Ann. Inst. Statist. Math.
Vol. 54, No. 3, 543–564 (2002)

POWER DIVERGENCE FAMILY OF TESTS FOR CATEGORICAL TIME SERIES MODELS

KONSTANTINOS FOKIANOS

Department of Mathematics & Statistics, University of Cyprus, P.O. Box 20537, 1678 Nicosia, Cyprus, e-mail: fokianos@ucy.ac.cy

(Received April 3, 2000; revised October 17, 2000)

Abstract. A fundamental issue that arises after fitting a regression model is that of testing the goodness of the fit. Our work brings together the power divergence family of goodness of fit tests and regression models for categorical time series. We show that under some reasonable assumptions, the asymptotic distribution of the power divergence family of goodness of fit tests converges to a normal random variable. This fact introduces a novel method for carrying out goodness of fit tests about a regression model for categorical time series. We couple the theory with some empirical results.

Key words and phrases: Stochastic time dependent covariates, partial likelihood, martingale, logistic regression, multinomial logits, proportional odds, power.

1. Introduction

The aim of this contribution is to link the power divergence family of goodness of fit tests with the regression theory of categorical time series. A categorical time series is a sequence of dependent observations taking qualitative values. The introduction of generalized linear models (see McCullagh and Nelder (1989)) had a profound impact on the modeling aspects of such data. Some examples can be found in the work of Bonney (1987), Fahrmeir and Kaufmann (1987), Kaufmann (1987), Korn and Whittemore (1979), Liang and Zeger (1989), Muenz and Rubinstein (1985), Stern and Coe (1984) and more recently Slud and Kedem (1994), Brillinger (1996), Fokianos and Kedem (1998, 1999). However a central question that arises after fitting a regression model is that of the adequacy of the fit. In the context of generalized linear models, goodness of fit is examined either by a Pearson chi-square test or by the residual deviance (see McCullagh and Nelder (1989), Fahrmeir and Tutz (1994)). However, in the special case of regression models for categorical time series this approximation can be poor due to the fact that data are sparse. Thus, some other techniques should be developed in order to take advantage of the fitting output for answering questions regarding the fit of the model.

The topic of testing the goodness of fit of a regression model for a categorical time series has been addressed by some authors either by conducting a chi-square test or by inspection of the residuals. For example, in Brillinger (1996), the author examines the so called uniform residuals which are based on the probability integral transformation under the fitted model. If the parameter values are known, then we have a uniform distribution and the uniform residuals can be used to examine the adequacy of the fit either by constructing probability plots or by using them in some other established way.

Another approach to the goodness of fit question is that of Slud and Kedem (1994) and Fokianos and Kedem (1998) that relies on an appealing idea of Schoenfeld (1980). These authors classify the response variable according to some partition of the covariate space. Then, the goodness of fit test is based on the new grouping of observations. Although this is a sensible approach, it depends on the partition of the covariate space which occasionally might be rather large.

Recent work in the area of independent and not identically distributed data shows that the family of power divergence tests can serve as a general framework for answering questions regarding the goodness of fit. The power divergence family of goodness of fit tests has been introduced by Cressie and Read (1984) as a generalization of the well known Pearson's X^2 and likelihood ratio G^2 test statistics. Denote by α_λ the deviation—or power divergence—between observed and expected counts, that is α_λ is a distance which is given by

$$\alpha_\lambda(\text{observed}, \text{expected}) = \frac{2}{\lambda(\lambda+1)} \text{ observed} \left[\left(\frac{\text{observed}}{\text{expected}} \right)^\lambda - 1 \right].$$

Notice that the above quantity compares the fraction of the observed counts divided by the expected counts raised to the power λ with 1. Then the power divergence family of test statistics indexed by a parameter $\lambda \in R$, say $I(\lambda)$, is just the sum over all cells of these deviations. Namely,

$$I(\lambda) = \sum_{\text{cells}} \alpha_\lambda(\text{observed}, \text{expected}).$$

It is straightforward to verify that the statistic $I(\lambda)$ reduces to Pearson's X^2 when $\lambda = 1$ and to likelihood ratio G^2 when $\lambda \to 0$. Some other interesting cases worth mentioning include $\lambda \to -1$, $\lambda = -1/2$ and $\lambda = -2$. For those particular values we obtain the minimum discrimination information statistic, the Freeman-Tukey statistic and the Neyman-modified X^2 statistic, respectively. An extended study of the properties of the power divergence family of goodness of fit tests is given by Read and Cressie (1988). The authors suggest the value of $\lambda = 2/3$ for checking adequacy of a hypothesized model for independent data. Notice that for $\lambda = 2/3$ the resulting test statistic lies between the Pearson's and likelihood ratio test statistics. Furthermore, the authors offer an extensive discussion of the large sample properties of the test statistic under both fixed and increasing cells assumptions (Read and Cressie (1988), Chapter 4). Fixed cells assumptions imply that the number of the cells remains fixed while all group sizes tend suitably to infinity. Increasing cells assumptions—or so called sparseness assumptions—signify the fact that the number of the cells tends to infinity. Under fixed cells assumption the asymptotic distribution of the test statistic is approximated by a chi-square distribution with some degrees of freedom given that the hypothesis of the correct model is true (Read and Cressie (1988), Chapter 4). In contrast, increasing cells assumptions imply that the asymptotic null distribution of the test statistic is approximated by a normal random variable with mean and variance depending on the parameter λ. Early work on necessary asymptotic theory under relatively different increasing cells assumption has been done by Morris (1975), Weiss (1976) and Holst (1972). References on the asymptotic distribution of the X^2 and G^2 test statistics under sparseness assumptions include Koehler (1986) who demonstrated asymptotic normality of the likelihood ratio statistic for log—linear models admitting closed—form maximum likelihood estimates, McCullagh (1986) who

pointed out the use of conditional distribution of the X^2 and G^2 statistics given a sufficient statistic in the context of linear exponential family models and Dale (1986) who proved the asymptotic normality of the X^2 and G^2 statistics for increasing cells with bounded expectations. More recently, Osius and Rojek (1992) generalized the aforementioned results by proving asymptotic normality of $I(\lambda)$ under sparseness assumptions (see also Osius (1985) for binary data). Finally, we mention that Gleser and Moore (1985) investigate the effect of serial dependence on the test statistic, under fixed cells assumptions. These authors demonstrate that positive dependence—according to their definition—is confounded with lack of fit. Therefore, the null asymptotic distribution of $I(\lambda)$ behaves differently according to whether or not the data are sparse. We refer the reader to Read and Cressie (1988) for more details on both small and large properties of the power divergence family $\{I(\lambda), \lambda \in R\}$.

Increasing cells asymptotics turn out to be very useful for regression models of categorical time series because those data are sparse. Our contribution is to extend the power divergence family to accommodate categorical time series data and study the asymptotic distribution of the test statistic under the null hypothesis of the correct model. We review some theory for regression models for categorical time series in the sequel while Section 3 covers the proposed test statistic and gives its asymptotic distribution. We conclude our presentation by illustrating some empirical results. The proof of the main theorem (Theorem 3.1) is given in the Appendix.

2. Regression models for categorical time series

Suppose we observe a nonstationary categorical time series, say $\{\boldsymbol{Y}_s, s = 1, \ldots, T\}$. Let m denote the possible number of categories for each observation. We assume that the s-th observation is given by the vector $\boldsymbol{y}_s = (y_{s1}, \ldots, y_{sq})'$ of length q, with elements

$$y_{sj} = \begin{cases} 1, & \text{if the } j\text{-th category is observed at time } s \\ 0, & \text{otherwise} \end{cases}$$

for $s = 1, \ldots, T$ and $q = m - 1$. In addition, we denote by $\boldsymbol{p}_s = (p_{s1}, \ldots, p_{sq})'$ the vector of conditional probabilities given $\mathcal{F}_{s-1}$, that is $p_{sj} = P(y_{sj} = 1 \mid \mathcal{F}_{s-1})$, $j = 1, \ldots, q$, $s = 1, \ldots, T$. Here $\mathcal{F}_{s-1}$ stands for the whole information up to and including time s. Clearly, $y_{sm} = 1 - \sum_{j=1}^{q} y_{sj}$ and $p_{sm} = 1 - \sum_{j=1}^{q} p_{sj}$. Finally, we let $\boldsymbol{Z}_{s-1}$ to denote a $p \times q$ matrix that represents a covariate process. The latter may include past values of the process or/and any other auxiliary processes. Let

$$\boldsymbol{p}_s(\boldsymbol{\beta}) = \boldsymbol{h}(\boldsymbol{Z}'_{s-1}\boldsymbol{\beta}). \tag{2.1}$$

Here $\boldsymbol{\beta}$ denotes a p-dimensional vector of time invariant unknown parameters and the function $\boldsymbol{h}(\cdot)$ is the so called link function.

Important models that fall under the above framework include the multinomial logits and the cumulative odds model. The multinomial logits model (see for example, Agresti (1990)), which is used for the analysis of nominal time series is a special case of (2.1). It is given by

$$p_{sj} = \frac{\exp(\boldsymbol{\beta}'_j \boldsymbol{z}_{s-1})}{1 + \sum_{i=1}^{q} \exp(\boldsymbol{\beta}'_i \boldsymbol{z}_{s-1})}, \qquad j = 1, \ldots, q \tag{2.2}$$

where $\boldsymbol{\beta}_j$ is a d-dimensional regression parameter and $\boldsymbol{z}_{s-1}$ is a vector of stochastic time dependent covariates of the same dimension.

