6. Some corollaries for binary time series. To give some discussion of the assumptions assuring asymptotic properties, we will see how they reduce for the binary models introduced in Section 3.

The conditions on the link functions h simplify to the following

ASSUMPTION H. The link function h maps an open interval (d_1, d_2), with possibly infinite endpoints, onto $(0, 1)$. It is two times continuously differentiable with a strictly positive derivative.

This assumption holds, e.g., for logit and probit models, where h is the logistic resp. normal distribution function, $(d_1, d_2) = (-\infty, +\infty)$, for the identity link $h(\gamma) = \gamma$, $(d_1, d_2) = (0, 1)$ and for the angular transform $h(\gamma) = \sin^2\gamma$, $(d_1, d_2) = (0, \pi/2)$. Moreover, for these link functions the inequalities

$$(h - 1)^{-1} \leq \ddot{h}\dot{h}^{-2} \leq h^{-1} \tag{6.1}$$

hold, where $\dot{h}$ ($\ddot{h}$) denotes the first (second) derivative. Under assumption H, (6.1) is equivalent to concavity of $\log h$ and $\log(1 - h)$ and implies concavity of the log-likelihood.

For the homogeneous autoregressive process (3.1), there are $n = 2^l$ possible values of Z_t, which are actually independent of t, namely

$$\begin{aligned} Z_{(1)} &= (1, 0, \ldots, 0, 0), \\ Z_{(2)} &= (1, 0, \ldots, 0, 1), \\ &\vdots \\ Z_{(n)} &= (1, 1, \ldots, 1, 1). \end{aligned} \tag{6.2}$$

The compact set C may be defined as the convex hull of these points. Then,

under assumption H, condition A(ii) reduces to

$$d_1 < Z'_{(j)}\beta < d_2, \qquad j = 1, \ldots, n,\ \beta \in B,$$

and even further to a system of two inequalities,

$$(6.3) \qquad d_1 - s_1 + s_2 < \beta_0 < d_2 - s_1 - s_2, \qquad \beta = (\beta_0, \beta_1, \ldots, \beta_l)' \in B,$$

where $s_1 = \frac{1}{2}\Sigma_1^l \beta_j$, $s_2 = \frac{1}{2}\Sigma_1^l |\beta_j|$. These inequalities can be solved for β choosing $\beta_1, \ldots, \beta_l$ subject to $2s_2 < d_2 - d_1$ and then computing the bounds for β_0 given in (6.3). If $d_1 = -\infty$ or $d_2 = +\infty$, then the corresponding inequality does not restrict the admissible set B. In the logit or probit model, for instance, condition A(ii) can be dropped under (3.1). Further, we have

$$(6.4) \qquad \lambda_{\min}\left(\sum_{s=l+1}^{t} \sum_{j=1}^{n(t)} Z_{s,j} Z'_{s,j} \right) = (t-l)\lambda_{\min}\left(\sum_{j=1}^{n} Z_{(j)} Z'_{(j)} \right).$$

It is well known that $\Sigma_{j=1}^n Z_{(j)} Z'_{(j)}$ is of full rank, whence A(iii) follows from (6.4). Collecting together, we have the following corollary.

COROLLARY 3. *For the homogeneous autoregressive process* (3.1), *condition* A *holds if and only if the link function h fulfills condition* H *and the parameter set B fulfills* (6.3).

The more general autoregressive model (3.2) allows for exogenous variables $x_t = (x_{t1}, \ldots, x_{tk})'$, without interaction between x_t and $y_{t-1}, \ldots, y_{t-l}$, however. The possible values of Z_t are now obtained by appending x_t to any vector of (6.2). They depend on t, if x_t assumes at least two values. With some additional arguments, A(ii) can be reduced again. For simplicity, we consider only $(d_1, d_2) = (-\infty, +\infty)$. Then A(ii) holds if and only if $\{x_t, t > l\}$ is bounded.

Regarding A(iii), we can apply Corollary 2. Defining $Y_j = Z_{(j)}$, $Z_{(j)}$ as in (6.2), yields $\mu = 1$, $\lambda = (1, 0, \ldots, 0)'$, to be a solution of $\mu = Y_j'\lambda$, $j = 1, \ldots, n$, in conflict with the assumptions of Corollary 2. Hence we must drop the constant in Y_j, $j = 1, \ldots, n$, and consider $(1, x_t')'$ as exogenous vector. Then, due to full rank of $\Sigma_1^n Z_{(j)} Z'_{(j)}$, the matrix $\Sigma_1^n Y_j Y_j'$ is also of full rank, and $\mu = 0$, $\lambda = 0$ is the only solution of $\mu = Y_j'\lambda$, $j = 1, \ldots, n$. In the divergence condition for the exogenous variables, utilizing the inversion formula for partitioned matrices and boundedness of $\{x_t\}$, we can eliminate the constant by centering the vectors x_t. Then A(iii) reduces to the assumption that these vectors are sufficiently scattered. This holds, for instance, if its empirical covariance matrix converges to a positive definite matrix.

