UiO : **Department of Mathematics**
University of Oslo

Focused Model Selection for Markov Chain Models

With an Application to Armed Conflict Data

Jens Kristoffer Haug

Master's Thesis, Spring 2019

This master's thesis is submitted under the master's programme *Computational Science*, with programme option *Applied Mathematics and Risk Analysis*, at the Department of Mathematics, University of Oslo. The scope of the thesis is 30 credits.

The front page depicts a section of the root system of the exceptional Lie group E_8, projected into the plane. Lie groups were invented by the Norwegian mathematician Sophus Lie (1842–1899) to express symmetries in differential equations and today they play a central role in various parts of mathematics.

Abstract

This thesis is devoted to the development of a focused information criterion for dynamic multinomial logit models. The achievements of the thesis are fourfold. First, a dynamic multinomial logit model is defined which admits the possibility of model misspecification. Then, approximate large sample distributions of maximum likelihood estimates of this model are deduced. The deduction is done both for correctly specified models and for misspecified models. On the basis of these approximate distributions, the Focused Information Criterion is constructed. The performance of the developed Focused Information Criterion is investigated through simulation experiments. It is shown that the developed information criterion indeed aims at selecting the models giving the most precise estimate of the focus parameter. As an application of the developed methodology, armed conflict data are analyzed. The focus parameter in this analysis is the probability of conflict escalation. The findings show that the level of democracy has no significant effect on conflict escalation probabilities.

Preface

The seeds of this Master's thesis were sown in late autumn 2017. During a lecture in Bayesian statistics my supervisor Nils Lid Hjort asked me if I would be interested in writing a thesis on Tolstoyan topics. Knowing that Hjort had analyzed word frequencies in late medieval literature, this suggestion did not surprise me. What Hjort had in mind, however, was the subject of statistical conflict analysis. Having studied philosophy and history in the past, the idea appealed to me. Hjort's suggestion resulted in an application for the UiO-PRIO student program, where I got accepted.

Since January 2018 I have been working on this thesis partly at the Department of Mathematics at UiO, partly at the Peace Research Institute Oslo. This has provided me with a stimulating interdisciplinary environment. It has been a pleasure to write this thesis amongst so many talented people.

I am grateful to my supervisor Nils Lid Hjort for encouragement, enthusiasm and excellent advice throughout the whole process. It has been very interesting to be associated with the FocuStat research group where so many great ideas have been conceived and discussed.

Also, thanks go to my co-supervisor Håvard Mokleiv Nygård at PRIO for constructive feedback and reliable guidance in the challenging field of quantitative conflict research. Being a part of the excellent team at PRIO has been a truly inspiring experience. I thank fellow master students at PRIO and especially the gang at the Learner's Loft. Without you, it would have been harder to write this Master's thesis.

I owe thanks to Mathis Mæhlum, Håvard Halland Fretheim and Amund Norli Løvik for good advice at crucial junctures in the process. A final 'thanks' goes to Henrik Haug Hytten.

Oslo, May 2019
Jens Kristoffer Haug

CHAPTER 1

Introduction

Statistical analysis of armed conflict is of interest to an audience much wider than just the scholars working on international relations. After all, questions of War and Peace are of the highest importance to everyone. Show me the person indifferent to the way Madame Fortuna is turning her wheel of violence!

No wonder then, that a global bestseller in recent years has been Steven Pinker's *The Better Angels of our Nature* (Pinker, 2011). In this impressive work, Pinker examines historical data from a variety of sources and concludes that the world has seen a steady decline in armed conflict. According to Pinker, the world has changed to the better: Wars have become less probable, the chances of violent death are reduced. Not the worst of messages to convey to the general public.

Among conflict researchers, the question of reduced conflict probabilities has been debated for a long time. A considerable amount of studies supports Pinker's joyful message. Gat (2006), Goldstein (2011) and Cunen, Hjort, and Nygård (2019) for example, all agree with Pinker that the world has become more peaceful. Cunen, Hjort and Nygård even give an estimate of when this change took place. Through statistical change point analysis, they argue that the war-generating mechanism got less intense somewhere during the sixties.

Other authors are less sanguine. Clauset (2017, 2018) is a case in point. Clauset argues that it is still too early to draw conclusions from the current trend of relative peace. According to him, this trend has to last another hundred years before we can state anything with confidence about reduced conflict probabilities. Still, even if the thesis of the 'long peace' (Gaddis, 1989) is somewhat contested, there seems to be clear evidence that *democracies* are highly unlikely to go to war with each other. (See Hegre (2014) for a summary of findings.) So even if Pinker's thesis may be too optimistic, statistical analyses reveal that free people in liberal democracies do not wage war against each other. A message hardly less delightful to convey to the general public.