For the analysis of ordinal time series the cumulative odds model (see McCullagh (1980)) is given by

$$P(\boldsymbol{Y}_s \leq j \mid \mathcal{F}_{s-1}) = F(\theta_j + \boldsymbol{\gamma}'\boldsymbol{z}_{s-1}) \tag{2.3}$$

where F denotes a cumulative distribution function, $\boldsymbol{\gamma}$ is a d-dimensional regression parameter, $\boldsymbol{z}_{s-1}$ stands for a vector of stochastic time dependent covariates of the same dimension while $-\infty = \theta_0 < \theta_1 < \cdots < \theta_m = \infty$. Common choices for F include the logistic—which gives rise to the so called proportional odds model—the normal and the complementary log-log distribution functions. Further examples and modeling strategies are discussed in Fahrmeir and Tutz ((1994), Chapter 6).

The central issue is to estimate the vector of parameters $\boldsymbol{\beta}$. Since the data are dependent, we attack the problem through the partial likelihood methodology which was suggested by Cox (1975). Partial likelihood approaches successfully the problem of estimation and testing by means of martingale theory. It has been proved a useful tool for time series following generalized linear models (see for example, Wong (1986), Slud and Kedem (1994), Fokianos and Kedem (1998) among others). According to Fokianos and Kedem (1998), the partial likelihood (PL) function relative to $\boldsymbol{\beta}$, $\mathcal{F}_s$, and the data $\{\boldsymbol{y}_s, s = 1, \ldots, T\}$, is given by

$$\mathrm{PL}(\boldsymbol{\beta}) = \prod_{s=1}^{T} \prod_{j=1}^{m} p_{sj}(\boldsymbol{\beta})^{y_{sj}}.$$

Hence the partial log-likelihood is

$$pl_T(\boldsymbol{\beta}) = \sum_{s=1}^{T} \sum_{j=1}^{m} y_{sj} \log p_{sj}(\boldsymbol{\beta}). \tag{2.4}$$

The partial score is the p-dimensional vector

$$\begin{aligned} S_T(\boldsymbol{\beta}) &= \sum_{s=1}^{T} \sum_{j=1}^{m} y_{sj} \frac{1}{p_{sj}(\boldsymbol{\beta})} \frac{\partial p_{sj}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} \\ &= \sum_{s=1}^{T} \boldsymbol{W}_s'(\boldsymbol{\beta})(\tilde{\boldsymbol{y}}_s - \tilde{\boldsymbol{p}}_s(\boldsymbol{\beta})), \end{aligned} \tag{2.5}$$

where $\tilde{\boldsymbol{y}}_s$ and $\tilde{\boldsymbol{p}}_s$ denote the m-dimensional vectors $(y_{s1}, \ldots, y_{sm})$ and $(p_{s1}, \ldots, p_{sm})$ respectively, while $\boldsymbol{W}_s(\boldsymbol{\beta})$ stands for the $m \times p$ matrix with rows

$$w_{sj}(\boldsymbol{\beta}) = \frac{1}{p_{sj}(\boldsymbol{\beta})} \frac{\partial p_{sj}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}, \qquad j = 1, \ldots, m. \tag{2.6}$$

The maximum partial likelihood estimator $\hat{\boldsymbol{\beta}}$ is a consistent root of the equation $S_T(\boldsymbol{\beta}) = 0$. For categorical time series data we have that the conditional information matrix is given by

$$\boldsymbol{G}_T(\boldsymbol{\beta}) = \sum_{s=1}^{T} \mathrm{Var}[\boldsymbol{W}_s'(\boldsymbol{\beta})(\tilde{\boldsymbol{y}}_s - \tilde{\boldsymbol{p}}_s(\boldsymbol{\beta})) \mid \mathcal{F}_{s-1}] \tag{2.7}$$

$$= \sum_{s=1}^{T} \boldsymbol{W}_s'(\boldsymbol{\beta})\boldsymbol{\Sigma}_s(\boldsymbol{\beta})\boldsymbol{W}_s(\boldsymbol{\beta})$$

with $\boldsymbol{\Sigma}_s(\boldsymbol{\beta})$ the conditional covariance matrix of $\tilde{\boldsymbol{y}}_s$ with generic element

$$\sigma_s^{(ij)}(\boldsymbol{\beta}) = \begin{cases} -p_{si}(\boldsymbol{\beta})p_{sj}(\boldsymbol{\beta}) & \text{if } i \neq j \\ p_{si}(\boldsymbol{\beta})(1 - p_{si}(\boldsymbol{\beta})) & \text{if } i = j \end{cases}$$

for $i, j = 1, \ldots, m$. Asymptotic properties of the maximum partial likelihood estimator $\hat{\boldsymbol{\beta}}$ are examined via the score function and the conditional information matrix (Arjas and Haara (1987), Andersen and Gill (1982), Wong (1986)). It turns out that the following theorem holds (Fokianos and Kedem (1998)):

THEOREM 2.1. *Consider model* (2.1) *and assume the following*: *Assumption* (A)

A.1 *The parameter* $\boldsymbol{\beta}$, *belongs to an open set* $B \subseteq R^p$.

A.2 *The covariate matrices* $\boldsymbol{Z}_{s-1}$ *almost surely lie in a nonrandom compact subset* Γ *of* $R^{p\times q}$ *such that* $P[\sum_{s=1}^{T} \boldsymbol{Z}_{s-1}\boldsymbol{Z}_{s-1}' > \boldsymbol{0}] = 1$. *Furthermore we assume that* $\boldsymbol{Z}_{s-1}'\boldsymbol{\beta}$ *lies almost surely in the domain of* h *for all* $\boldsymbol{Z}_{s-1} \in \Gamma$ *and* $\boldsymbol{\beta} \in B$.

A.3 *The true probability measure is given by* $P_{\boldsymbol{\beta}}$.

A.4 *The link function* $\boldsymbol{h}$ *is twice continuously differentiable*, $\det[\partial \boldsymbol{h}(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma}] \neq 0$

A.5 *There is a probability measure* μ *on* $R^{p\times q}$ *such that* $\int_{R^{p\times q}} \boldsymbol{Z}\boldsymbol{Z}'\mu(d\boldsymbol{Z})$ *is positive definite, such that under* (2.1) *for all Borel sets* $A \subset R^{p\times q}$ *we have*

$$\frac{1}{T}\sum_{s=1}^{T} I_{[\boldsymbol{Z}_{s-1}\in A]} \xrightarrow{p} \mu(A), \qquad as \quad T \to \infty$$

at the true parameter $\boldsymbol{\beta}$.

Then:

1. *There exists a locally unique maximum partial likelihood estimator*, $\hat{\boldsymbol{\beta}}$, *with probability tending to* 1, *as* $N \to \infty$.

2. *The estimator is consistent and asymptotically normal*,

$$\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$$

and

$$\sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} \mathcal{N}(0, \boldsymbol{G}^{-1}(\boldsymbol{\beta})),$$

as $T \to \infty$.

3. *The following is true*:

$$\sqrt{T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \frac{1}{\sqrt{T}}\boldsymbol{G}(\boldsymbol{\beta})^{-1}S_T(\boldsymbol{\beta}) \xrightarrow{p} 0.$$

Notice that $\boldsymbol{G}$ denotes the limit in probability of the conditional information matrix, that is

$$\frac{\boldsymbol{G}_T(\boldsymbol{\beta})}{T} \xrightarrow{p} \int_{R^{p\times q}} \boldsymbol{W}'(\boldsymbol{\beta})\boldsymbol{\Sigma}(\boldsymbol{\beta})\boldsymbol{W}(\boldsymbol{\beta})\mu(d\boldsymbol{Z}) \equiv \boldsymbol{G}(\boldsymbol{\beta}),$$

where $\boldsymbol{W}(\boldsymbol{\beta})$ denotes that $m \times p$ matrix with rows

$$w_j(\boldsymbol{\beta}) = \frac{1}{h_j(\boldsymbol{Z}'\boldsymbol{\beta})}\frac{\partial h_j(\boldsymbol{Z}'\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}, \qquad j = 1, \ldots, m, \tag{2.8}$$

and Σ has generic element

$$\sigma^{(ij)}(\boldsymbol{\beta}) = \begin{cases} -h_i(\boldsymbol{Z}'\boldsymbol{\beta})h_j(\boldsymbol{Z}'\boldsymbol{\beta}) & \text{if } i \neq j \\ h_i(\boldsymbol{Z}'\boldsymbol{\beta})(1 - h_i(\boldsymbol{Z}'\boldsymbol{\beta})) & \text{if } i = j \end{cases}$$

for $i, j = 1, \ldots, m$. The proof of Theorem 2.1 and some further discussion on the regularity conditions can be found in Fokianos and Kedem (1998). It is essential to point out that we do not assume stationarity of the observed time series. In addition, we do not postulate any Markov property on the process. Notice that we use slightly different notation from Fokianos and Kedem (1998).

In conclusion, we gave a brief account of asymptotic theory for the general regression model for categorical time series. We established some notation that will be found useful in the rest of the article and stated some asymptotic results regarding the maximum partial likelihood estimator. An essential question that is posed immediately is the quality of the fit which is answered in the next section.