COROLLARY 4. *For the nonhomogeneous autoregressive process* (3.2), *assume that the link function h fulfills assumption* H *with* $d_1 = -\infty$, $d_2 = +\infty$. *Then assumption* A *holds if and only if the exogenous variables* $x_t = (x_{t1}, \ldots, x_{tk})'$, $t > l$, *are bounded, and the smallest eigenvalue of their scatter*

matrix diverges:

$$\lambda_{\min}\left(\sum_{s=l+1}^{t}(x_s-\bar{x}_t)(x_s-\bar{x}_t)'\right)\to\infty,\qquad \bar{x}_t=(t-l)^{-1}\sum_{l+1}^{t}x_s.$$

For the model (3.3) and its submodels, we assume again $d_1=-\infty$, $d_2=+\infty$. For the homogeneous model (3.5), condition A(iii) holds without further assumptions, with similar arguments as those preceding Corollary 3. Next we consider the nonhomogeneous submodels of (3.3) without an interaction term between x_t and y_{t-1} or y_{t-2}. Again, A(ii) holds if and only if $\{x_t,\ t>l\}$ is bounded. Under this assumption, applying Corollary 2, it follows that

$$\sum_{s=l+1}^{t}(x_s-\bar{x}_t)^2\to\infty,\qquad \bar{x}_t=(t-l)^{-1}\sum_{l+1}^{t}x_s, \tag{6.5}$$

is necessary and sufficient for A(iii). It turns out that the same statement holds for (3.3) and (3.4), although there are interaction terms x_ty_{t-1} or x_ty_{t-2}: for the simpler model (3.4), the possible values of Z_t are

$$Z_{t,1}=(1,x_t,0,0)',\qquad Z_{t;2}=(1,x_t,1,x_t)',$$

after interchanging coordinates two and three. With

$$C_t=\begin{bmatrix} t-l & \sum_{l+1}^{t}x_s \\ \sum_{l+1}^{t}x_s & \sum_{l+1}^{t}x_s^2 \end{bmatrix},$$

we have

$$\sum_{l+1}^{t}(Z_{s,1}Z_{s,1}'+Z_{s,2}Z_{s,2}')=\begin{bmatrix} 2C_t & C_t \\ C_t & C_t \end{bmatrix}.$$

Since

$$\begin{bmatrix} C_t & C_t \\ C_t & C_t \end{bmatrix}=\sum_{l+1}^{t}Z_{s,2}Z_{s,2}'$$

is positive semidefinite, we can apply Lemma 4 of Section 8 with $\alpha=\frac{1}{2}$ and obtain that A(iii) is equivalent to $\lambda_{\min}(C_t)\to\infty$, which is in turn equivalent to (6.5), if $\{x_t,\ t>l\}$ is bounded. By repeated application of Lemma 4, this can also be shown for (3.3).

COROLLARY 5. *For each submodel of* (3.3), *assume that the link function h fulfills assumption* H *with* $d_1=-\infty$, $d_2=+\infty$. *For the homogeneous model* (3.5) *and its submodels, this implies condition* A. *For the nonhomogeneous models,* A *is valid if and only if* $\{x_t,\ t>l\}$ *is bounded and* (6.5) *holds.*

7. A general theorem on maximum likelihood estimation. The following theorem addresses maximum likelihood estimation for general stochastic

processes with discrete time. Let $\{y_t,\ t = 1,2,\dots\}$ be such a process on a probability space $(\Omega, \mathfrak{A}, P)$. The σ-field generated by the first t observations $y_1,\dots, y_t$ is denoted by $\mathfrak{A}_t$, $\mathfrak{A}_0 = \{\varnothing, \Omega\}$. The probability measure P is assumed to belong to a parametric family $\{P_\beta,\ \beta \in B\}$, where the parameter space B is an open subset of $\mathbb{R}^p$, $p \in \mathbb{N}$. For fixed t, let the projections $\{P_{t,\beta},\ \beta \in B\}$ on the first t observations be mutually absolutely continuous. Then the corresponding likelihood exists. If the likelihood is continuous, it is nonzero for all $\beta \in B$ and unique up to a factor which does not depend on β, a.s. Beyond continuity, the likelihood is assumed to be two times continuously differentiable. Let $l_t(\beta)$, $s_t(\beta)$, $-H_t(\beta)$ denote the log-likelihood and its first and second derivatives, respectively; in addition, define $a_t(\beta) = s_t(\beta) - s_{t-1}(\beta)$.

Theorem 3 below parallels Theorem 1 and refers also to local maxima. Hence the same definition of a MLE as in Section 4 applies. Asymptotic properties can be obtained under the following condition N. This condition refers to the true probability measure $P = P_\beta$. Usually, the assumptions have to be checked for all $\beta \in B$; the constants involved may depend on β. To simplify notation, dependence on the true parameter vector β is mostly suppressed.

ASSUMPTION N. (i) The score function $\{s_t\}$, evaluated at β, is a square integrable zero mean martingale with respect to $\{\mathfrak{A}_t\}$.

