Analyzing Associations in Multivariate Binary Time Series

Roland Fried[1], Silvia Kuhls[2], and Isabel Molina[1]

[1]Depto. de Estadística, Universidad Carlos III de Madrid, Spain

[2] Fachbereich Statistik, Universität Dortmund, Germany

Summary. We analyze multivariate binary time series using a mixed parameterization in terms of the conditional expectations given the past and the pairwise canonical interactions among contemporaneous variables. This allows consistent inference on the influence of past variables even if the contemporaneous associations are misspecified. Particularly, we can detect and test Granger non-causalities since they correspond to zero parameter values.

1 Introduction

In multivariate time series analysis we want to measure the associations among the variables and identify lead-lag relationships. This is useful for predicting future outcomes and for identifying causal mechanisms using the concept of Granger causality (Granger, 1969). This concept is based on the common sense that a cause must precede its effect in time.

Observation-driven transitional models are natural candidates for modelling dynamic interactions within multivariate binary time series since they condition explicitly on the history of the process. We use a likelihood approach assuming the conditional distribution of the contemporaneous variables given the past to lie within the quadratic exponential family (Zhao and Prentice, 1990), adapting the mixed parameterization in terms of conditional expectations and the canonical pairwise interactions (Fitzmaurice and Laird, 1993) to the dynamic situation. This allows consistent estimation of the influences of the past even if the contemporaneous association structure is misspecified. Additionally, it allows straightforward detection of conditional independencies implying Granger non-causalities among the process variables (Dahlhaus and Eichler, 2003).

Section 2 introduces the mixed parameterization model for multivariate binary time series. Section 3 proposes model fitting based on conditional likelihood and asymptotic tests. Section 4 presents some simulation results.

2 The model

Let $M_V(t) = (M_1(t), \ldots, M_d(t))'$, $t \in \mathbb{Z}$, be a d-variate process of binary variables indexed by $v \in V = \{1, \ldots, d\}$. We restrict ourselves to multiplicative models (Cox, 1972) for the contemporaneous association structure, which belong to the quadratic exponential family (Zhao and Prentice, 1990) and only consider pairwise interactions. We further assume the process to have a memory of p observations. Denoting the past of the process at time t by $\overline{m}_V(t-1) = (m_V(t-1), m_V(t-2), \ldots)$, the conditional multivariate model for the outcome $m_V(t) \in \{0,1\}^d$ reads

$$\begin{aligned} pr(m_V(t)|\overline{m}_V(t-1)) &= \Delta(t)^{-1} \exp\left\{m_V(t)'\psi_V(t) + w_V(t)'\lambda_V\right\}, \qquad (1) \\ \text{where } \Delta(t) &= \sum_{m_V(t)\in\{0,1\}^d} \exp\left\{m_V(t)'\psi_V(t) + w_V(t)'\lambda_V\right\}, \\ \text{and } w_V(t) &= (m_1(t)m_2(t), \ldots, m_2(t)m_3(t), \ldots, m_{d-1}(t)m_d(t))' \end{aligned}$$

is a $d(d-1)/2 \times 1$-vector of all different pairwise products from $m_V(t)$. By $W_V(t)$ as usual we denote the corresponding random vector. The time-variant main effects $\psi_V(t) = (\psi_1(t), \ldots, \psi_d(t))'$ include the influences of the past, while the elements of λ_V are log conditional odds ratios describing pairwise interactions, which are assumed to be time-invariant. The normalizing constant $\Delta(t)$ depends on the past $\overline{m}_V(t-1)$ and the model parameters. The parameters in this log-linear representation are not restricted.

While Liang and Zeger (1989) implicitly model the canonical parameters $\psi_V(t)$ and λ_V, we follow Zeger and Liang (1991) and model the conditional expectation $\pi_V(t) = E(M_V(t)|\overline{m}_V(t-1))$ given the past. In this we make use of the mixed parameterization of the quadratic exponential family in terms of the (marginal) expectations and the log conditional odds ratios λ_V (Fitzmaurice and Laird, 1993). Conditioning on the past we introduce past effects via the conditional mean $\pi_V(t) = (\pi_v(t))_{v \in V}$ as it is common practice in the Gaussian framework. We transform the canonical parameters $(\psi_V(t), \lambda_V)$ to $(\pi_V(t), \lambda_V)$ and use the logit link $h(\pi_v(t)) = \log\{\pi_v(t)/[1 - \pi_v(t)]\} = \mu_v(t)$ for the conditional mean,

$$\begin{aligned} \pi_v(t) = E\left[M_v(t)|\overline{m}_V(t-1)\right] &= h^{-1}(\mu_v(t)) = \frac{\exp(\mu_v(t))}{1+\exp(\mu_v(t))} \qquad (2) \\ \text{where } \mu_V(t) &= \phi_V + \sum_{h=1}^{p} \Phi_V(h) m_V(t-h)\,. \end{aligned}$$