The potential of statistical analysis in conflict research is however greater than just being a tool for testing hypotheses about the progress of the world. Statistical analysis may in itself be a contributor to a decline in violence. Consider all armed conflicts that have taken place in modern times. Of course, each of these conflicts is unique in its own right. Each conflict has its particular agents, its particular stakes, its particular historical causes. Nevertheless, it would be very strange if there were no common patterns across these conflicts. We would be very surprised if an increase in arms expenditure did not have any effect on the probability of war for example. (If this were the case, all states are wasting a lot of money). We would also be surprised if the form of government had absolutely nothing to say for the war chances.

To identify such common patterns, dynamic regression is the appropriate choice of method. Dynamic regression methods are able to identify numerical patterns in the data, decide which effects are significant and even make us able *predict* violent conflict in the future!

It is not difficult to see the practical value of this. The international community, for example the United Nations, could use such dynamic regression models to monitor current conflict levels around the globe. Aid and attention could be directed to the areas identified as hot spots. In this way, conflicts could be stopped even before they erupted. The general public would be thrilled!

Dynamic regression models for conflict prediction is no idealist's dream of the future. Hegre, Karlsen, Nygård, Strand, and Urdal (2013) have used *multinomial* Markov chain models with a *logit* link to predict future of civil wars. Basing their analysis on data on civil wars after 1946, they use such dynamic regression models to predict the future. Their predictions are encouraging. They predict that the coming years will see a decline in intrastate violence. They estimate probabilities of civil war eruption in different countries, they identify potential hot spots and they even identify effects that tend to increase the probability of an outbreak of civil war. This is exactly what would be of great use to the international community.

The question of model selection is important, to such dynamic conflict modeling, as it is to all statistical modeling. There is a multitude of models that can be fitted to the data. How to select the best one? We could of course use traditional information criteria, such as the *Aikake's Information Criterion* (AIC), the *Bayesian Information Criterion* (BIC) or the *Deviance Information Criterion* (DIC). These well-known and widely used criteria aim at selecting the model closest to the true data-generating mechanism.

The potential problem is that the model deemed closest to the true data-generating mechanism not necessarily is the model best at estimating the parameter of interest, the *focus parameter* of the analysis. It may be that the model deemed closest to the true data-generating mechanism includes so many parameters that it renders the final estimates of the focus parameter imprecise. Simpler models may be preferable. Such simple models would probably be biased, but due to their simplicity, they may involve so much less variance that they nevertheless give more stable estimates of the focus parameter. This is all the more true in the modeling of conflict dynamics. War rarely breaks out, peace observations are abundant. The model chosen by the AIC may be much too wide to estimate war-related parameters, as the probability of escalation from minor conflict to war, for example.

The *Focused Information Criterion* (FIC) is an information criterion that aims at selecting the models in regard to the precision of the focus parameter estimator, rather than the closeness to the true data-generating mechanism. Compared to the AIC and its relatives, however, the FIC is a mathematically complex criterion. It is not readily available through a general formula, as are the handy AIC and BIC. Being based on large sample asymptotics of maximum likelihood estimators under misspecification, it needs to be worked out uniquely for each class of parametric models.

In this Master's thesis I develop such a *Focused Information Criterion* for the dynamic multinomial *logit* model of Markov chains. Inspired by the analysis of Hegre et al. (2013) I will use the developed FIC to analyze interstate conflict dynamics in the period between 1950 and 2010. The data to be analyzed will be the Military Interstate Disputes data set of the Correlates of War project (Maoz, Johnson, Kaplan, Ogunkoya, & Shreve, 2018). The focus parameter of the analysis will be the probability of escalation from minor conflict to war. We will be particularly interested in assessing the effect of democracy on this escalation probability.

1.1 Markov Chains

Markov chain models are a natural choice when it comes to the modeling of conflict dynamics. Current conflict probabilities may be dependent on past conflict levels and we should allow for such dependency on the past in our models.

Recall that a Markov Chain is a stochastic process where the probability distribution of the current event depends on the states of past events. Consider a time series $\{y_t\}$ for $t = 0, 1, \ldots n$. Define K different categories such that y_t may take values $1, \ldots, K$. A categorical time series is a p'th order finite Markov Chain if it is the case that

$$P(y_t = j | y_{t-1}, \ldots, y_0) = P(y_t = j | y_{t-1}, \ldots, y_{t-p}), \qquad j = 1, \ldots K,$$

with initial probabilities

$$P(y_0 = j), \qquad k = 1, \ldots K.$$

Thus in a p'th order Markov chain the probability distribution of y_t is conditioned on the p past values $y_{t-1}, \ldots, y_{t-p}$.