(ii) With some nonrandom nonsingular norming sequence $\{A_t^{1/2}\}$, the conditional information $G_t(\beta) = \sum_1^t \operatorname{cov}_\beta(a_s(\beta)|\mathfrak{A}_{s-1})$ converges to a random a.s. positive definite matrix,

$$A_t^{-1/2} G_t(\beta) A_t^{-T/2} \to_p V(\beta).$$

(iii) The Lindeberg condition holds, i.e., for any $\varepsilon > 0$,

$$\sum_1^t E\big(a_s' A_t^{-1} a_s I_{ts}(\varepsilon) | \mathfrak{A}_{s-1}\big) \to_p 0,$$

where $I_{ts}(\varepsilon)$ is the indicator of $\{a_s' A_t^{-1} a_s \geq \varepsilon^2\}$.

(iv) The continuity condition

$$\sup_{\tilde\beta \in N_t(\delta)} \big\| A_t^{-1/2}\big(H_t(\tilde\beta) - G_t\big) A_t^{-T/2} \big\| \to_p 0,$$

with $N_t(\delta) = \{\tilde\beta\colon \|A_t^{T/2}(\tilde\beta - \beta)\| \leq \delta\}$, holds for any $\delta > 0$.

THEOREM 3. *Under assumption* N, *the probability that a locally unique MLE exists converges to one. Moreover, there exists a sequence* $\{\hat\beta_t\}$ *of MLE's which is consistent and asymptotically normal,*

$$G_t^{T/2}\big(\hat\beta_t - \beta\big) \to_{\cdot} N(0, I), \tag{7.1}$$

with an appropriate square root $G_t^{T/2}$.

REMARKS. (i) The limit law (7.1) holds, for instance, if $G_t^{T/2} A_t^{-T/2}$ is the right Cholesky square root of $A_t^{-1/2} G_t A_t^{-T/2}$. Condition N(ii) implies

nonsingularity of $A_t^{-1/2}G_tA_t^{-T/2}$ in the probability limit, hence such a version of $G_t^{T/2}$ exists, with probability converging to one.

(ii) The situation where the limiting matrix $V(\beta)$ is truly random is referred to as nonergodic, whereas $V(\beta)$ a.s. constant is referred to as ergodic [e.g., Basawa and Scott (1983)]. In the latter case, one can renorm to obtain $V(\beta) = I$. Then N(ii) holds for any version of $A_t^{1/2}$, if it holds at all, and in (7.1) one can choose the Cholesky square root $G_t^{T/2}$.

(iii) In the limit law (7.1), the matrix G_t can be replaced by $G_t(\hat{\beta}_t)$ if, additionally, the continuity condition N(iv) holds with $G_t(\tilde{\beta})$ instead of $H_t(\tilde{\beta})$. Concerning appropriate square roots, Remarks (i) and (ii) remain valid, with $G_t(\hat{\beta}_t)$ instead of G_t.

(iv) Similar results have been given by Sweeting (1980) and Basawa and Scott (1983). In the theorem presented here a martingale approach is used instead of requiring uniformity or continuity of convergence in N(ii) and N(iv), which may sometimes be difficult to check. A similar martingale approach is also used by Jeganathan (1982), for an asymptotically centering sequence of estimators. However, he does not provide conditions under which the MLE is asymptotically centering. In contrast to the papers mentioned, asymptotic normality of the MLE is stated here with random norming. This is facilitated by considering versions of the square root which are different from the symmetric positive definite square root.

8. Proofs. In establishing limiting properties of nonstationary Markov chains, the δ-coefficient of Dobrushin (1956) plays an important role. For a stochastic $m \times m$-matrix $Q = (q_{jk})$, it is defined by

$$\delta(Q) = \tfrac{1}{2}\max_{i,j}\sum_{k=1}^{m}|q_{ik} - q_{jk}|.$$

From Isaacson and Madsen (1976, Lemmas V.2.2 and V.2.3) it follows easily that $\delta(Q) \leq 1$,

$$\min_{j,k} q_{jk} \geq c \Rightarrow \delta(Q) \leq 1 - mc, \tag{8.1}$$

$$\delta(Q_1Q_2) \leq \delta(Q_1)\delta(Q_2), \tag{8.2}$$

for a product of stochastic matrices Q_1, Q_2. With the δ-coefficient, the following mixing inequality can be obtained.

LEMMA 2. *Let y_1, y_2 be random variables with the same finite state space $\{1,\ldots,m\}$ and transition matrix $Q = (q_{jk})$, $q_{jk} = P(y_2 = k|y_1 = j)$. If the random variable x_i is a function of y_i, $i = 1, 2$, then*

$$|\operatorname{cov}(x_1, x_2)| \leq 2\delta(Q)E|x_1|\max|x_2|.$$

PROOF. Setting $(p_{j+}), (p_{+k})$ for the marginal distributions of y_1 resp. y_2, we have

$$\operatorname{cov}(x_1, x_2) = \sum_j x_1(j) p_{j+} \sum_k x_2(k)(q_{jk} - p_{+k}),$$

$$|\operatorname{cov}(x_1, x_2)| \le E|x_1| \max_k |x_2(k)| \max_j \sum_k |q_{jk} - p_{+k}|.$$

With $\Sigma_k |q_{jk} - p_{+k}| \le \Sigma_i p_{i+} \Sigma_k |q_{jk} - q_{ik}| \le \max_i \Sigma_k |q_{jk} - q_{ik}|$, the assertion follows. □

REMARK. The restriction to a finite state space is made for simplicity only. It could be removed similarly as in the proof of Theorem A6 in Hall and Heyde (1980).