3. A family of goodness of fit tests

Suppose that we observe a categorical time series, say $\{\boldsymbol{Y}_s, s = 1, \ldots, T\}$, as in Section 2. We generalize the family of power divergence tests to accommodate dependent data by introducing the following quantity

$$u_s(\boldsymbol{\beta}) \equiv \frac{2}{\lambda(\lambda+1)} \sum_{j=1}^{m} y_{sj} \left[\left(\frac{y_{sj}}{p_{sj}(\boldsymbol{\beta})} \right)^{\lambda} - 1 \right]. \tag{3.1}$$

Notice that we drop notation that depends on λ for ease of presentation. Equation (3.1) simply states that at each time instance s, we calculate the deviation between the observed data and the corresponding transition probabilities. Now, since each $\{y_{sj}, j = 1, \ldots, m\}$ can take values 0 or 1, for $s = 1, \ldots, T$ we obtain that equation (3.1) is defined for values of λ greater than -1. Summing up all the deviations across time we conclude that the quantity $\sum_{s=1}^{T} u_s(\boldsymbol{\beta})$ is the analog of the power divergence statistic for dependent categorical data. Thus, if we calculate the conditional expectation of $\{u_s(\cdot), s = 1, \ldots, T\}$ under the correct model, the residual process should evolve around 0. Let us be more specific. If

$$e_s(\boldsymbol{\beta}) \equiv E[u_s(\boldsymbol{\beta}) \mid \mathcal{F}_{s-1}] = \frac{2}{\lambda(\lambda+1)} \sum_{j=1}^{m} p_{sj}(\boldsymbol{\beta}) \left[\left(\frac{1}{p_{sj}(\boldsymbol{\beta})} \right)^{\lambda} - 1 \right], \tag{3.2}$$

denotes the conditional expectation of $u_s(\boldsymbol{\beta})$ given the past process, then the difference $\sum_{s=1}^{T} u_s(\boldsymbol{\beta}) - \sum_{s=1}^{T} e_s(\boldsymbol{\beta})$ is clearly a zero mean martingale, by construction. If the model is correct then the centered process should fluctuate around 0. Therefore any large value of $|\sum_{s=1}^{T} u_s(\boldsymbol{\beta}) - \sum_{s=1}^{T} e_s(\boldsymbol{\beta})|$ will lead evidence against the null hypothesis. However, we replace $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$, the maximum partial likelihood estimator. It turns out that the quantity

$$I_T(\cdot) = \sum_{s=1}^{T} [u_s(\cdot) - e_s(\cdot)] \tag{3.3}$$

evaluated at the maximum partial likelihood estimator $\hat{\boldsymbol{\beta}}$, is approximated by a zero mean square integrable martingale that satisfies all the conditions for an application of

a central limit theorem. In addition, the asymptotic variance of this martingale can be calculated explicitly by

$$\xi_T(\boldsymbol{\beta}) = \frac{4}{\lambda^2(\lambda+1)^2}\left\{\sum_{s=1}^{t} v_s'(\boldsymbol{\beta})\boldsymbol{\Sigma}_s(\boldsymbol{\beta})v_s(\boldsymbol{\beta}) - c_T'(\boldsymbol{\beta})\boldsymbol{G}_T^{-1}(\boldsymbol{\beta})c_T'(\boldsymbol{\beta})\right\}. \tag{3.4}$$

The quantities $c_T(\cdot)$, $v_s(\cdot)$ are defined in the Appendix through equations (A.4) and (A.9), while $\boldsymbol{\Sigma}_s(\boldsymbol{\beta})$ is the conditional covariance matrix of $\tilde{\boldsymbol{y}}_s$ and $\boldsymbol{G}_T(\boldsymbol{\beta})$ is the conditional information matrix (see (2.7)). Our basic result affirms that the process (3.3) normalized by the expression (3.4) converges to the standard normal distribution. The proof of this theorem is given in the Appendix.

THEOREM 3.1. *Assume Assumption* (A) *as in Theorem* 2.1. *Then, under the hypothesis that model* (2.1) *holds and for* $\lambda > -1$, *the following is true*:

$$\boldsymbol{I}_\lambda = \frac{I_T(\hat{\boldsymbol{\beta}})}{\sqrt{\xi_T(\hat{\boldsymbol{\beta}})}} \to \mathcal{N}$$

in distribution, as $T \to \infty$, *where* $\mathcal{N}$ *is a standard normal random variable. Recall that* $\hat{\boldsymbol{\beta}}$ *stands for the maximum partial likelihood estimator while the quantities* $\{I_T(\hat{\boldsymbol{\beta}}), \xi_T(\hat{\boldsymbol{\beta}})\}$ *are defined by the equations* (3.3), (3.4) *respectively. Large values of* $|\boldsymbol{I}_\lambda|$ *lead to the rejection of the null hypothesis.*

Theorem 3.1 states that the asymptotic distribution of the process I_T depends on λ. Thus, the choice of λ is of considerable importance in applications. Based on our simulations, we recommend values that fall between -1 and 1. The approximation is quite sensitive to values of λ greater than 1 as some of the simulated examples show. This fact is explained by observing that the definition of the test statistic requires all the transition probabilities to stay bounded away from 0. In addition, the derivatives of the link function with respect to the parameter $\boldsymbol{\beta}$ have to be bounded, for all t. Both of these conditions are met because of Assumption (A). Indeed, the compactness Assumption (A.2) jointly with the differentiability Assumption (A.4) provide necessary bounding conditions. However the approximation deviates from the asserted normality when the transition probabilities are small because of the fact that the function $x^{-\lambda}$, for x between 0 and 1 and $\lambda \geq 1$, grows to infinity when x takes values near 0. Therefore, the test statistic is sensitive to values of $\lambda > 1$. The next section—which investigates the empirical performance of the test statistic—lists some examples where this situation occurs. In addition we provide some empirical assessment of the power of the test. There are several questions that need to be addressed in the context of the test statistic's power. We leave untouched this area mentioning that even for the case of independent data, optimality of the test has been proved in certain cases (equiprobable model) while studies about the power of the tests introduced by the power divergence family are sparse in the literature (see Read and Cressie (1988), Chapter 8).

However, optimality results regarding the test statistic can be derived when the latter is viewed as a score test. To make this point clear consider first equations (3.2) and (3.3) to get that

$$I_T(\boldsymbol{\beta}) = \sum_{s=1}^{T}\sum_{j=1}^{m}(y_{sj} - p_{sj}(\boldsymbol{\beta}))\left[\left(\frac{1}{p_{sj}(\boldsymbol{\beta})}\right)^{\lambda} - 1\right] \tag{3.5}$$

$$= \sum_{s=1}^{T}\sum_{j=1}^{q}(y_{sj} - p_{sj}(\boldsymbol{\beta}))\left[\left(\frac{1}{p_{sj}(\boldsymbol{\beta})}\right)^{\lambda} - \left(\frac{1}{1-\sum_{i=1}^{q} p_{si}(\boldsymbol{\beta})}\right)^{\lambda}\right]$$

by recalling that $\sum_{j=1}^{m} y_{sj} = 1$, $\sum_{j=1}^{m} p_{sj}(\boldsymbol{\beta}) = 1$, $q = m - 1$ and ignoring the factor $2/\lambda(\lambda+1)$. If we define the following $q \times 1$ vector

$$\boldsymbol{d}_{\lambda}(\boldsymbol{x}) = \left[\left(\frac{1}{x_1}\right)^{\lambda} - \left(\frac{1}{1-\sum_{i=1}^{q} x_i}\right)^{\lambda}, \ldots, \left(\frac{1}{x_q}\right)^{\lambda} - \left(\frac{1}{1-\sum_{i=1}^{q} x_i}\right)^{\lambda}\right]'$$

for $\boldsymbol{x} = (x_1, \ldots, x_q)'$ with $0 < x_i < 1$ for $i = 1, \ldots, q$ and $\sum_i x_i < 1$, then equation (3.5) is rewritten as

$$I_T(\boldsymbol{\beta}) = \sum_{s=1}^{T} \boldsymbol{d}_{\lambda}'(\boldsymbol{p}_s(\boldsymbol{\beta}))(\boldsymbol{y}_s - \boldsymbol{p}_s(\boldsymbol{\beta})). \tag{3.6}$$

We define the $q \times q$ matrices

$$\boldsymbol{H} = \begin{bmatrix} h_1(1-h_1) & -h_1h_2 & \ldots & -h_1h_q \\ -h_1h_2 & h_2(1-h_2) & \ldots & -h_2h_q \\ \vdots & \vdots & \vdots & \vdots \\ h_1h_q & -h_2h_q & \ldots & h_q(1-h_q) \end{bmatrix} \tag{3.7}$$

and

$$\nabla \boldsymbol{h} = \begin{bmatrix} \nabla' h_1 \\ \nabla' h_2 \\ \vdots \\ \nabla' h_q \end{bmatrix}. \tag{3.8}$$

Equations (3.6)–(3.8) are useful on specifying the following q-dimensional function

$$\boldsymbol{f}_{\lambda}'(\boldsymbol{x}) = \boldsymbol{d}_{\lambda}'(\boldsymbol{h}(\boldsymbol{x}))(\nabla \boldsymbol{h}(\boldsymbol{x}))^{-1}\boldsymbol{H}(\boldsymbol{x}). \tag{3.9}$$

In particular, if we let

$$\boldsymbol{h}(\boldsymbol{x}) = \left(\frac{\exp(x_1)}{1+\sum_{i=1}^{q}\exp(x_i)}, \ldots, \frac{\exp(x_q)}{1+\sum_{i=1}^{q}\exp(x_i)}\right)'$$

and

$$\mathbf{logit}(\boldsymbol{p}_s(\boldsymbol{\beta})) = \left[\log\left(\frac{p_{s1}(\boldsymbol{\beta})}{1-\sum_{i=1}^{q} p_{si}(\boldsymbol{\beta})}\right), \ldots, \log\left(\frac{p_{sq}(\boldsymbol{\beta})}{1-\sum_{i=1}^{q} p_{si}(\boldsymbol{\beta})}\right)\right]',$$

then we obtain

$$\boldsymbol{f}_{\lambda}'(\boldsymbol{h}(\mathbf{logit}(\boldsymbol{p}_s(\boldsymbol{\beta})))) = \boldsymbol{d}_{\lambda}'(\boldsymbol{p}_s(\boldsymbol{\beta}))$$

from (3.9).