The matrices $\Phi_V(h) = (\phi_{vw}(h))_{v,w \in V}$, $h = 1, \ldots, p$, describe the time-invariant effects of past observations, and ϕ_V determines the probabilities of ones which are not induced by others.

In the mixed parameterization model arising from (1) and (2), $\lambda_{vw} = 0$ means that the variables $M_v(t)$ and $M_w(t)$ are conditionally independent given the past $\overline{M}_V(t-1)$ and the other variables at time t, $M_{V\setminus\{v,w\}}(t)$. More interestingly, $\phi_{vw}(h) = 0$ implies that $M_v(t)$ is independent from $M_w(t-h)$ given the other past observations, $\overline{M}_V(t-1) \setminus \{M_w(t-h)\}$. We can further deduce that $M_v(t)$ is conditionally independent from the past $\overline{M}_w(t-1)$ of variable w given the past of the other variables if all $\phi_{vw}(h) = 0$, $h = 1, \ldots, p$. Variable w is Granger non-causal for v in this case. This model can be seen as a binary analogue of a Gaussian VAR model parameterized via the inverse covariance matrix instead of the covariance matrix as investigated in (Dahlhaus and Eichler, 2003).

3 Model fitting

We derive conditional likelihood estimates and asymptotic tests for the mixed parameterization model. The tests allow to detect zero parameters, and this in turn allows specification of (non-)causalities among the process variables.

3.1 Likelihood estimation

Let in the following β be a vector of mean parameters comprising all elements of $\phi_V, \Phi_V(1), \ldots, \Phi_V(p)$, while $\theta = (\beta', \lambda')'$ contains all parameters in the conditional means $\pi(t)$ and the contemporaneous associations λ. We drop the index V in this section for simplicity, always referring to the set of all variables. Denoting the follow-up time by T and conditioning on the first p observations, the conditional log-likelihood reads

$$\ell(\theta) = \sum_{t=p+1}^{T} \ell_t(\theta) = \sum_{t=p+1}^{T} \left[m(t)'\psi(t) + w(t)'\lambda - \log \Delta(t) \right].$$

Let $X(t)$ be the design matrix with past observations, $\eta(t) = E(W(t)|\overline{m}(t-1))$, $V_{11}(t) = cov(M(t)|\overline{m}(t-1))$, $V_{21}(t) = cov(W(t), M(t)|\overline{m}(t-1))$, and $V_{22}(t) = cov(W(t)|\overline{m}(t-1))$. Evaluating the partial derivatives, $(\hat{\beta}, \hat{\lambda})$ solves

$$\sum_{t=p+1}^{T} \begin{pmatrix} \frac{\partial \ell_t}{\partial \beta} \\ \frac{\partial \ell_t}{\partial \lambda} \end{pmatrix} = \sum_{t=p+1}^{T} \begin{pmatrix} X(t)' D(t) V_{11}^{-1}(t) \left[m(t) - \pi(t)\right] \\ \left\{ w(t) - \eta(t) - V_{21}(t) V_{11}^{-1}(t) \left[m(t) - \pi(t)\right] \right\} \end{pmatrix} = 0 \, .$$

Here, $D(t) = diag(var(M(t)|\overline{m}(t-1))$ is a diagonal matrix, which, like $\pi(t)$, depends only on β, while $\eta(t), V_{11}(t), V_{21}(t)$ and $V_{22}(t)$ depend on β and λ.

The conditional information matrix G_T turns out to be block-diagonal,

$$G_T = \sum_{t=p+1}^{T} \begin{pmatrix} X(t)' D(t) V_{11}^{-1}(t) D(t) X(t) & 0 \\ 0 & V_{22}(t) - V_{21}(t) V_{11}^{-1}(t) V_{21}(t)' \end{pmatrix},$$

implying that the parameters β and λ modelling the influences of the past and the contemporaneous associations are orthogonal: estimation of β is robust against misspecification of the contemporaneous associations, and the asymptotic variance of $\hat{\beta}$ remains the same whether λ is known or estimated (Cox and Reid, 1987, 1989).