To prove Theorem 1, we will verify the assumptions of Theorem 3 for given starting values, since β is estimated from the conditional likelihood. Hence the conclusions of Theorem 3 at first hold conditionally. Since there are only a finite number of possible starting values, the unconditional statements follow easily by integration. Time $t - l$ of Section 7 is to be identified with time t of the earlier sections; indices always refer to the latter one. For instance, $\mathfrak{A}_t$ denotes the σ-field generated by $y_1, \ldots, y_t$. As in Section 7, the argument β is dropped whenever possible.

More specifically, we will verify condition N with $A_{t-l} = E_t(\beta)$, $V(\beta) = I$. An important intermediate step is made in the following lemma, which states in particular N(ii).

LEMMA 3. *Condition* A *implies*

$$\lambda_{\min}(F_t) \to \infty, \qquad \lambda_{\min}(E_t) \to \infty, \tag{8.3}$$

the latter for any starting values. Moreover, under assumption A,

$$F_t^{-1/2} G_t F_t^{-T/2} \to_p I, \qquad E_t^{-1/2} G_t E_t^{-T/2} \to_p I \tag{8.4}$$

both hold unconditionally, the latter also for given starting values, for arbitrary square roots $\{F_t^{1/2}\}$, $\{E_t^{1/2}\}$.

PROOF. From (4.1), it is easily seen that $\{s_t,\ t > l\}$ is a square integrable zero mean martingale, under P as well as under $P(\cdot|\mathfrak{A}_l)$. Hence $\{s_t\}$ has orthogonal increments $\{a_t\}$, and $G_t\ (= \Sigma Z_s V_s Z_s') = \Sigma E(a_s a_s' | \mathfrak{A}_{s-1})$. By integration,

$$F_t = EG_t \quad \text{and} \quad E_t = E(G_t|\mathfrak{A}_l). \tag{8.5}$$

In the sequel, we assume $l > 0$ without loss of generality. Consider $w_t = (y_t, \ldots, y_{t-l+1})$, $t \ge l$. This process forms a Markov chain of first order, with state space W, say. Since $Z_t \in C$ compact, it holds that $\inf_{t>l,\, j} P(y_{tj} = 1 | \mathfrak{A}_{t-1}) > 0$. By induction and summation, it follows that

$$\inf_{t,\, s,\, w} P(w_{t+l} = w | \mathfrak{A}_s) > 0, \tag{8.6}$$

subject to $t > l$, $t \geq s$ and $w \in W$. In particular, we have

$$\inf P(w_t = w) > 0, \qquad \inf P(w_t = w | \mathfrak{A}_l) > 0, \tag{8.7}$$

where the infimum is over $w \in W$, $t > 2l$. The compactness condition A(ii), together with $\det \partial h(\gamma)/\partial \gamma \neq 0$ resp. $\det \partial u(\gamma)/\partial \gamma \neq 0$, also implies that

$$\inf_{t>l} \lambda_{\min}(V_t) > 0. \tag{8.8}$$

Since $F_t = \sum E Z_s V_s Z_s'$ [see (8.5)], the inequalities (8.7) and (8.8) yield, with some constant $c > 0$,

$$\lambda_{\min}(F_t) \geq c \lambda_{\min}\left(\sum_{s=2l+1}^{t} \sum_{j=1}^{n(t)} Z_{s,j} Z_{s,j}' \right).$$

Hence $\lambda_{\min}(F_t) \to \infty$ follows from A(iii). In particular, F_t is nonsingular for all $t \geq t_0$, say. The second inequality of (8.7) gives $\lambda_{\min}(E_t) \to \infty$, with analogous arguments.

To show $F_t^{-1/2} G_t F_t^{-T/2} \to_p I$, consider the triangular array

$$v_{st} = \lambda' F_t^{-1/2} Z_s V_s Z_s' F_t^{-T/2} \lambda, \qquad l < s \leq t,\ t \geq t_0,$$

where λ is fixed, $\lambda'\lambda = 1$. We have

$$\sum_{l+1}^{t} v_{st} = \lambda' F_t^{-1/2} G_t F_t^{-T/2} \lambda, \qquad \sum_{l+1}^{t} E v_{st} = 1.$$

For $F_t^{-1/2} G_t F_t^{-T/2} \to_p I$, it is sufficient that

$$\operatorname{var} \sum_{l+1}^{t} v_{st} \to 0, \tag{8.9}$$

for any fixed λ with $\lambda'\lambda = 1$. This can be shown by an application of Lemma 2. Since v_{st} is a nonnegative random variable depending only on w_{s-1}, we obtain

$$|\operatorname{cov}(v_{rt}, v_{st})| \leq 2 E v_{rt} M_t \delta(Q_r \cdot \cdots \cdot Q_{s-1}), \qquad l < r \leq s \leq t, \tag{8.10}$$

where $M_t = \max_{l < s \leq t} v_{st}$, and Q_t denotes the transition matrix $w_{t-1} \to w_t$. From (8.6), it can be inferred that for products $Q_t \cdot \cdots \cdot Q_{t+l-1}$ of length l, all entries are bounded away from zero, uniformly for all $t > l$. In view of (8.1), there exists a $\gamma < 1$ with $\delta(Q_t \cdot \cdots \cdot Q_{t+l-1}) \leq \gamma$, for all $t > l$. From (8.2), we obtain

$$\delta(Q_r \cdot \cdots \cdot Q_{s-1}) \leq \gamma^n, \qquad nl \leq s - r < (n+1)l. \tag{8.11}$$