Consider the enlarged model

$$\boldsymbol{p}_s(\boldsymbol{\beta}) = \boldsymbol{h}(\boldsymbol{Z}_{s-1}'\boldsymbol{\beta} + \psi \boldsymbol{f}_{\lambda}(\boldsymbol{Z}_{s-1}'\boldsymbol{\beta})). \tag{3.10}$$

We argue that the score test for testing the hypothesis $\psi = 0$ is equivalent to (3.6). Recall (2.4) to obtain

$$\begin{aligned}\left[\frac{dpl_T(\beta;\psi)}{d\psi}\right]_{\psi=0} &= \sum_{t=1}^{N}\sum_{j=1}^{m}\frac{y_{sj}}{h_j(Z_{s-1}'\beta)}\nabla' h_j((Z_{s-1}'\beta)\boldsymbol{f}_\lambda(Z_{s-1}'\beta)) \\ &= \sum_{t=1}^{N}\Bigg[y_{s1}\left(1-p_{s1}(\beta),-p_{s2}(\beta),\ldots,-p_{sq}(\beta)\right) \\ &\quad +y_{s2}\left(-p_{s1}(\beta),1-p_{s2}(\beta),\ldots,-p_{sq}(\beta)\right) \\ &\quad +\cdots+\left(1-\sum_{i=1}^{q}y_{si}\right)\left(-p_{s1}(\beta),-p_{s2}(\beta),\ldots,-p_{sq}(\beta)\right)\Bigg]\boldsymbol{d}_\lambda(\boldsymbol{p}_s(\beta)) \\ &= \sum_{s=1}^{T}\boldsymbol{d}_\lambda'(\boldsymbol{p}_s(\beta))(\boldsymbol{y}_s-\boldsymbol{p}_s(\beta))\end{aligned}$$

after some calculations. Thus the family of power divergence test statistics can be derived as a score test by considering the enlarged model (3.10). Obviously, the family of test statistics (3.6) can be used to detect whether or not additional covariates are needed in a regression model. However, we recommend some other established ways (see for example, Kaufmann (1987) and Li (1991)) for performing tests regarding additional covariates to the model since our earlier discussion shows that the choice of parameter λ for (3.6) is crucial.

Remark 3.1. Another work that addresses the question of goodness of fit for a regression model in the context of categorical time series is that of Fokianos and Kedem (1998). The authors generalize the results by Slud and Kedem (1994) who dealt only with the case of binary time series. Their method relies on earlier work by Schoenfeld (1980) who classify the responses according to a partition of the covariate space . We only mention that Fokianos and Kedem (1998) prove that if the covariate space is partitioned into k sets, then a certain goodness of fit test is asymptotically chi-square distributed with degrees of freedom equal to kq (see Fokianos and Kedem (1998), Proposition 4.1). We compare this test with the family of power divergence test statistics in the next section.

Let us summarize the main results. We proposed and examined a goodness of fit test for regression models for categorical time series. We showed that testing the null model can be based on the result of Theorem 3.1. The test is a simple extension of the power divergence family of goodness of fit tests for independent data. Furthermore, our method can be viewed as a score test for a certain enlarged model. We study its performance in the next section.

4. Empirical results

We present a limited simulation study to demonstrate empirically some aspects of the theoretical results. All the simulations were run 500 times.

Initially, we simulate a categorical time series with $m = 3$ categories according to the multinomial logits model (2.2). Here we choose β_1 and β_2 to be the three dimensional

Table 1. Achieved and nominal significance levels for testing the fit of the multinomial logits model with different values of λ and for different sample sizes. The data have been generated according to model (2.2) with $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)' = (-0.25, 0.50, 1, 0.50, -0.25, -1)'$ and $\boldsymbol{z}_t = (1, x_t, \cos(\pi t/12))$. Here x_t stands for an autoregressive process of order 1 and coefficient equal to 0.2. The number of simulations is 500.

λ	$T = 50$			$T = 300$		
	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
−0.8	0.108	0.062	0.012	0.104	0.054	0.018
−0.6	0.088	0.046	0.014	0.128	0.058	0.010
−0.4	0.088	0.048	0.012	0.094	0.056	0.010
−0.2	0.080	0.040	0.012	0.072	0.024	0.012
0	0.090	0.046	0.018	0.090	0.052	0.008
0.2	0.072	0.038	0.014	0.114	0.058	0.018
0.4	0.058	0.038	0.014	0.106	0.046	0.008
0.6	0.060	0.042	0.016	0.082	0.038	0.008
0.8	0.066	0.042	0.016	0.082	0.046	0.018
1	0.048	0.030	0.018	0.070	0.034	0.014

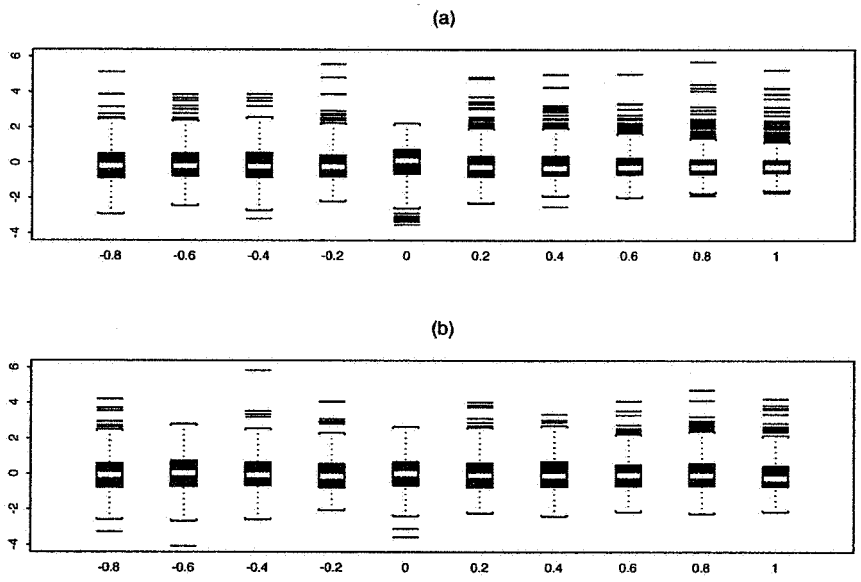

Fig. 1. Boxplots of the values of the power divergence statistic for testing the fit of the multinomial logits model. Each boxplot corresponds to a different value of the parameter λ which varies from -0.8 to 1 with step equal to 0.2. The data have been generated according to model (2.2) with $\boldsymbol{\beta} = (-0.25, 0.50, 1, 0.50, -0.25, -1)'$ and $\boldsymbol{z}_t = (1, x_t, \cos(\pi t/12))$. Here x_t stands for an autoregressive process of order 1 and coefficient equal to 0.2 and the number of simulations equals to 500. (a) $T = 50$. (b) $T = 300$.

vectors $(-0.25, 0.50, 1)'$ and $(0.50, -0.25, -1)'$, respectively. Thus $\boldsymbol{\beta}$ is the six dimensional vector which consists of all the elements of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. The covariate process is given by $\boldsymbol{z}_{t-1} = (1, x_t, \cos(\pi t/12))'$ with $\{x_t, t = 1, \ldots, T\}$ an autoregressive process of order 1

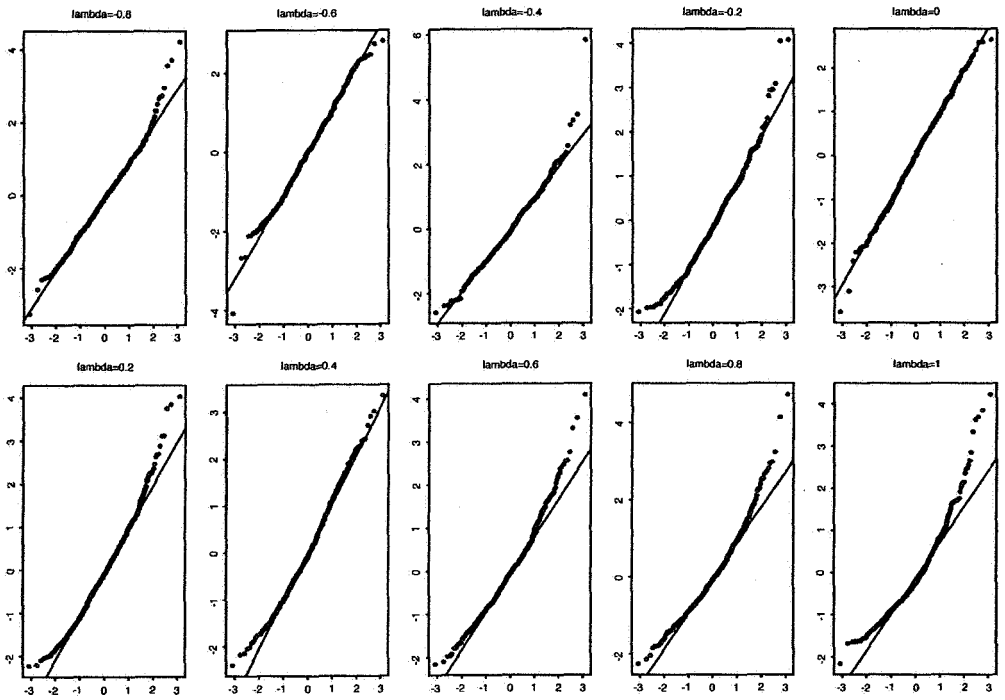

Fig. 2. QQ plots of the values of the power divergence statistic for testing the fit of the multinomial logits model. Each plot corresponds to a different value of the parameter λ which varies from -0.8 to 1 with step equal to 0.2. The data have been generated according to model (2.2) with $\boldsymbol{\beta} = (-0.25, 0.50, 1, 0.50, -0.25, -1)'$ and $\boldsymbol{z}_t = (1, x_t, \cos(\pi t/12))$. Here x_t stands for an autoregressive process of order 1 and coefficient equal to 0.2. The length of each time series is 300 while the number of simulations is 500.

and coefficient equal to 0.2. We investigate both the small and large sample performance of the test statistic by generating $T = 50$ and $T = 300$ observations, respectively. The achieved significance levels of the test statistic I_λ with different values of the parameter λ and for different sample sizes are tabulated in Table 1. Notice that λ assumes values from -0.8 to 1 with step equal to 0.2. In particular, recall that for $\lambda = 0$ we obtain the deviance while for $\lambda = 1$ we obtain the Pearson's chi-square statistic. We observe that the achieved significance levels are in close agreement with the nominal levels when $T = 300$ and deviate when $T = 50$, especially for positive values of the parameter λ. In particular notice that the nominal level of significance $\alpha = 0.01$ is attained in most of the cases for both large and small sample sizes.