For regression models of higher order Markov chains, the number of parameters quickly becomes immense. We therefore restrict ourselves to Markov chains of the first order in this thesis. For such first-order Markov chains, we denote the transition probabilities

$$\pi_{kj} = P(y_t = j | y_{t-1} = k), \qquad k, j = 1, \ldots K.$$

We denote the $K \times K$ transition probability

$$\mathbf{P}(t) = \begin{pmatrix} \pi_{11}(t) & \pi_{12}(t) & \cdots & \pi_{1K}(t) \\ \pi_{21}(t) & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \pi_{K1}(t) & \cdots & \cdots & \pi_{KK}(t) \end{pmatrix}.$$

The probability of the chain being in state k at time t and then in state j at time $t+s$ we denote $P_{kj}^{(s)}(t)$. This probability is the (k, j)'th element of the *forward matrix*, which for $s = 0$ is $P^{(0)}(t) = I$. For $s \geq 1$ it is

$$\mathbf{P}^{(s)}(t) = \mathbf{P}(t+1)\mathbf{P}(t+2) \cdots \mathbf{P}(t+s).$$

Now, a Markov chain may either be *homogeneous* or *inhomogeneous*. A Markov chain is *homogeneous* if the probability of going from one state to another state is independent of the time t at which the transition takes place. In this case $\mathbf{P}(t) = \mathbf{P}$ for all t. The constancy of the transition matrix ensures that the limiting behavior of the *homogeneous* Markov chain has a clean mathematical formulation. Due to this desirable property, *homogeneous* Markov chains are extensively studied. See for example Karlin and Taylor (1975), Meyn and Tweedie (1993).

The *inhomogeneous* Markov chain is a much more complicated creature. In this class of Markov chains the transition probabilities change with t so that the transition matrix $\mathbf{P}(t)$ is *not* constant. The changing nature of the transition matrix may result in very complicated limiting behavior of *inhomogeneous* Markov chains, if such a limiting behavior exists at all. Early students of *inhomogeneous* Markov chains are Dobrushin (1956), Sarymsakov (1953) and Hajnal (1956, 1958). For a summary of fundamental concepts see Seneta (2014).

Homogeneous and *inhomogeneous* Markov chains need to be ergodi to have a limiting behavior. According to Hajnal (1958) a Markov chain is *weakly ergodic* if there for each t exists a $K \times 1$ vector function $\pi(t) = (\pi_1(t), \ldots, \pi_K(t))^{\mathrm{t}}$ of limiting probabilities. This is equivalent to saying that

$$\lim_{s \to \infty} \left| P_{kj}^{(s)}(t) - \pi_j(t) \right| = 0.$$

When the number of transitions goes to infinity, the probability of a *weakly ergodic* Markov chain ending in state j will be independent of the state k where it started. In other words, the chain has forgotten where it started. The chain has not completely forgotten its past, however, as it still remembers the probability operators associated with t.

A complete loss of memory the chain has only if it is *strongly ergodic*. For *strongly ergodic* Markov chains there exists a $K \times 1$ vector $\pi = (\pi_1, \ldots, \pi_K)^{\mathrm{t}}$ of limiting probabilities which is independent of t. This is equivalent to saying that

$$\lim_{s\to\infty}\left|P_{kj}^{(s)}(t)-\pi_j\right|=0.$$

We see that a *strongly ergodic* chain has completely forgotten its past as the long term behavior of the chain is the same at each time t.

To be able to demonstrate the asymptotic behavior of Markov chain models under misspecification, a first task will be to show that the models under study fulfill the conditions of *strong ergodicity*.

1.2 Regression Models for Markov Chains

Regression models for Markov chains have their applications in all fields where dynamic systems are studied. Examples are as diverse as medicine, genetics, engineering, economics and meteorology, in addition of course to the study of international relations.

Parametric Markov chain models may be elegantly expressed in the framework of *generalized linear models*. In this framework, the transition probabilities $\pi_{kj}(t)$ are modeled as a function of a covariate vector x_t and a parameter vector β such that

$$\pi_{kj}(t,\beta)=h(x_t^{\mathrm{t}}\beta),\qquad k,j=0,\ldots,K,$$

where $h(\cdot)$ is an appropriate *link* function. In the cases where the covariate vector x is constant with t, the resulting Markov chain will be *homogeneous*. In the cases where the covariate vector x varies with t, the resulting Markov chain will be *inhomogeneous*.