Inserting (8.11) into (8.10) and summing up yields

$$\sum_r \sum_s |\operatorname{cov}(v_{rt}, v_{st})| \leq \frac{2l}{1-\gamma} \sum_r E v_{rt} M_t = \frac{2l}{1-\gamma} M_t.$$

In view of

$$\operatorname{var} \sum v_{st} \leq 2 \sum_r \sum_s |\operatorname{cov}(v_{rt}, v_{st})|,$$

it follows that, with $c = 4l/(1 - \gamma)$,

$$\operatorname{var} \sum_{s=l+1}^{t} v_{st} \le cM_t \le c\Big(\sup_{s>l} \lambda_{\max} Z_s V_s Z_s'\Big)/\lambda_{\min} F_t.$$

The numerator of the right side is finite, from $Z_s \in C$ compact. The denominator diverges, hence (8.9) holds.

The second part of (8.4) can be shown with analogous arguments, for given starting values. The unconditional result follows by integration. □

PROOF OF THEOREM 1. It is easy to see that any model of Section 2 fits into the general setting of Section 7, provided the link function is two times continuously differentiable. Hence we concentrate on condition N, for given starting values. Assumption N(i) follows immediately from (4.1) and has already been used in the proof of Lemma 3, which contains N(ii) with $A_{t-l} = E_t(\beta)$, $V(\beta) = I$.

Next we demonstrate the Lindeberg condition N(iii). The increment of the score function is $a_t = Z_t U_t(y_t - \pi_t)$. Since $Z_t \in C$ compact, it follows that $\sup_{t>l} \|a_t\| < \infty$. This implies $a_s' E_t^{-1} a_s \le c/\lambda_{\min}(E_t)$, with some constant c. From $\lambda_{\min}(E_t) \to \infty$, it follows that $I_{ts}(\varepsilon) = 0$, $l < s \le t$, if t is sufficiently large, $\varepsilon > 0$ fixed. Hence N(iii) holds.

Equivalently to the continuity condition N(iv), it can be shown that, for any $\delta > 0$ and any fixed but arbitrary λ,

$$\sup_{\tilde\beta \in N_t(\delta)} \lambda' E_t^{-1/2}\big(H_t(\tilde\beta) - G_t\big)E_t^{-T/2}\lambda \to_p 0. \tag{8.12}$$

To verify (8.12), we decompose $H_t(\tilde\beta)$ [compare Fahrmeir and Kaufmann (1985, proof of Theorem 4)]. Since $H_t(\tilde\beta) = G_t(\tilde\beta) - R_t(\tilde\beta)$,

$$g_t := \sup_{N_t(\delta)} \lambda' E_t^{-1/2}\big(G_t(\tilde\beta) - G_t\big)E_t^{-T/2}\lambda \to_p 0, \tag{8.13}$$

$$\sup_{N_t(\delta)} \lambda' E_t^{-1/2} R_t(\tilde\beta) E_t^{-T/2}\lambda \to_p 0 \tag{8.14}$$

together are sufficient for (8.12). Let t_0 be such that E_t, $t \ge t_0$, is nonsingular. With the vectors $w_{st}' = \lambda' E_t^{-1/2} Z_s$, $l < s \le t$, and $w_t = \sum_{s=l+1}^{t} w_{st}' w_{st}$, we have

$$g_t = \sup_{N_t(\delta)} \sum_s w_{st}'\big(V_s(\tilde\beta) - V_s\big)w_{st},$$

$$g_t \le w_t \sup_{\tilde\beta \in N_t(\delta),\ s>l} \big\|V_s(\tilde\beta) - V_s\big\|, \qquad t \ge t_0.$$

Using A(ii), $\sup_{s>l}\|V_s(\tilde\beta) - V_s\|$ can be estimated from above by a continuous function of $\tilde\beta$ with a zero at $\tilde\beta = \beta$. Since $\{N_t(\delta)\}$ shrinks to β, we obtain

$$\sup_{\tilde\beta \in N_t(\delta),\ s>l} \big\|V_s(\tilde\beta) - V_s\big\| \to 0.$$

On the other hand, using (8.8), it can be shown that Ew_t is bounded, uniformly in t. By an application of the Markov inequality, (8.13) follows.