Figures 1(a) and 1(b) display box and whisker plots of the values of the test statistic I_λ for different values of the parameter λ when $T = 50$ and $T = 300$ respectively. Thus the first boxplot—starting from the left side of the figure—illustrates boxplot of the values of I_λ for $\lambda = -0.8$. The second boxplot demonstrates the same information but now for $\lambda = -0.6$, and so on. Notice that positive values of λ do indicate a skewed distribution for $T = 50$ (see Fig. 1(a)) while as T grows larger the plots show that the limiting distribution is symmetric centered at zero with most of the data lying from -3 to 3. This fact is in accordance with Fig. 2 which displays QQ-plots of the simulated values of the power divergence statistic for $\lambda = -0.8, -0.6, \ldots, 0.8, 1$. We do not detect any

Table 2. Achieved and nominal significance levels for testing the fit of the proportional odds model with different values of λ and for different sample sizes. The data have been generated according to model (2.3) with $\boldsymbol{\beta} = (1, 1.50, -2, 1)'$ and $\boldsymbol{z}_t = (1, x_t, \cos(\pi t/12))$. Here x_t stands for an autoregressive process of order 1 with coefficient 0.2. The number of simulations is 500.

λ	$T = 50$			$T = 300$		
	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
−0.8	0.178	0.122	0.054	0.118	0.062	0.022
−0.6	0.210	0.164	0.092	0.120	0.064	0.018
−0.4	0.212	0.152	0.068	0.104	0.048	0.016
−0.2	0.120	0.086	0.040	0.130	0.075	0.026
0	0.084	0.042	0.008	0.102	0.042	0.006
0.2	0.090	0.078	0.048	0.137	0.091	0.036
0.4	0.130	0.096	0.048	0.084	0.044	0.014
0.6	0.102	0.084	0.056	0.042	0.026	0.012
0.8	0.118	0.098	0.068	0.040	0.020	0.012
1	0.080	0.068	0.046	0.022	0.014	0.006

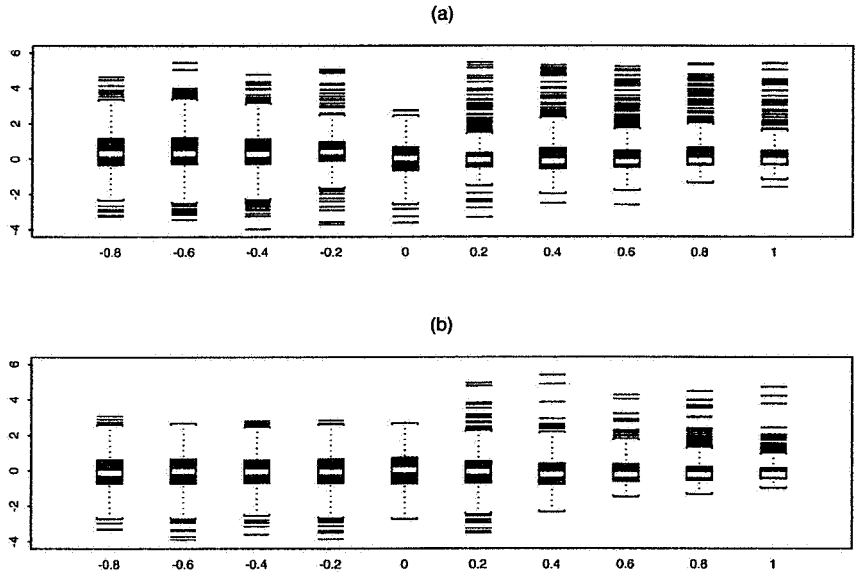

Fig. 3. Boxplots of the values of the power divergence statistic for testing the fit of the proportional odds model. Each boxplot corresponds to a different value of the parameter λ which varies from −0.8 to 1 with step equal to 0.2. The data have been generated according to model (2.3) with $\boldsymbol{\beta} = (1, 1.50, -2, 1)'$ and $\boldsymbol{z}_t = (1, x_t, \cos(\pi t/12))$. Here x_t stands for an autoregressive process of order 1 with coefficient 0.2 and the number of simulations is 500. (a) $T = 50$. (b) $T = 300$.

gross departures from the asserted asymptotic normality in most of the cases. Notice that for $\lambda = 0.6, 0.8$, and 1, the plots indicate moderate departure from the normality pointing out some large values of the test statistic. This is in agreement with the discussion immediately proceeding Theorem 3.1.

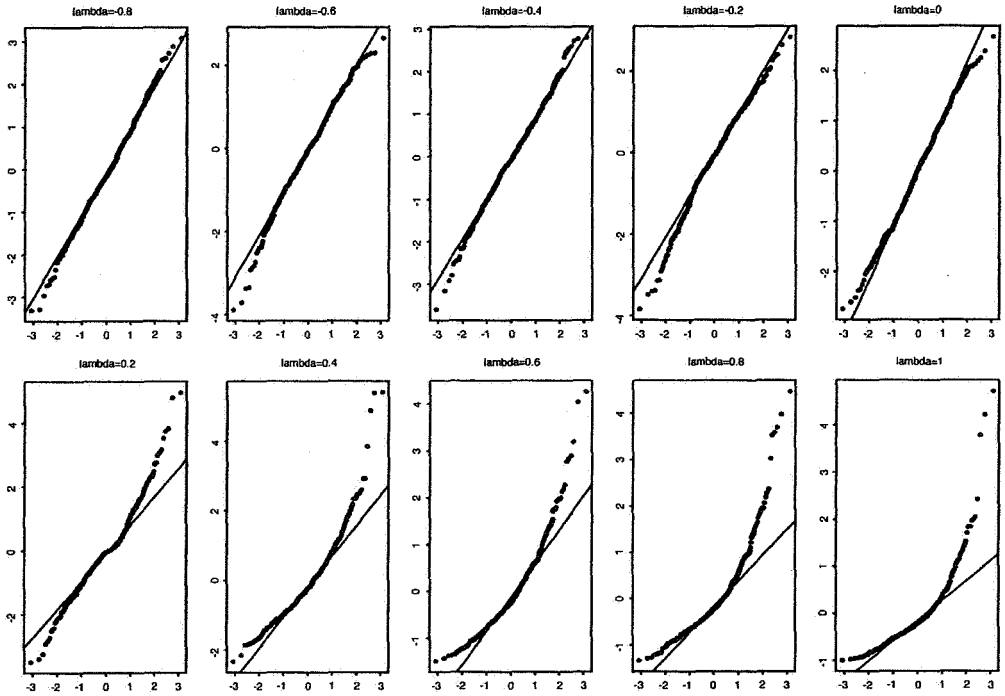

Fig. 4. QQ plots of the values of the power divergence statistic for testing the fit of the proportional odds model. Each plot corresponds to a different value of the parameter λ which varies from -0.8 to 1 with step equal to 0.2. The data have been generated according to model (2.3) with $\boldsymbol{\beta} = (1, 1.50, -2, 1)'$ and $\boldsymbol{z}_t = (1, x_t, \cos(\pi t/12))$. Here x_t stands for an autoregressive process of order 1 with coefficient 0.2. The length of each time series is 300 while the number of simulations is 500.

We illustrate similar results for the proportional odds model. Here, we simulate a time series with three categories according to (2.3) and use the same covariates as before but now $\boldsymbol{\beta} = (1, 1.5, -2, 1)'$. Table 2 reports our findings for different sample sizes. The approximation is satisfactory for $T = 300$, but we notice that the nominal significance levels $\alpha = 0.10$ and 0.05 are underestimated when λ takes on values greater than 0.6. In contrast, for $T = 50$, the achieved significance levels deviate from the asserted nominal levels in most of the cases. For instance when $\lambda = -0.6$, we obtain achieved significance levels 0.210, 0.164 and 0.092 respectively which shows that the approximation of the test statistic may not be reliable for small samples. Furthermore, Fig. 3(a) displays boxplots of the values of the power divergence statistic for different λ. Observe that for negative values of λ, the resulting distribution is symmetric with heavy tails while as λ shifts to the positive axis, the resulting distribution takes a skewed shape as opposed to Fig. 3(b) which exhibits a symmetric limiting distribution centered around 0, at least for $\lambda \leq 0.2$. However, notice that values of $\lambda \geq 0.4$, result to large positive numbers (see also Fig. 4, for $\lambda = 0.6$, 0.8 and 1).

The power of the test statistic was investigated by performing two experiments. We initially generate data according to the following model

$$p_{sj} = \frac{(\boldsymbol{\beta}_j' \boldsymbol{z}_{s-1})^2}{1 + (\boldsymbol{\beta}_1' \boldsymbol{z}_{s-1})^2 + (\boldsymbol{\beta}_2' \boldsymbol{z}_{s-1})^2}, \tag{4.1}$$

Table 3. Empirical power of the power divergence statistic for testing the fit of the multinomial logits model with different values of λ and for different sample sizes. The data have been generated according to model (4.1) with $\boldsymbol{\beta} = (-0.25, 0.50, 1, 0.50, -0.25, -1)'$ and $\boldsymbol{z}_t = (1, x_t, \cos(\pi t/12))$. Here x_t stands for an autoregressive process of order 1 and coefficient equal to 0.2. The number of simulations is 500.