The Markov model used by Hegre et al. (2013) to study of civil war is on this elegant form of generalized linear models. These authors use a *multinomial* Markov model with *logit* link. Statistical theory for this model is developed in Kaufmann (1987), Fahrmeir and Kaufmann (1987), and more recently in Fokianos and Kedem (1998, 2003) and Kedem and Fokianos (2002). Letting the number of states be $K=3$, defining level 2 as a baseline category and letting $\beta=(\beta_0^{\mathrm{t}},\beta_1^{\mathrm{t}})^{\mathrm{t}}$ be the total parameter vector, the transition probabilities may be expressed as

$$\pi_{tj}(\beta)=\frac{\exp(z_t^{\mathrm{t}}\beta_j)}{1+\exp(z_t^{\mathrm{t}}\beta_0)+\exp(z_t^{\mathrm{t}}\beta_1)}.\tag{1.1}$$

Here z_t is a vector that may consist of elements from a vector of covariates x_t, but also elements from the vector of interaction with past values $x_t y_{t-1,k}$, where $k=0,1,2$.

Other versions of *multinomial* Markov chain models may be found by choosing different *link* functions than the *logit* function. Przeworski, Alvarez, Cheibub, and Limongi (2000) for example, suggest using a multinomial Markov model with *probit* link to analyze the relationship between development and democracy. Kedem and Fokianos (2002) present an overview of common *link* functions that may be used within the framework. Brillinger (1996) considers the case for ordinal data.

There are also regression models for *inhomogeneous* Markov chains other than the *multinomial* models. In the case of survival data a non-parametric model is suggested by Aalen and Johansen (1978). A healthy variety of parametric *Hazards Models* is presented in Martinussen and Scheike (2006). Other examples of parameteric regression models for Markov chains are Mixture Transition Distribution models proposed in Berchtold and Raftery (2002) and Time transformed Markov models proposed in Hubbard, Inoue, and Fann (2008). Brillinger, Morettin, Irizarry, and Chiann (2000) devevlop a Wavelet-based approach.

In this thesis we restrict ourselves to the *multinomial* regression model with *logit* link. Generalizations should be possible to reach.

1.3 Statistical Inference for Markov Chain Models

We will fit the dynamic multinomial *logit* model with maximum likelihood estimation. Given data y_{obs}, a model may be constructed which has joint distribution $f_{\text{joint}}(y|\theta)$, where θ is the model parameter to be estimated. The maximum likelihood estimate $\hat{\theta}$ of this constructed parametric model is the parameter value that maximizes the likelihood

$$L(\theta|y_{\text{obs}}) = f(y_{\text{obs}}|\theta).$$

This means that $\hat{\theta}$ is the point in the parameter space at which the observed sample is the most likely. The invariance property of the maximum likelihood estimator ensures that for any function $\tau(\theta)$, the maximum likelihood estimate of $\tau(\theta)$ is $\tau(\hat{\theta})$. See Casella and Berger (2002, pp. 320–1)

In the case of independent and identically distributed data, the maximum likelihood estimator $\hat{\theta}$ is *consistent* under mild regularity conditions. Consistency here means that when the number of observations n grows, the maximum likelihood estimator converges almost surely to the *true* parameter value θ_{true}. This we may write as

$$\hat{\theta} \overset{a.s.}{\to} \theta_{\text{true}}.$$

Further, under the same mild regularity conditions, it is the case that

$$\sqrt{n}(\hat{\theta} - \theta) \to_p \mathrm{N}\Big(0, J^{-1}\Big)$$

where J is the Fisher information matrix of the model. See Casella and Berger (2002, Section 10.1). This crucial result implies that maximum likelihood estimates are approximately normally distributed about the true value when the sample size n is large.

Outside the assumptions of i.i.d. data, things are somewhat more complicated. Billingsley (1961a, 1961b) and Basawa and Prakasa Rao (1980) show that maximum likelihood estimation is applicable also to data from *homogeneous* Markov chains. Via Martingale arguments they show that asymptotic consistency and normality is the case also in the situation of *homogeneous* Markov chains.

Dobrushin (1956) develops a central limit theorem for *inhomogeneous* Markov chains. This theorem is proven with Martingale arguments by Sethuraman and Varadhan (2005). Kaufmann (1987) and Fahrmeir and Kaufmann (1987) show that maximum likelihood estimation is applicable to the inhomogeneous Markov Chain Model (1.1) under the condition of *ergodicity*.

All this is for the situation where the constructed parametric model is correctly specified. The data-generating mechanism may however be different from the parametric model chosen to analyze the data. In this case we have to take model misspecification into consideration.

For i.i.d. data, White (1982) shows that the maximum likelihood estimators $\hat{\theta}$ will be consistent and normally distributed asymptotically also under model misspecification. The maximum likelihood estimator $\hat{\theta}$ now converges, not to the *true* parameter value, but to the least false parameter value θ_0 that minimizes the Kullback-Leibler distance from the parametric model to the true data-generating mechanism. This we may write as

$$\hat{\theta} \overset{a.s.}{\to} \theta_0.$$

For a misspecified model it is also the case that

$$\sqrt{n}(\hat{\theta} - \theta_0) \to_p \mathrm{N}\Big(0, J^{-1}KJ^{-1}\Big),$$

where the matrix K is the variance of the random score function. In the case of correctly specified models $K = J$, but this is not the case when models are misspecified. Thus the variance matrix will be a ‘sandwich’ matrix $J^{-1}KJ^{-1}$. See also Claeskens and Hjort (2008b, Chapter 2).