By further decomposition, we obtain that

$$\sup_{N_t(\delta)} \sum_s w'_{st}\big(W_{sj}(\tilde{\beta}) - W_{sj}\big)w_{st}(y_{sj} - \pi_{sj}) \to_p 0, \tag{8.15}$$

$$\sup_{N_t(\delta)} \sum_s w'_{st}W_{sj}(\tilde{\beta})w_{st}\big(\pi_{sj} - \pi_{sj}(\tilde{\beta})\big) \to_p 0, \tag{8.16}$$

$$\sum w'_{st}W_{sj}w_{st}(y_{sj} - \pi_{sj}) \to_p 0, \tag{8.17}$$

for any j, $l \le j \le q$, jointly are sufficient for (8.14). Statements (8.15) and (8.16) can be shown similarly as (8.13). For the increments of (8.17), $v_{st} := w'_{st}W_{sj}w_{st}(y_{sj} - \pi_{sj})$, it holds that

$$E(v_{st}|\mathfrak{A}_{s-1}) = 0, \qquad \operatorname{var}(v_{st}|\mathfrak{A}_{s-1}) \le c(w'_{st}w_{st})^2, \qquad l < s \le t, \tag{8.18}$$

where $c < \infty$ is an upper bound on $\|W_{sj}\|^2 \operatorname{var}(y_{sj} - \pi_{sj}|\mathfrak{A}_{s-1})$, $s > l$. From A(ii) and (iv) and the boundedness of $\{y_{sj}\}$, such a bound exists. In particular, (8.18) states that, for any $t > t_0$, $\{v_{st},\ l < s \le t\}$ are the (orthogonal) increments of a square integrable zero mean martingale. Integrating out $y_{l+1}, \ldots, y_{s-1}$ and summing up yields

$$\begin{aligned} E\left(\sum_{s=l+1}^{t} v_{st}|\mathfrak{A}_l\right) &= 0, \\ \operatorname{var}\left(\sum_{s=l+1}^{t} v_{st}|\mathfrak{A}_l\right) &\le cw_t \sup_{l<s\le t} w'_{st}w_{st}. \end{aligned} \tag{8.19}$$

Using A(ii) again,

$$\sup w'_{st}w_{st} \le \sup_{Z\in C} \|Z\|^2/\lambda_{\min}(E_t) \to 0.$$

Since Ew_t is bounded, uniformly in t, (8.17) holds for any j with $1 \le j \le q$, and the remaining condition N(iv) is established under assumption A.

Finally, the Cholesky square root $G_t^{T/2}(\beta)$ can indeed be used as norming quantity in (5.2), since Remark (ii) after Theorem 3 applies. □

PROOF OF REMARK 1. For the moment, assume that $G_t^{T/2}$ denotes the Cholesky square root. Applying the continuity theorem, we can replace $G_t^{T/2}$ in (5.2) by any $p \times p$-matrix M_t with

$$M_t G_t^{-T/2} \to_p I. \tag{8.20}$$

If M_t is the right Cholesky square root of F_t or E_t, then (8.4), the arguments of Fahrmeir and Kaufmann [1985, page 350, Remark (iii)] and repeated application of the continuity theorem imply first $G_t^{T/2}M_t^{-1} \to_p I$ and then (8.20). The more general statements on F_t and E_t follow from Remark (i) on page 349 of the same reference. Formulas (8.12), (8.13) together with (8.26) imply that (8.20) holds with the Cholesky square roots $M_t = G_t^{T/2}(\hat{\beta}_t)$, $H_t^{T/2}$, $H_t^{T/2}(\hat{\beta}_t)$. □

PROOF OF LEMMA 1. Keeping the starting values fixed, it is possible to apply to the conditional likelihood the definition of local asymptotic normality

as given, e.g., in Ibragimov and Has'minskii (1981, page 120). Conditions N(i)–(iii), hence condition A, are sufficient for the application of a central limit theorem for martingales [e.g., Hall and Heyde (1980, Corollary 3.1)]. We obtain $E_t^{-1/2}s_t \to_d N(0, I)$. By Taylor expansion, using (8.4) and (8.12), the LAN condition follows. □

PROOF OF THEOREM 2. Conditionally on the starting values, it follows from the LAN condition and part (iii) of the proof of Theorem 3 that $\{\hat{\beta}_t\}$ is an asymptotically centering sequence of estimators; see Definition 2 of Jeganathan (1982). The restriction to symmetric positive definite square roots made by that author can be dropped. Applying his Theorem 2 it follows that the MLE is regular. The optimality result is obtained utilizing Corollary 1 of the same reference, and Anderson's lemma [e.g., Basawa and Scott (1983), page 51)]. □

Corollary 2 can be shown with the following lemma.

LEMMA 4. *Let the positive semidefinite matrix* A_t *be partitioned as*

$$A_t = \begin{bmatrix} B_t & D_t \\ D_t' & C_t \end{bmatrix}, \qquad t \geq 1.$$

Divergence of the smallest eigenvalue, $\lambda_{\min}(A_t) \to \infty$, *implies*

$$\lambda_{\min}(B_t) \to \infty, \qquad \lambda_{\min}(C_t) \to \infty. \tag{8.21}$$

If, with some constant $\alpha < 1$, *the matrices*

$$\begin{bmatrix} B_t & D_t \\ D_t' & \alpha C_t \end{bmatrix}, \qquad t \geq 1, \tag{8.22}$$

are positive semidefinite, then conversely (8.21) *implies* $\lambda_{\min}(A_t) \to \infty$.