λ	$T = 50$			$T = 300$		
	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
−0.8	0.250	0.202	0.112	0.458	0.346	0.210
−0.6	0.192	0.162	0.096	0.418	0.290	0.166
−0.4	0.312	0.230	0.138	0.486	0.356	0.202
−0.2	0.258	0.234	0.140	0.486	0.396	0.222
0	0.090	0.040	0.020	0.080	0.040	0.010
0.2	0.212	0.120	0.084	0.548	0.452	0.270
0.4	0.242	0.200	0.120	0.554	0.424	0.278
0.6	0.238	0.194	0.112	0.558	0.446	0.262
0.8	0.180	0.162	0.090	0.494	0.410	0.248
1	0.194	0.154	0.094	0.574	0.488	0.324

Table 4. Empirical power of the chi–square goodness of fit test based on Fokianos and Kedem ((1998), Proposition 4.1) for testing the fit of the multinomial logits model with different partitions and different sample sizes. The data have been generated according to model (4.1) with $\boldsymbol{\beta} = (-0.25, 0.50, 1, 0.50, -0.25, -1)'$ and $\boldsymbol{z}_t = (1, x_t, \cos(\pi t/12))$. Here x_t stands for an autoregressive process of order 1 and coefficient equal to 0.2.The number of simulations is 500.

Partition	$T = 50$			$T = 300$		
	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
$\mathcal{C}$	0.092	0.064	0.032	0.584	0.488	0.296
$\mathcal{D}$	0.120	0.084	0.044	0.844	0.776	0.532

for $j = 1, 2$. In other words, we generate a time series with three categories. The parameters $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2$ and the covariate process are specified as before. That is, we let $\boldsymbol{\beta}_1 = (-0.25, 0.50, 1)'$, $\boldsymbol{\beta}_2 = (0.50, -0.25, -1)$ and $\boldsymbol{z}_t = (1, x_t, \cos(\pi t/12))'$, for $t = 1, \ldots, T$. We fit the multinomial logits model to these data with the same covariates entering the regression equation, that is we apply model (2.2). Table 3 summarizes our findings regarding the power of the test statistics for different sample sizes. Notice that for $T = 50$, the power of the test statistic is low and it increases as T equals to 300. Observe that for $T = 300$, the power of the test increases as λ grows in the positive direction for this particular model. An essential observation is that when $\lambda = 0$ we obtain the significance levels which confirms the fact that the deviance does not offer an adequate way to measure departures from the hypothesized model.

To compare the power of the proposed test statistic with the power of the goodness of fit procedure proposed by Fokianos and Kedem ((1998) Proposition 4.1), we consider the following two partitions of the covariate space which in this case is 3-dimensional.

Table 5. Empirical power of the power divergence statistic for testing the fit of the proportional odds model with different values of λ and for different sample sizes. The data have been generated according to model (2.3) with $\boldsymbol{\beta} = (1, 1.50, -2, 1)'$, $\boldsymbol{z}_t = (1, x_t, \cos(\pi t/12))$ and F the Cauchy distribution. Here x_t stands for an autoregressive process of order 1 with coefficient equal to 0.2. The number of simulations is 500.

λ	$T = 50$			$T = 300$		
	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
−0.8	0.258	0.242	0.210	0.498	0.378	0.218
−0.6	0.244	0.124	0.022	0.424	0.338	0.188
−0.4	0.208	0.174	0.112	0.360	0.264	0.122
−0.2	0.192	0.134	0.064	0.292	0.190	0.078
0	0.154	0.060	0.018	0.082	0.040	0.010
0.2	0.122	0.112	0.050	0.744	0.674	0.526
0.4	0.228	0.198	0.138	0.566	0.514	0.374
0.6	0.256	0.222	0.180	0.464	0.402	0.296
0.8	0.206	0.168	0.110	0.350	0.282	0.200
1	0.232	0.232	0.196	0.308	0.248	0.180

First, we let $\mathcal{C} = (\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4)$ to be given by

$$\begin{aligned}
\mathcal{C}_1 &= \{\boldsymbol{z} : z_1 = 1, z_2 \leq -1.5, -1 \leq z_3 \leq -0.5\} \\
\mathcal{C}_2 &= \{\boldsymbol{z} : z_1 = 1, -1.5 < z_2 \leq 0.5, -0.5 < z_3 \leq 0\} \\
\mathcal{C}_3 &= \{\boldsymbol{z} : z_1 = 1, 0.5 < z_2 \leq 2, 0 < z_3 \leq 0.5\} \\
\mathcal{C}_4 &= \{\boldsymbol{z} : z_1 = 1, 2 < z_2, 0.5 < z_3 \leq 1\}
\end{aligned}$$

and then we let $\mathcal{D} = (\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4)$ to be given by

$$\begin{aligned}
\mathcal{D}_1 &= \{\boldsymbol{z} : z_1 = 1, z_2 \leq -1, -1 \leq z_3 \leq -0.2\} \\
\mathcal{D}_2 &= \{\boldsymbol{z} : z_1 = 1, -1 < z_2 \leq 1, -0.2 < z_3 \leq 0.4\} \\
\mathcal{D}_3 &= \{\boldsymbol{z} : z_1 = 1, 1 < z_2 \leq 2.5, 0.4 < z_3 \leq 0.8\} \\
\mathcal{D}_4 &= \{\boldsymbol{z} : z_1 = 1, 2.5 < z_2, 0.8 < z_3 \leq 1\}.
\end{aligned}$$

Notice that the asymptotic reference distribution for both partitions $\mathcal{C}$ and $\mathcal{D}$ is the chi-square with 8 degrees of freedom. The results reported in Table 4 show that the power divergence family of goodness of fit tests performs better for small sample sizes for this particular model. In contrast the chi-square test perform better when $T = 300$ except in a few cases. For instance when $T = 300$, $\alpha = 0.01$, $\lambda = 1$ and we choose partition $\mathcal{C}$ we notice that the power divergence family attains higher power. However Table 4 illustrates that the power the chi-square test depends upon the partition of the covariate space which occasionally may be large.

Our investigation continues by simulating data according to model (2.3) with F being the cumulative distribution function of a Cauchy random variable. We employ the same covariates, that is $\boldsymbol{z}_t = (1, x_t, \cos(\pi t/12))'$ and we let $\boldsymbol{\beta} = (1, 1.50, -2, 1)'$. We first fit the proportional odds model and subsequently the multinomial logits model. Tables 5 and 7 report our results. Notice that when T is small, the power of the test statistic for fitting the wrong cumulative odds model is higher in most of the cases than the

Table 6. Empirical power of the chi-square goodness of fit test based on Fokianos and Kedem ((1998), Proposition 4.1) for testing the fit of the proportional odds model with different partitions and for different sample sizes. The data have been generated according to model (2.3) with $\boldsymbol{\beta} = (1, 1.50, -2, 1)'$, $\boldsymbol{z}_t = (1, x_t, \cos(\pi t/12))$ and F the Cauchy distribution. Here x_t stands for an autoregressive process of order 1 with coefficient equal to 0.2. The number of simulations is 500.

Partition	$T = 50$			$T = 300$		
	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
$\mathcal{C}$	0.044	0.030	0.024	0.176	0.098	0.049
$\mathcal{D}$	0.030	0.012	0.008	0.064	0.056	0.022

Table 7. Empirical power of the power divergence statistic for testing the fit of the multinomial logits model with different values of λ and different sample sizes. The data have been generated according to model (2.3) with $\boldsymbol{\beta} = (1, 1.50, -2, 1)'$, $\boldsymbol{z}_t = (1, x_t, \cos(\pi t/12))$ and F the Cauchy distribution. Here x_t stands for an autoregressive process of order 1 with coefficient equal to 0.2. The number of simulations is 500.

λ	$T = 50$			$T = 300$		
	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
−0.8	0.190	0.142	0.092	0.772	0.706	0.550
−0.6	0.232	0.180	0.110	0.782	0.716	0.556
−0.4	0.284	0.212	0.134	0.736	0.680	0.548
−0.2	0.208	0.156	0.080	0.688	0.636	0.484
0	0.112	0.040	0.014	0.096	0.054	0.008
0.2	0.100	0.094	0.066	0.632	0.542	0.418
0.4	0.112	0.102	0.044	0.568	0.500	0.384
0.6	0.092	0.072	0.038	0.534	0.474	0.390
0.8	0.088	0.060	0.034	0.450	0.394	0.324
1	0.094	0.098	0.076	0.414	0.368	0.294

Table 8. Empirical power of the chi-square goodness of fit test based on Fokianos and Kedem ((1998), Proposition 4.1) for testing the fit of the multinomial logits model with different partitions and for different sample sizes. The data have been generated according to model (2.3) with $\boldsymbol{\beta} = (1, 1.50, -2, 1)'$, $\boldsymbol{z}_t = (1, x_t, \cos(\pi t/12))$ and F the Cauchy distribution. Here x_t stands for an autoregressive process of order 1 with coefficient equal to 0.2. The number of simulations is 500.