The *Focused Information Criterion* is based on the asymptotic properties of maximum likelihood estimators under misspecification. An important part of this thesis will therefore be to show that the maximum likelihood estimator $\hat{\beta}$ of dynamic *multinomial logit* models for Markov chains will be consistent and normally distributed asymptotically also under model misspecification.

1.4 The Focused Information Criterion

The main objective of this thesis is to develop a *Focused Information Criterion* for dynamic multinomial logit models. The FIC is an information criterion different in its essence from the well known and widely used criteria AIC, BIC and DIC. Whereas the latter criteria aim at selecting the model closest to the true data-generating mechanism, the FIC aims at selecting the model with the most precise estimator of a *focus parameter*. The AIC and its relatives work in an 'overall' modus, they are off the shelf methods *prêt-à-porter*. The FIC takes on board the intended use of the models, it is a criterion tailored to the modeling purpose, it is *haute couture*.

Which criterion to use depends on the goal of the selection procedure. If the objective of model selection mainly is to understand the true data-generating mechanism, the AIC, BIC and the DIC are proper choices of selection strategies. These criteria are based on the likelihood $\ell^{(n)}(\beta)$ of the fitted models. As such they are blessed with simple formulas that remain the same across all likelihood based models. The formula for the AIC is for example $AIC = 2\ell^{(n)}(\hat{\beta}) - 2p$, where p is the number of parameters in the fitted model. The simplicity and uniformity of these criteria contribute surely to their popularity and wide-spread use.

The FIC is not based directly on the likelihood of the fitted models. Rather, it is based on the estimated mean squared error of the maximum likelihood estimator $\hat{\mu}$ of the *focus parameter*.

The mean squared error of $\hat{\mu}$ is given by

$$\mathrm{mse}(\hat{\mu}) = \mathrm{Var}\ \hat{\mu} + (\mathrm{E}\ \hat{\mu} - \mu)^2.$$

The second term is here the bias of $\hat{\mu}$. If a fitted model is far from the true data-generating mechanism, this bias will typically (but not always) be considerable. Still, if the model is simple it may render very little variance in estimates, which will result in a low value of the mean squared error of $\hat{\mu}$, despite the bias. On the other hand, a model close to the true data-generating mechanism will typically have estimators $\hat{\mu}$ that are close to unbiased. But the model may be so complex that it has high variance of $\hat{\mu}$. The resulting mean squared error of $\hat{\mu}$ may be considerable, although the model involves no bias. So even though models far from the *true* data-generating mechanism may be biased, they may still be the models giving the most precise estimates of the focus parameter μ. The FIC aims at selecting the model which strikes the best balance between bias and variance.

The fic score is defined as

$$\mathrm{fic} = \widehat{\mathrm{mse}}(\hat{\mu}) = \widehat{\mathrm{Var}}\,\hat{\mu} + \widehat{\mathrm{bsq}}.$$

The model with the lowest fic score is the model estimated to give the most precise estimates of μ and is thus the model selected by the FIC.

To calculate the estimated mean squared error, a *true* model need to be presumed. This should be a rather complex model, a model that includes all effects of possible explanatory value to the response variable. In the FIC literature, the model chosen to be the true model is called the *wide* model. We follow this usage in this thesis. Models fitted to the data that are different from the *wide* model are called candidate models. These models will be misspecified under the true *wide* model. The $\widehat{\mathrm{mse}}\,\hat{\mu}$ of candidate models is then calculated based on the large-sample approximations of the maximum likelihood estimators under this mispecification.

The FIC comes in two versions. The difference between the two resides in how the misspecification context is defined. See, e.g., Claeskens, Cunen, and Hjort (2019) for an presentation of both versions.

The first version, originally developed in N. L. Hjort and Claeskens (2003) and Claeskens and Hjort (2008b), takes place in a local misspecification context. In this context the *wide* model is considered to be only $O(n^{-\frac{1}{2}})$ away from a narrow model. Thus, the *wide* model will change with sample size, coming closer and closer to the narrow model. Candidate models are supposed to lie between the narrow and the *wide* models.

The virtue of this original approach is that it results in clean mathematical formulas for the fic values. On the other hand, it may place too heavy restrictions on the class of potential candidate models. Candidate models need to be submodels of the *wide* model, and for some model classes, this may exclude many interesting cases. Nevertheless, this original FIC scheme has been applied with success to a wide range of model types. It has for example been developed for generalized linear models by Claeskens and Hjort (2008a), Cox regression models by N. Hjort and Claeskens (2006) and generalized additive linear models by Zhang and Liang (2011).