The conditional ML estimates can be obtained using Fisher scoring. Denoting the estimates obtained in step j of the algorithm by the superscript (j), the recursions read

$$\begin{aligned} \hat{\beta}^{(j+1)} &= \hat{\beta}^{(j)} + \left\{ \sum X(t)' D(t)^{(j)} V_{11}^{-1}(t)^{(j)} D(t)^{(j)} X(t) \right\}^{-1} \\ &\quad \times \sum X(t)' D(t)^{(j)} V_{11}^{-1}(t)^{(j)} \left[m(t) - \pi(t)^{(j)} \right] \\ \hat{\lambda}^{(j+1)} &= \hat{\lambda}^{(j)} + \left\{ \sum \left[V_{22}(t)^{(j)} - V_{21}(t)^{(j)} V_{11}^{-1}(t)^{(j)} \left(V_{21}(t)^{(j)} \right)' \right] \right\}^{-1} \\ &\quad \times \sum_{t=1}^{T} \left\{ w(t) - \eta(t)^{(j)} - V_{21}(t)^{(j)} V_{11}^{-1}(t)^{(j)} \left[m(t) - \pi(t)^{(j)} \right] \right\} \, . \end{aligned}$$

Evaluation of these scoring equations is complicated by the fact that the conditional probability distribution is needed for all time points t to calculate moments of up to fourth order, but there is no closed form expressing the joint probabilities as function of $\pi(t)$ and λ. We follow Fitzmaurice and Laird (1993) and apply iterative proportional fitting (Deming and Stephan, 1940) of the cell probabilities given the current estimates within each Fisher scoring step. Given $\lambda^{(j)}$, we construct 2^d tables, $S_t(\lambda^{(j)})$ with these conditional log odds ratios. Using iterative proportional fitting with $S_t(\lambda^{(j)})$ as start table, we then fit the margins $\pi(t)^{(j)}$ to the table. For the resulting 2^d tables of cell probabilities we get $E[M(t)|\overline{m}(t-1)] = \pi(t)^{(j)}$ and conditional log-odds $\lambda^{(j)}$, and thus we can use them to update $\eta(t), V_{11}(t), V_{21}(t)$ and $V_{22}(t)$. We can save computation time by once evaluating all 2^{dp} possible such tables in each step j instead of calculating the tables for all $T-p$ time points individually.

3.2 Parameter tests

For derivation of the asymptotic distribution of the estimators we can adapt results of Kaufmann (1987) for categorical time series to our context: The probability that a unique conditional maximum likelihood estimate exists converges to one. Any sequence $\{(\hat{\beta}_T', \hat{\lambda}_T')'\}$ of MLEs is consistent and asymptotically normal,

$$\left(G_T^{1/2}\right)' \begin{pmatrix} \hat{\beta}_T \\ \hat{\lambda}_T \end{pmatrix} \xrightarrow{d} N(0, I),$$

where we use a square root like the Cholesky one for standardization.

Application of the Wald statistic

$$\hat{\theta}' C' [C G_T^{-1}(\hat{\theta}) C']^{-1} C \hat{\theta}$$

allows to test general linear hypotheses $C\theta = 0$ for arbitrary matrices C with appropriate dimensions; particularly, for a given order p we can test the significance of subsets of past variables using the estimates obtained for the saturated model.

As noted before, $\hat{\beta}$ is consistent regardless whether the contemporaneous dependence structure has been correctly specified or not since β and λ are orthogonal. However, the standardization by the conditional Fisher information matrix can give inconsistent estimates of the asymptotic variance of $\hat{\beta}$. A robust alternative is the sandwich estimate, which is not sensitive to the specification of the association structure (Boos, 1992).

4 Simulations

For illustration, we fit the mixed parameterization model to 3-variate time series, using the memory $p = 1$ and the logit link. We hence need to estimate a total of 15 parameters in $\beta = (\phi_1, \phi_{11}(1), \ldots, \phi_{13}(1), \ldots, \phi_3, \phi_{31}(1), \ldots, \phi_{33}(1))'$ and $\lambda = (\lambda_{12}, \lambda_{13}, \lambda_{23})' \in \mathbb{R}^3$. The design matrix for the conditional mean is

$$X(t) = \begin{pmatrix} 1 & m_1(t-1) & \ldots & m_d(t-1) \\ \vdots & \vdots & & \vdots \\ 1 & m_1(t-1) & \ldots & m_d(t-1) \end{pmatrix}.$$

4.1 White noise

First we check whether the Wald tests for the individual significance of the parameters preserve their level in finite samples of several sizes $T = 200, 500, 1000$. We generate 3-variate binary time series of independent observations by discretizing a Gaussian white noise process $\{Y_V(t)\}$ with covariance matrix being the identity, using 1.0 (2.0) as threshold, i.e. we set $m_v(t) = 1$ iff for the corresponding Gaussian observation $y_v(t) > 1.0$ (2.0). Thus, all the binary variables are independent, and we expect about 16% (2.5%) of the binary observations to be 1's for each variable.