Partition	$T = 50$			$T = 300$		
	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
$\mathcal{C}$	0.076	0.064	0.042	0.112	0.102	0.040
$\mathcal{D}$	0.064	0.056	0.032	0.078	0.050	0.022

obtained power when we fit a multinomial logits model to those data. However, when T is large the multinomial logits model is rejected more frequently than the proportional odds model. In other words, when there is more data available, the power divergence

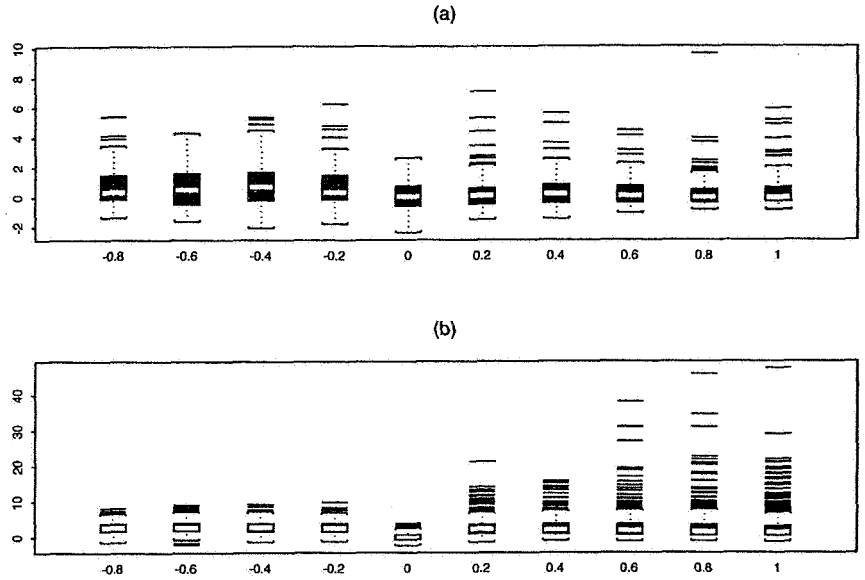

Fig. 5. Boxplots of the values of the power divergence statistic for testing the fit of multinomial logits model. Each boxplot corresponds to a different value of the parameter λ which varies from -0.8 to 1 with step equal to 0.2. The data have been generated according to model (2.3) with $\boldsymbol{\beta} = (1, 1.50, -2, 1)'$, $\boldsymbol{z}_t = (1, x_t, \cos(\pi t/12))$ and F the Cauchy distribution. Here x_t stands for an autoregressive process of order 1 with coefficient equal to 0.2 and the number of simulations is 500. (a) $T = 50$. (b) $T = 300$.

statistic discriminates among the different classes of models. The results from Tables 3, 5 and 7 indicate that there exists an appropriate value of the parameter λ that should optimize the power of the test. For instance, Table 7 shows that if $T = 300$ and $\lambda = -.6$ then we obtain the maximum power. The optimal value of the parameter λ will depend on the sample size and on the model at hand. Tables 6 and 8 report the power of the chi-square goodness of fit for different partitions. For both of these examples the power divergence family clearly performs better. Notice that different partitions would have altered the results.

Finally, Fig. 5 displays boxplots of the values of the test statistic for different sample sizes that correspond to Table 7. Apparently under the alternative hypothesis, positive values of λ indicate skewed distributions with most of the data lying in the positive axis. However, we notice that negative values of the parameter point to a symmetric distribution with mean shifted to the positive axis.

5. Concluding remarks

The family of power divergence test statistics can serve—as we already demonstrated—as an additional tool for testing the goodness of fit of a categorical regression model for time series. We saw by simulation that the particular choice of λ is of practical importance both in achieving the nominal significance level and in obtaining good power. Consequently, choice of the parameter λ is an issue that needs further investigation. Our limited simulation examples indicate that a reasonable range is between -1 and 1. Further insight is also needed for the theoretical properties of the power of the test

statistic. This work puts the power divergence family of test statistics in the framework of categorical regression models and examines some specific examples.

Acknowledgements

We would like to thank the reviewers for their useful and constructive remarks.

Appendix

PROOF OF THEOREM 3.1. The proof of theorem is based on the proof of Osius and Rojek (1992) for independent and not identically distributed data. However, we make use of the martingale limit theory for the proof of asymptotics and certain modifications apply.

Notice that the proof of Theorem 2.1 (see Fokianos and Kedem (1998)) leads to

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = o_p(T^{-1/2}). \tag{A.1}$$

Now, we turn to the numerator of the test statistic $I_T(\hat{\boldsymbol{\beta}})$. Then, a first order Taylor expansion yields

$$I_T(\hat{\boldsymbol{\beta}}) = I_T(\boldsymbol{\beta}) + [\nabla I_T(\boldsymbol{\beta})]'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + o_p(1). \tag{A.2}$$

However

$$\begin{aligned}\nabla I_T(\boldsymbol{\beta}) = &-\frac{2}{\lambda(\lambda+1)}\sum_{s=1}^{T}\sum_{j=1}^{m}\left[\left(\frac{1}{p_{sj}(\boldsymbol{\beta})}\right)^{\lambda} - 1\right]\frac{\partial p_{sj}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} \\ &-\frac{2}{(\lambda+1)}\sum_{s=1}^{t}\sum_{j=1}^{m}(y_{sj} - p_{sj}(\boldsymbol{\beta}))\frac{1}{p_{sj}^{\lambda+1}(\boldsymbol{\beta})}\frac{\partial p_{sj}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}.\end{aligned} \tag{A.3}$$

Now put

$$c_T(\boldsymbol{\beta}) = \sum_{s=1}^{T}\sum_{j=1}^{m}\frac{\partial p_{sj}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}\left[\left(\frac{1}{p_{sj}(\boldsymbol{\beta})}\right)^{\lambda} - 1\right]. \tag{A.4}$$

Hence, if we show that

$$-\frac{2}{(\lambda+1)}\sum_{s=1}^{T}\sum_{j=1}^{m}(y_{sj} - p_{sj}(\boldsymbol{\beta}))\frac{1}{{p_{sj}}^{\lambda+1}(\boldsymbol{\beta})}\frac{\partial p_{sj}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} = o_p(T^{1/2}) \tag{A.5}$$

then we conclude that an equivalent representation of equation (A.2) is given by

$$I_T(\hat{\boldsymbol{\beta}}) = I_T(\boldsymbol{\beta}) - \frac{2}{\lambda(\lambda+1)}c_T'(\boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + o_p(1). \tag{A.6}$$

We show that equation (A.5) holds. Consider the following quantity

$$\frac{1}{\sqrt{T}}\sum_{s=1}^{T}\sum_{j=1}^{m}(y_{sj} - p_{sj}(\boldsymbol{\beta}))\frac{1}{{p_{sj}}^{\lambda+1}(\boldsymbol{\beta})}\frac{\partial p_{sj}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}.$$

Its mean is obviously zero while the variance-covariance matrix can be explicitly calculated and it is given by

$$\frac{1}{T}\sum_{s=1}^{T}\left\{\sum_{j=1}^{m}\frac{(1-p_{sj}(\boldsymbol{\beta}))}{p_{sj}^{2\lambda+1}(\boldsymbol{\beta})}\frac{\partial p_{sj}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}\frac{\partial p_{sj}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right.$$
$$\left.-\sum_{j\neq k}^{m}p_{sj}(\boldsymbol{\beta})p_{sk}(\boldsymbol{\beta})\frac{1}{p_{sj}(\boldsymbol{\beta})}\frac{1}{p_{sk}(\boldsymbol{\beta})}\frac{\partial p_{sj}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}\frac{\partial p_{sk}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right\}$$

which converges in probability to some limit, as $T \to \infty$, in the presence of Assumption (A). Thus, equation (A.5) follows. The result (A.6) coupled with Theorem 2.1 yields to following useful representation of the numerator of the test statistic

$$I_T(\hat{\boldsymbol{\beta}}) = I_T(\boldsymbol{\beta}) - \frac{2}{\lambda(\lambda+1)}c_T'(\boldsymbol{\beta})\boldsymbol{G}_T^{-1}(\boldsymbol{\beta})S_T(\boldsymbol{\beta}) + o_p(1). \tag{A.7}$$

We now show that the term on the right hand side of Equation (A.7) is a zero-mean square integrable martingale that satisfies the necessary conditions for an application of the central limit theorem for martingales. Rewrite the right hand side of (A.7) as

$$\Xi_T(\boldsymbol{\beta}) = \frac{2}{\lambda(\lambda+1)}\left\{\sum_{s=1}^{T}(v_s'(\boldsymbol{\beta}) - c_T'\boldsymbol{G}_T^{-1}(\boldsymbol{\beta})\boldsymbol{W}_s'(\boldsymbol{\beta}))(\tilde{\boldsymbol{y}}_s - \tilde{\boldsymbol{p}}_s(\boldsymbol{\beta}))\right\}, \tag{A.8}$$

with $v_s(\boldsymbol{\beta})$ an $m \times 1$ vector with components

$$v_{sj}(\boldsymbol{\beta}) = \frac{1}{p_{sj}^{\lambda}(\boldsymbol{\beta})} - 1, \qquad j = 1, \ldots, m \tag{A.9}$$

and $\boldsymbol{W}_s(\boldsymbol{\beta})$ the $m \times p$ matrix with rows given by (2.6). Put

$$\alpha_s(\boldsymbol{\beta}) = (v_s'(\boldsymbol{\beta}) - c_T'\boldsymbol{G}_T^{-1}(\boldsymbol{\beta})\boldsymbol{W}_s'(\boldsymbol{\beta}))\,(\tilde{\boldsymbol{y}}_s - \tilde{\boldsymbol{p}}_s(\boldsymbol{\beta}))$$

for the increments of (A.8). Notice that $\{\Xi_s(\boldsymbol{\beta}),\ s = 1, 2, \ldots, T\}$ is a zero mean square integrable martingale with respect to the increasing family of σ-fields, $\mathcal{F}_s$. Indeed, Equation (A.8) yields that $\mathrm{E}[\Xi_t(\boldsymbol{\beta}) \mid \mathcal{F}_{t-1}] = \Xi_{t-1}(\boldsymbol{\beta})$ while $\mathrm{E}[\Xi_t(\boldsymbol{\beta})] = 0$. In the presence of Assumption (A) we have almost surely that $|\operatorname{Var}[\alpha_t(\boldsymbol{\beta})] \mid \mathcal{F}_{t-1}|| \leq M_1$, with M_1 a constant. Put