Recently a second, more flexible Information Criterion has been developed. In this FIC scheme, a *fixed* model is considered to be the true, *wide* model. Jullum and Hjort (2017) have developed such a version for i.i.d data where the empirical distribution plays the role of the fixed true model. Ko, Hjort, and Hobæk Haff (2019) use a fixed model approach in the development of FIC for copulae models. Cunen, Walløe, and Hjort (2019) have developed a FIC with a fixed true model for Linear Mixed Models.

The virtue of this second type of FIC with a fixed wide model is that candidate models may be at any distance from the *wide* model. The possible disadvantage is that expressions may get very complicated, as we will see.

According to Cunen, Walløe, and Hjort (2019) and Claeskens et al. (2019) the misspecification of candidate models under the fixed *wide* model, should lead to the approximate joint normal distribution of maximum likelihood estimates on the form

$$\begin{pmatrix} \sqrt{n}(\hat{\theta} - \theta_{\mathrm{true}}) \\ \sqrt{n}(\hat{\theta}_M - \theta_{M,0,n}) \end{pmatrix} \approx_d \mathrm{N}\left(0, \begin{pmatrix} J_n^{-1} & J_n^{-1} C_{M,n} J_{M,n}^{-1} \\ J_{M,n}^{-1} C_{M,n} J_n^{-1} & J_{M,n}^{-1} K_{M,n} J_{M,n}^{-1} \end{pmatrix}\right). \tag{1.2}$$

Here $\hat{\theta}$ is the ML estimator of the wide model, $\hat{\theta}_M$ is the ML estimator of the misspecified candidate model, whereas $\theta_{M,n,0}$ is the least false parameter value of the candidate model. The matrices $J_n, J_{M,n}, C_{M,n}$ and $K_{M,n}$ are appropriate information matrices and variance matrices of random score functions.

The central point in the second FIC scheme, is that this approximate joint distribution of maximum liklihood estimates, via delta arguments leads to the following approximate joint distribution of maximum likelihood estimates of focus parameters $\hat{\mu}$ of the *wide* model and $\hat{\mu}_M$ of the candidate model. This approximate joint distribution will be on the form:

$$\begin{pmatrix} \sqrt{n}(\hat{\mu} - \mu_{\text{true}}) \\ \sqrt{n}(\hat{\mu}_M - \mu_{M,0,n}) \end{pmatrix} \approx_d \mathrm{N}\left(0, \begin{pmatrix} \nu_{\text{wide}} & \nu_{M,c} \\ \nu_{M,c} & \nu_M \end{pmatrix}\right). \tag{1.3}$$

where $\mu_{M,n,0}$ is the least false focus parameter value and ν_{wide} and ν_M the appropriate variances, and $\nu_{M,c}$ the appropriate covariance. This approximate joint distribution of the focus parameter estimates enables estimation of mse$\hat{\mu}$ of the *wide* model as well as mse$\hat{\mu}_M$ of misspecified candidate models.

The *Focused Information Criterion* we develop in this thesis for dynamic multinomial logit models will be a FIC with a fixed true model. This means that we need to show that (1.2) and (1.3) hold also for misspecified dynamic multinomial logit models. This will be no trivial undertaking. As a guide through the wilderness, we will rely heavily on the proceedings in Cunen, Walløe, and Hjort (2019).

1.5 Outline

The rest of the thesis is organized as follows:

Chapter 2 In this chapter, I describe the model setup of the thesis in detail. Data are considered to be independent *inhomogeneous* Markov Chains of order one. I describe how the dynamic multinomial logit model can be used to model the transition matrices of such *inhomogeneous* Markov Chains. I find expressions for log-likelihood, score function and Fisher information Matrix of the dynamic multinomial logit model. I need to take a somewhat different approach than Fokiamos and Kedem, Kaufmann and Fahrmeir as the model must allow for misspecification. In this chapter, I also state regularity assumptions that need to be in place for the large sample theory under misspecification to go through.

Chapter 3 In this chapter, I work out the large sample theory for the dynamic multinomial logit model described in Chapter 2. This I do both for correctly specified models and for misspecified models under a true dynamic multinomial model. I show that maximum likelihood estimators are consistent in that they tend to the least false parameter value and I show that maximum likelihood estimates are approximately normally distributed about this least false parameter.

Chapter 4 In this chapter, I develop the *Focused Information Criterion* for the dynamic multinomial Markov model. I show that the formula (1.3) holds also in the case of this Markov model. I suggest an estimation strategy, as the matrices involved are rather complicated.