PROOF. Due to the inversion formula for partitioned matrices and positive semidefiniteness of A_t, $\lambda_{\min}(A_t) \to \infty$ is equivalent to

$$\lambda_{\min}\left(B_t - D_t C_t^{-1} D_t'\right) \to \infty, \qquad \lambda_{\min}\left(C_t - D_t' B_t^{-1} D_t\right) \to \infty.$$

This is apparently stronger than (8.21). For the converse, we note that (8.21) and positive semidefiniteness of (8.22) imply that $\alpha C_t - D_t' B_t^{-1} D_t$ is positive semidefinite. Hence, in view of

$$\lambda_{\min}\left(C_t - D_t' B_t^{-1} D_t\right) \geq (1 - \alpha)\lambda_{\min}(C_t) + \lambda_{\min}\left(\alpha C_t - D_t' B_t^{-1} D_t\right),$$

$\lambda_{\min}(C_t - D_t' B_t^{-1} D_t) \to \infty$ follows from $\lambda_{\min}(C_t) \to \infty$. In (8.22), we can assume $\alpha > 0$, since with some value of α, (8.22) is positive semidefinite also for all greater values of α. Rescaling (8.22) from both sides with the matrix $\operatorname{diag}(\alpha^{1/2}, \ldots, \alpha^{1/2}, \alpha^{-1/2}, \ldots, \alpha^{-1/2})$, $\lambda_{\min}(B_t - D_t C_t^{-1} D_t') \to \infty$ follows analogously. □

PROOF OF COROLLARY 2. Applying Lemma 4 with $B_t = (t - l)\Sigma_1^n Y_j Y_j'$, $C_t = n\Sigma_{l+1}^t X_s X_s'$, $D_t = \Sigma_1^n Y_j \Sigma_{l+1}^t X_s'$, it follows immediately that A(iii) implies (5.4).

The converse can be established by showing that (5.4) implies (8.22), with some $\alpha < 1$.

With the vector (λ', μ') partitioned corresponding to $Z_{t,j}$, the increment of the quadratic form (8.22) is

$$\sum_{j=1}^{n} (\lambda' Y_j + \mu' X_t)(Y_j'\lambda + X_t'\mu) - n(1-\alpha)\mu' X_t X_t'\mu. \tag{8.23}$$

For fixed μ, unconstrained minimization of (8.23) over λ is an ordinary least-squares problem. With $M_j = I - Y_j'(\Sigma_k Y_k Y_k')^{-1}\Sigma_k Y_k$, $j = 1, \ldots, n$, we obtain the minimum

$$\mu' X_t \left(\sum_{1}^{n} M_j' M_j \right) X_t'\mu - n(1-\alpha)\mu' X_t X_t'\mu.$$

From the assumption that the system $\mu = Y_j'\lambda$, $j = 1, \ldots, n$, has only solutions with $\mu = 0$, it follows that the matrix $(M_1', \ldots, M_n')$ has full rank. This is equivalent to $\lambda_* = \lambda_{\min}(\Sigma M_j' M_j) > 0$. Choosing $\alpha = 1 - \lambda_*/n$, (8.23) will be nonnegative for any λ, μ. By summation, (8.22) follows. □

PROOF OF THEOREM 3. (i) Asymptotic distribution of the score function: Conditions N(i)–(iii) are sufficient for the application of a central limit theorem for martingales [e.g., Hall and Heyde (1980, Corollary 3.1)]. We obtain

$$\left(A_t^{-1/2} s_t, A_t^{-1/2} G_t A_t^{-T/2}\right) \to_d (V^{1/2} z, V),$$

where z is a standard normal vector independent of $V^{1/2}$. Choosing $V^{1/2}$ as the Cholesky square root and $G_t^{1/2}$ such that $A_t^{-1/2} G_t^{1/2}$ is the Cholesky square root of $A_t^{-1/2} G_t A_t^{-T/2}$, it follows from continuity of this square root that

$$\left(A_t^{-1/2} s_t, A_t^{-1/2} G_t^{1/2}\right) \to_d (V^{1/2} z, V^{1/2}), \tag{8.24}$$

z as above.

(ii) Asymptotic existence and consistency: With the common argument, the Lindeberg condition N(ii) implies the Feller condition

$$\max_{s=1,\ldots,t} E\left(x_s' A_t^{-1} x_s | \mathfrak{A}_{s-1}\right) \to_p 0.$$

Together with N(ii), this implies

$$\lambda_{\min}(A_t) \to \infty. \tag{8.25}$$

By Taylor expansion, using N(ii) and N(iv) and noting that (8.24) implies that $\{\|A_t^{-1/2} s_t\|\}$ is bounded in probability, it follows that for any $\eta > 0$, there exist $\delta > 0$, t_1 with

$$P\left(l_t(\tilde{\beta}) < l_t(\beta) \text{ for all } \tilde{\beta} \in \partial N_t(\delta)\right) \geq 1 - \eta, \qquad t \geq t_1,$$

similar to the proof of Theorem 1 of Fahrmeir and Kaufmann (1985). This includes asymptotic existence of a MLE as well as consistency, in view of (8.25). More precisely, there exists a sequence $\{\hat{\beta}_t\}$ of MLE's such that for any $\eta > 0$,

there exist $\delta > 0$, t_1 with

$$P\big(\hat{\beta}_t \in N_t(\delta)\big) \geq 1 - \eta, \qquad t \geq t_1. \tag{8.26}$$

Since N(ii) and N(iv) imply that $H_t(\beta)$ is positive definite throughout $N_t(\delta)$, with probability converging to one, for any $\delta > 0$, the MLE $\hat{\beta}_t$ is also locally unique.