$$\xi_T(\boldsymbol{\beta}) = \frac{4}{\lambda^2(\lambda+1)^2}\sum_{s=1}^{T}\operatorname{Var}[\alpha_s(\boldsymbol{\beta}) \mid \mathcal{F}_{s-1}]$$
$$= \frac{4}{\lambda^2(\lambda+1)^2}\left\{\sum_{s=1}^{t}v_s'(\boldsymbol{\beta})\boldsymbol{\Sigma}_s(\boldsymbol{\beta})v_s(\boldsymbol{\beta}) - c_T'(\boldsymbol{\beta})\boldsymbol{G}_T^{-1}(\boldsymbol{\beta})c_T(\boldsymbol{\beta})\right\},$$

that is the cumulative variance of the martingale $\{\Xi_s(\boldsymbol{\beta}), s = 1, \ldots, T\}$, (see 3.4). Then, Assumption (A) guarantees the existence of an almost surely positive limit, say $\xi(\boldsymbol{\beta})$, such that

$$\frac{\xi_T(\boldsymbol{\beta})}{T} \to \xi(\boldsymbol{\beta}) \tag{A.10}$$

in probability, as $T \to \infty$. Here,

$$\xi(\boldsymbol{\beta}) = \int_{R^{p\times q}} [v'(\boldsymbol{\beta})\boldsymbol{\Sigma}(\boldsymbol{\beta})v(\boldsymbol{\beta})]\,\mu(dZ) - c'(\boldsymbol{\beta})\boldsymbol{G}^{-1}(\boldsymbol{\beta})c(\boldsymbol{\beta})$$

where $v(\boldsymbol{\beta})$ is an m-dimensional vector with components

$$v_j(\boldsymbol{\beta}) = \frac{1}{h_j(Z'\boldsymbol{\beta})^{\lambda}} - 1, \qquad j = 1, \ldots, m,$$

while the p-dimensional vector $c(\boldsymbol{\beta})$ is given by

$$c(\boldsymbol{\beta}) = \int_{R^{p\times q}} \left\{ \sum_{j=1}^{m} \frac{\partial h_j(Z'\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} \left[\left(\frac{1}{h_j(Z\boldsymbol{\beta})} \right)^{\lambda} - 1 \right] \right\} \mu(dZ), \qquad k = 1, \ldots, p,$$

and $\boldsymbol{G}(\boldsymbol{\beta})$ denotes the limiting information matrix. Furthermore if $C_s(\epsilon)$ denotes the indicator of the set $\{|\lambda'\alpha_s(\boldsymbol{\beta})|^2 \geq (\lambda'\xi_s(\boldsymbol{\beta})\lambda)^{1/2}\epsilon\}$, with λ arbitrary p-dimensional vector, we have

$$\begin{aligned} \frac{1}{\xi_t(\boldsymbol{\beta})} \sum_{s=1}^{t} E[|\lambda'\alpha_s(\boldsymbol{\beta})|^2 C_s(\epsilon) \mid \mathcal{F}_{s-1}] &\leq \frac{1}{(\xi_t(\boldsymbol{\beta}))^{3/2}\epsilon} \sum_{s=1}^{t} E[|\lambda'\alpha(\boldsymbol{\beta})|^3 \mid \mathcal{F}_{s-1}] \\ &\leq \frac{TM_2}{(\xi_t(\boldsymbol{\beta}))^{3/2}\epsilon}, \end{aligned}$$

with M_2 some constant. The last expression tends to 0, as $T \to \infty$ which implies that the Lindeberg's condition is satisfied and thus a central limit theorem for martingales is applicable (Hall and Heyde (1980), Corollary 3.1). In other words, we proved that

$$\frac{\Xi_T(\boldsymbol{\beta})}{\sqrt{\xi_T(\boldsymbol{\beta})}} \xrightarrow{D} \mathcal{N} \tag{A.11}$$

as $T \to \infty$. Here $\mathcal{N}$ stands for a standard normal random variable. Thus, in order to complete the proof of the theorem we need to show that $\xi_t(\hat{\boldsymbol{\beta}})$ is a consistent estimator of $\xi_t(\boldsymbol{\beta})$. We have though the following

$$\frac{1}{T} \left\{ \sum_{s=1}^{T} v_s'(\hat{\boldsymbol{\beta}})\boldsymbol{\Sigma}_s(\hat{\boldsymbol{\beta}})v_s(\hat{\boldsymbol{\beta}}) - \sum_{s=1}^{T} v_s'(\boldsymbol{\beta})\boldsymbol{\Sigma}_s(\boldsymbol{\beta})v_s(\boldsymbol{\beta}) \right\} = O_p(1)$$

and

$$\frac{1}{T}(c_T'(\hat{\boldsymbol{\beta}})\boldsymbol{G}_T{}^{-1}(\hat{\boldsymbol{\beta}})c_T(\hat{\boldsymbol{\beta}}) - c_T'(\boldsymbol{\beta})\boldsymbol{G}_T{}^{-1}(\boldsymbol{\beta})c_T(\boldsymbol{\beta})) = O_p(1)$$

due to Theorem 2.1 and Assumption (A). Therefore the theorem is proved.

REFERENCES

Agresti, A. (1990). *Categorical Data Analysis*, Wiley, New York.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression models for counting process: A large sample approach, *Ann. Statist.*, **10**, 1100–1120.

Arjas, E. and Haara, P. (1987). A logistic regression model for hazard: Asymptotic results, *Scand. J. Statist.*, **14**, 1–18.

Bonney, E. G. (1987). Logistic regression for dependent binary observations, *Biometrics*, **43**, 951–973.

Brillinger, D. R. (1996). An analysis of ordinal–valued time series, *Athens Conference on Applied Probability and Time Series Analysis, Vol. II: Time Series Analysis in Memory of E. J. Hannan*, Lecture Notes in Statist., No. 115, 73–87, Springer, New York.

Cox, D. R. (1975). Partial likelihood, *Biometrika*, **62**, 69–76.

Cressie, N. A. C. and Read, T. R. C. (1984). Multinomial goodness–of–fit tests, *J. Roy. Statist. Soc. Ser. B*, **46**, 440–464.

Dale, J. R. (1986). Asymptotic normality of goodness–of–fit statistics for sparse product multinomials, *J. Roy. Statist. Soc. Ser. B*, **48**, 48–59.

Fahrmeir, L. and Kaufmann, H. (1987). Regression models for nonstationary categorical time series, *J. Time Ser. Anal.*, **8**, 147–160.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modeling Based on Generalized Linear Models*, Springer, New York.

Fokianos, K. and Kedem, B. (1998). Prediction and classification of non-stationary categorical time series, *J. Multivariate Anal.*, **67**, 277–296.

Fokianos, K. and Kedem, B. (1999). A stochastic approximation algorithm for the adaptive control of time series following generalized linear models, *J. Time Ser. Anal.*, **20**, 289–308.

Gleser, L. J. and Moore, D. S. (1985). The effect of positive dependence on chi–square tests for categorical data, *J. Roy. Statist. Soc. Ser. B*, **47**, 459–465.

Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and Its Applications*, Academic Press, New York.

Holst, L. (1972). Asymptotic normality and efficiency for certain goodness of fit tests, *Biometrika*, **59**, 137–145.

Kaufmann, H. (1987). Regression models for nonstationary categorical time series: Asymptotic estimation theory, *Ann. Statist.*, **15**, 79–98.

Koehler, K. J. (1986). Goodness–of–fit tests for log–linear models in sparse contingency tables, *J. Amer. Statist. Assoc.*, **81**, 483–493.

Korn, E. L. and Whittemore, A. S. (1979). Methods for analyzing panel studies of acute health effects of air pollution, *Biometrics*, **35**, 795–802.

Li, W. K. (1991). Testing model adequacy for some markov regression models for time series, *Biometrika*, **78**, 83–89.

Liang, K.-Y. and Zeger, S. L. (1989). A class of logistic regression models for multivariate binary time series, *J. Amer. Statist. Assoc.*, **84**, 447–451.

McCullagh, P. (1980). Regression models for ordinal data (with discussion), *J. Roy. Statist. Soc. Ser. B*, **42**, 109–142.

McCullagh, P. (1986). The conditional distribution of goodness–of–fit statistics for discrete data, *J. Amer. Statist. Assoc.*, **81**, 104–107.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed., Chapman and Hall, London.

Morris, C. (1975). Central limit theorem for multinomial sums, *Ann. Statist.*, **3**, 165–188.

Muenz, L. R. and Rubinstein, L. (1985). Markov models for covariate dependence of binary sequences, *Biometrics*, **41**, 91–101.

Osius, G. (1985). Goodness–of–fit tests for binary data with (possible) small expectations but large degrees of freedom, Supplement to *Statist. Decisions*, **2**, 213–224.

Osius, G. and Rojek, D. (1992). Normal goodness–of–fit tests for multinomial models with large degrees of freedom, *J. Amer. Statist. Assoc.*, **87**, 1145–1152.

Read, T. R. C. and Cressie, N. A. C. (1988). *Goodness-of-fit Statistics for Discrete Multivariate Data*, Springer, New York.

Schoenfeld, D. (1980). Chi-square goodness-of-fit test for the proportional hazards regression model, *Biometrika*, **67**, 145–153.

Slud, E. and Kedem, B. (1994). Partial likelihood analysis of logistic regression and autoregression, *Statist. Sinica*, **4**, 89–106.

Stern, R. D. and Coe, R. (1984). A model fitting analysis of daily rainfall data, *J. Roy. Statist. Soc. Ser. A*, **147**, 1–34.

Weiss, L. (1976). The normal approximation to the multinomial with increasing number of classes, *Naval Res. Logist. Quarterly*, **23**, 139–149.

Wong, W. H. (1986). Theory of partial likelihood, *Ann. Statist.*, **14**, 88–123.