Chapter 5 In this chapter, I illustrate the developed theory by simulation studies. I verify that maximum likelihood estimates are normally distributed about the least false parameter value. I also show that the FIC procedure succeeds in selecting the model with the lowest true mean squared error in the estimator of the focus parameters.

Chapter 6 In this chapter, I analyze the MID data set. The focus parameter of the analysis is the probability of conflict escalation. I do focused model selection with two mathematical interpretations of this focus parameter. We will see that the preferred model predicts a decline in escalation probabilities with increasing democracy levels. The decline is however not significant.

Chapter 7 In this chapter, I summarize the achivements of this thesis. I point also to some additional topics that deserve to be further investigated.

Appendix A Proofs and expressions too comprehensive to be included in the text are given in the appendix.

CHAPTER 2

Dynamic Multinomial Models

Multinomial regression models are well-known tools in the statistician's toolbox. Particularly favored are multinomial models with a baseline category. These models are elegantly formulated in the framework of *generalized linear models*. See Agresti (2013, 2015) for an introduction. Such multinomial regression models with logit link may also be used to analyze *inhomogeneous* Markov chains. A GLM inspired class of *dynamic* multinomial regression models are developed by Fahrmeir and Kaufmann (1987), Kaufmann (1987), Fokianos and Kedem (1998, 2003) and Kedem and Fokianos (2002).

The dynamic multinomial regression model developed by these authors is a false start for the objective of developing a *Focused Information Criterion*, however. The FIC calls for the asymptotic distribution of ML estimates under misspecification. Derivation of asymptotic distributions under misspecification will be challenging in the setup of Kaufmann, Fahrmeir, Kedem and Fokianos. These authors allow past responses to be treated as covariates. A sensible thing to do when the model is correctly specified. Under misspecification, however, the correlation between all responses in the chain needs to be accounted for. That will be difficult in the scheme of autoregressive multinomial logit models.

In this chapter, I define a dynamic multinomial logit model in the framework of *generalized linear models*, as do Kaufmann, Fahrmeir, Kedem and Fokianos. I take a slightly different approach when it comes to the treatment of past values, however.

In Section 2.1 I describe the general setup of *inhomogeneous* Markov chains. In section Section 2.2 I define the dynamic multinomial logit model that allows for large sample asymptotics under misspecification. In Section 2.3 I describe the likelihood function, the score vector function and the Fisher Information Matrix of the defined dynamic multinomial logit model. I also state regularity conditions on the covariate distribution that will ensure ergodicity of the Markov chain. In Section 2.4 I show that under these assumptions on the covariate distribution, the dynamic multinomial logit model indeed is strongly ergodic, even though it still constitutes an *inhomogeneous* Markov chain.

2.1 Setup

Consider Markov chains of order one and of length $n+1$ such that $t = 0, 1, \dots, n$. Let there be m independent chains and let $\{y_{i,t}\}$ represent the i'th Markov Chain, where $i = 1, \dots m$.

Let the Markov chains have $K = 3$ categories. Denote each categorical level by $j = 0, 1, 2$. The particular observations $y_{i,t}$ may then be expressed as a 3×1 vector $y_{i,t} = (y_{i,t,0}, y_{i,t,1}, y_{i,t,2})^{\mathrm{t}}$. The elements of this vector are

$$y_{i,t,j} = \begin{cases} 1 & \text{if chain is in state } j \text{ at time } t, \\ 0 & \text{else.} \end{cases}$$

The categorical levels $j = 0, 1, 2$ of the chain are then represented by

$$\begin{aligned} y_{i,t} &= (1, 0, 0)^{\mathrm{t}}, \\ y_{i,t} &= (0, 1, 0)^{\mathrm{t}}, \\ y_{i,t} &= ((0, 0, 1)^{\mathrm{t}}, \end{aligned}$$

respectively.

Markov chains $\{y_{i,t}\}$ are allowed to be in one and only one conflict level at each time t. We therefore have that

$$\sum_{j=0}^{2} y_{i,t,j} = 1.$$

For each Markov chain $\{y_{i,t}\}$, let there be a time series of covariates $\{x_{i,t}\}$. Assume that there are p different explanatory variables, such that each element of $x_{i,t}$ will be a $p \times 1$ vector $(x_{i,t,1}, \dots x_{i,t,p})^{\mathrm{t}}$. The first element in this vector may be an intercept such that $x_{i,t,1} = 1$. The remaining variables may either be quantitative or qualitative.