(iii) Asymptotic normality: By Taylor expansion, we obtain

$$s_t = \tilde{H}_t\big(\hat{\beta}_t - \beta\big),$$

with $\tilde{H}_t = \int_0^1 H_t(\beta + u(\hat{\beta}_t - \beta))\, du$. Somewhat more generally than in Fahrmeir and Kaufmann (1985, proof of Theorem 3), it can be shown that N(ii) and N(iv) and (8.26) imply that (8.24) can be enlarged by

$$A_t^{-1/2}\tilde{H}_t A_t^{-T/2} \to_d V.$$

Applying the continuity theorem to

$$A_t^{-1/2}s_t = \big(A_t^{-1/2}\tilde{H}_t A_t^{-T/2}\big)\big(A_t^{T/2}\big(\hat{\beta}_t - \beta\big)\big),$$

it follows that

$$\big(A_t^{T/2}\big(\hat{\beta}_t - \beta\big), A_t^{-1/2}G_t^{1/2}\big) \to_d (V^{-T/2}z, V^{1/2}),$$

z as above. Premultiplying $A_t^{T/2}(\hat{\beta}_t - \beta)$ by $G_t^{T/2}A_t^{-T/2}$, (7.1) follows with the square root $G_t^{T/2}$ given in Remark (i) after Theorem 3, see part (i) of this proof. If $A_t^{T/2}$ and $G_t^{T/2}A_t^{-T/2}$ are Cholesky square roots, then $G_t^{T/2}$ shares this property. This proves in, particular, Remark (ii). Remark (iii) can be demonstrated similarly as Remark 1. □

Acknowledgments. I thank Ludwig Fahrmeir for many useful discussions on the subject of the paper, and a referee for detailed criticism and the reference to Brillinger and Segundo (1979).

REFERENCES

BASAWA, I. V. and SCOTT, D. J. (1983). *Asymptotic Optimal Inference for Non-ergodic Models. Lectures Notes in Statist.* **17** Springer, Berlin.

BILLINGSLEY, P. (1961). *Statistical Inference for Markov Processes*. Univ. Chicago Press, Chicago.

BRILLINGER, D. R. and SEGUNDO, J. P. (1979). Empirical examination of the threshold model of neuron firing. *Biol. Cybern.* **35** 213–220.

DOBRUSHIN, R. L. (1956). Central limit theorem for nonstationary Markov chains, I, II. *Theory Probab. Appl.* **1** 65–80, 329–383.

DRYGAS, H. (1976). Weak and strong consistency of the least squares estimators in regression models. *Z. Wahrsch. verw. Gebiete* **34** 119–127.

FAHRMEIR, L. and KAUFMANN, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist.* **13** 342–368.

FAHRMEIR, L. and KAUFMANN, H. (1987). Regression models for non-stationary categorical time series. *J. Time Ser. Anal.* To appear.

GOURIEROUX, C., MONFORT, A. and TROGNON, A. (1983). Estimation and test in probit models with serial correlation. CEPREMAP Discussion Paper 8220.

GRETHER, D. M. and MADDALA, G. S. (1982). A time series model with qualitative variables. In *Games, Economic Dynamics and Time Series Analysis* (M. Deistler et al., eds.). Physica, Wien.

HALL, P. and HEYDE, C. C. (1980). *Martingale Limit Theory and Its Application*. Academic, New York.

HAUSER, J. R. and WISNIEWSKI, K. J. (1982). Dynamic analysis of consumer response to marketing strategies. *Management Sci.* **28** 455–486.

HECKMAN, J. J. (1981). Dynamic discrete probability models. In *Structural Analysis of Discrete Data with Econometric Applications* (C. F. Manski and D. McFadden, eds.). MIT Press, Cambridge, Mass.

IBRAGIMOV, I. A. and HAS'MINSKII, R. Z. (1981). *Statistical Estimation, Asymptotic Theory*. Springer, Berlin.

ISAACSON, D. L. and MADSEN, R. W. (1976). *Markov Chains: Theory and Applications*. Wiley, New York.

JACOBS, P. A. and LEWIS, P. A. W. (1983). Stationary discrete autoregressive-moving average time series generated by mixtures. *J. Time Ser. Anal.* **4** 19–36.

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

KAUFMANN, H. (1987). On the uniqueness of the maximum likelihood estimator in quantal and ordinal response models. Preprint.

LAI, T. L., ROBBINS, H. and WEI, C. Z. (1979). Strong consistency of least squares estimates in multiple regression, II. *J. Multivariate Anal.* **9** 343–361.

SWEETING, T. J. (1980). Uniform asymptotic normality of the maximum likelihood estimator. *Ann. Statist.* **8** 1375–1381.

WEDDERBURN, R. W. M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika* **63** 27–32.

INSTITUT FÜR STATISTIK UND
WIRTSCHAFTSGESCHICHTE
UNIVERSITÄT REGENSBURG
UNIVERSITÄTSSTRASSE 31
8400 REGENSBURG
FEDERAL REPUBLIC OF GERMANY