Table 1 presents the percentage of cases a parameter describing a past influence was (incorrectly) found to be distinct from zero within 500 repetitions for each of $T = 200, 500, 1000$, averaging over all $\phi_{vw}(1)$, $v, w \in \{1, 2, 3\}$. The likelihood-based tests using the conditional information matrix for standardization seem to be slightly conservative, both in case of a large and a small probability of ones.

The tests based on the sandwich covariance matrix also preserve the significance level well if the probabilities of zeros and ones are about the same. If the probability of a one is small, however, i.e. if a one is a rare event, the tests based on the sandwich covariance matrix exceed their levels largely rejecting the null hypothesis in most of the cases. A large sample size seems necessary to distinguish zero and non-zero parameters using the sandwich estimate.

The left hand side of Table 2 shows the percentage of cases in which the order $p = 1$ was found to be significant against the null hypothesis $p = 0$. Again a Wald test based on either the conditional information or the sandwich covariance matrix with a significance level $\alpha = 10\%$ is used instead of a simpler score test. The test based on the conditional information seems to be conservative in samples of moderate size, but this conservatism diminishes with increasing sample size. In moderately large samples, the test based on the sandwich estimator seems only useful if zeros and ones are approximately balanced, see the huge percentage of cases $p = 1$

Table 1: Empirical significance levels of tests for the dependence parameters $\phi_{vw}(1)$ at $\alpha = 5\%$ for several sample sizes T, using the conditional information matrix (CI) or the sandwich estimate (SE) for standardization. Results for 16% (left) and for 2.5% probability of a 'one' (right).

T	CI	SE	CI	SE
200	3.2%	3.8%	4.0%	94.7%
500	3.7%	3.7%	3.6%	80.1%
1000	4.1%	4.1%	3.9%	63.1%

is incorrectly found to be necessary if there are 2.5% ones; even in case of a balanced occurrence of zeros and ones it can be somewhat liberal if the sample is small. This behavior can be explained by the often much larger variance of the sandwich estimator as compared to a correctly specified model-based variance estimator, see Kauermann and Carroll (2001).

4.2 VAR(1) process

To check the power of the tests for detecting relevant parameters we consider two settings with $p = 1$ and contemporaneously conditionally independent observations, i.e. $\lambda_{12} = \lambda_{13} = \lambda_{23} = 0$. The second setting corresponds to ones being rare events:

1) $\phi_v = -2$, $v = 1, 2, 3$; $\phi_{21}(1) = 1$, $\phi_{32}(1) = 0.5$, all other $\phi_{vw}(1) = 0$. This implies $P(M_2(t) = 1|M_1(t-1) = 0) = P(M_3(t) = 1|M_2(t-1) = 0) = 11.9\%$, but $P(M_2(t) = 1|M_1(t-1) = 1) = 26.9\%$ and $P(M_3(t) = 1|M_2(t-1) = 1) = 18.2\%$, i.e. an increase of 125% and 53%, respectively.

Table 2: Percentage of cases (out of 500) in which $p = 1$ was found to be significant vs. the null $p = 0$ at level $\alpha = 10\%$ if the true $p = 0$ or $p = 1$. Results for 16% (left) and for 2.5% probability of a one (right). Tests based on the conditional information (CI) and the sandwich estimate (SE).

	$p=0$		$p=1$		$p=0$		$p=1$	
T	CI	SE	CI	SE	CI	SE	CI	SE
200	5.4%	15.2%	33.4%	67.0%	2.0%	82.0%	4.8%	83.2%
500	7.0%	9.6%	70.6%	73.6%	5.8%	88.6%	10.0%	89.6%
1000	8.0%	8.2%	94.8%	95.0%	6.4%	95.8%	10.6%	91.4%

2) $\phi_v = -4$, $v = 1, 2, 3$; $\phi_{21}(1) = 1$, $\phi_{32}(1) = 0.5$, all other $\phi_{vw}(1) = 0$. This implies $P(M_2(t) = 1|M_1(t-1) = 0) = P(M_3(t) = 1|M_2(t-1) = 0) = 1.8\%$, but $P(M_2(t) = 1|M_1(t-1) = 1) = 4.75\%$ and $P(M_3(t) = 1|M_2(t-1) = 1) = 2.93\%$, i.e. an increase of 164% and 61%, respectively.

Table 3 presents the percentage of cases in 500 repetitions in which the parameters were found to be individually significant. Inference based on the conditional information matrix has considerable power in setting 1), but much smaller power in setting 2) where preceding ones in the causing variables cause a similar relative increase of risk, but on a much smaller level. Anyway, the power increases with increasing series length, and it is higher for a larger effect $\phi_{vw}(1)$. The empirical error rates of first type strengthen the previous statements on the significance level. Similarly, the necessity of a memory $p = 1$ can be detected with high certainty in setting 1), while in setting 2) longer series seem necessary given the small magnitudes of the

Table 3: Percentage of cases in which the parameters were significant at level $\alpha = 5\%$ for several sample sizes T, using the conditional information matrix (CI, left) or the sandwich estimate (SE, right). Results for setting 1) (top) and 2) (bottom).

T	$\phi_{21}(1)$	$\phi_{32}(1)$	others	$\phi_{21}(1)$	$\phi_{32}(1)$	others
200	45.6%	20.2%	4.1%	46.2%	22.0%	9.9%
500	80.6%	34.4%	3.9%	80.8%	33.8%	4.0%
1000	97.6%	52.2%	4.7%	97.6%	52.2%	4.7%
200	10.8%	7.2%	4.9%	95.4%	96.2%	96.6%
500	15.0%	8.0%	5.1%	84.0%	84.2%	90.1%
1000	17.0%	8.2%	3.4%	62.8%	70.4%	77.0%
5000	46.6%	13.4%	3.5%	48.2%	21.0%	15.3%

effects, see Table 2.

When using the sandwich estimate in setting 1), we get almost the same results as before if the length of the series is at least $T = 500$. For $T = 200$ we get almost the same detection rate, but at the expense of a considerably larger false discovery rate. In setting 2), the tests using the sandwich standardization are not able to distinguish zero and non-zero parameter values unless we have several thousand observations available.

5 Conclusion

We have proposed a likelihood approach to modelling multivariate binary time series that allows detection of Granger (non-)causalities among the process variables. Asymptotic tests for zero parameter values and hence non-causalities have been provided. For a simple and feasible approach, we have applied some restrictive, but common assumptions like the non-existence of higher order interactions and invariance of conditional odds ratios. The severeness of these assumptions is reduced by the orthogonality of the past and the contemporaneous associations. Likelihood estimation of the former is consistent even if the contemporaneous association structure is misspecified. Nevertheless, testing for zero parameter values depends on the validity of the underlying assumptions since the sandwich estimate seems unreliable in samples of moderate size if ones are rare events. This observation and the many other possible modelling assumptions underline the importance of careful model checking and comparison.

Acknowledgements

The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures") is gratefully acknowledged.

References

Boos, D.D. (1992). On generalized score tests. *Amer. Statist.*, **46**, 327–333.

Cox, D.R. (1972). The analysis of multivariate binary data. *Appl. Statist.*, **21**, 113–120.

Cox, D.R., and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *J. R. Statist. Soc.* B, **49**, 1–39.

Cox, D.R., and Reid, N. (1989). On the stability of maximum-likelihood estimators of orthogonal parameters. *Can. J. Statist.*, **17**, 229–233.

Dahlhaus, R., and Eichler, M. (2003). Causality and graphical models for time series. In: Green, P., Hjort, N., Richardson, S. (eds) *Highly Structured Stochastic Systems*. University Press, Oxford.

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

Fitzmaurice, G.M., and Laird, N.M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, **80**, 141–151.

Granger, C.W.J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, **37**, 424–438.

Kauermann, G., and Carroll, R.J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *J. Amer. Statist. Assoc.*, **96**, 1387–1396.

Kaufmann, H. (1987). Regression models for non-stationary categorical time series: asymptotic estimation theory. *Annals of Statistics*, **15**, 79–98.

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

Zeger, S.L., and Liang, K.-Y. (1991). Feedback models for discrete and continuous time series. *Statistica Sinica*, **1**, 51–64.

Zhao, L.P., and Prentice, R.L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika*, **77**, 642–648.