Assume now that the transition probabilities of the Markov chains $\{y_{i,t}\}$ depend on $x_{i,t}$. Let $\pi_{kj}(x_{i,t})$ denote the transition probability from state k at time $t-1$ to state j at time t. For $k = 0, 1, 2$ and $j = 0, 1$ the transition probability is then

$$\pi_{k,j}(x_{i,t}) = P(y_{i,t,j} = 1 \mid y_{i,t-1,k} = 1, x_{i,t}).$$

The transition matrix of the chain is

$$\mathbf{P}(x_{i,t}) = \begin{pmatrix} \pi_{0,0}(x_{i,t}) & \pi_{0,1}(x_{i,t}) & \pi_{0,2}(x_{i,t}) \\ \pi_{1,0}(x_{i,t}) & \pi_{1,1}(x_{i,t}) & \pi_{1,2}(x_{i,t}) \\ \pi_{2,0}(x_{i,t}) & \pi_{2,1}(x_{i,t}) & \pi_{2,2}(x_{i,t}) \end{pmatrix}.$$

Each row in this matrix sum to one such that

$$\sum_{j=0}^{2} \pi_{k,j}(x_{i,t}) = 1 \quad k = 0, 1, 2.$$

As covariate values $x_{i,t}$ may change with time, $\mathbf{P}(x_{i,t})$ varies with t and the resulting Markov chain is *inhomogeneous*.

2.2 The Dynamic Multinomial Logit Model

The multinomial regression model for *inhomogeneous* Markov chains developed in Fahrmeir and Kaufmann (1987), Kaufmann (1987) and Kedem and Fokianos (2002) is elegantly framed in the scheme of *generalized linear models*. With *logit* link this class of models is expressed as

$$\pi_{tj}(\beta) = \frac{\exp(z_{i,t}^{\mathrm{t}}\beta_j)}{1 + \exp(z_{i,t}^{\mathrm{t}}\beta_0) + \exp(z_{i,t}^{\mathrm{t}}\beta_1)}, \tag{2.1}$$

Here $\pi_{tj}(\beta)$ denotes the probability of $y_{t,j} = 1$. The vector $z_{i,t}$ is a vector that may include elements from covariate vector $x_{i,t}$, but also elements from the vector of interactions with past levels of observation $x_{i,t}y_{i,t-1}$ (See Kedem and Fokianos (2002, p. 93).

A particular advantage of this autoregressive Markov chain model is its potential sparsity. Typically, Markov chain models will be rather baroque. Markov chain models typically need a high number of parameters to incorporate the dependency of past observation level $y_{i,t-1}$. For a first-order Markov chain model, there will typically be a set of parameters β_k for each past observation level k.

The autoregressive multinomial model (2.1) allows for a drastic reduction in the number of model parameters. In this model, not all effects of covariates need to incorporate the Markov assumption. Covariates $x_{i,t}$ which are represented in the vector $z_{i,t}$ without interaction $x_{i,t}y_{i,t-1}$ elements have effects that do not change with past observation levels $y_{i,t-1}$. We call these effects *Markov independent effects*.

Covariates $x_{i,t}$ which are represented in $z_{i,t}$ with interaction element $x_{i,t}y_{i,t-1}$ do have effects which are dependent on past observation level $y_{i,t-1}$. We call these effects *Markov dependent effects*. Thus the resulting dynamic multinomial logit model (2.1) may include both Markov dependent effects and Markov independent effects.

This implies that the class of models on the form of (2.1) is comprehensive. The widest models in the class are the models with full Markov dependency in all effects. The most parsimonious models in the class are the models with no Markov dependency in any effects. This is in fact the standard multinomial model. Between those two extremes, there is a huge subclass of intermediate models with a mixture of Markov dependent and Markov independent effects.

As a consequence of this, the model class renders particular flexibility in the modeling process. We are not constrained by any Markov assumption. If the Markov dependency is wrong, or if it is fully explained by the covariates, the simple *multinomial* model may be chosen as the best model, although the Markov assumption was fully reasonable *a priori*.

To develop a FIC for dynamic multinomial logit models, we need to take a slightly different approach than Kedem, Fokianos, Kaufmann and Fahrmeir. We need a model setup that allows for model misspecification. As mentioned in the introduction to this chapter, this will be difficult when past observations are allowed to be treated as covariates. But we would like our approach to render the same flexibility as the model (2.1) defined by these authors. Our class of dynamic multinomial models should also include all models from the widest models with Markov dependency in all effects to the narrowes multinomial model with no Markov dependency in any effects.

We approach the issue by first fitting a separate multinomial model to each line in the transition matrix $\mathbf{P}(x)$ of the chain. Conditioned on past conflict level $y_{i,t-1,k}$ and covariate values $x_{i,t}$, the stochastic variable $y_{i,t}$ has then a *multinomial* distribution with one trial. This conditional distribution may consequently be expressed as

$$f(y_{i,t}|x_{i,t}, y_{i,t-1,k} = 1) = \prod_{j=0}^{2} \pi_{k,j}(x_{i,t})^{y_{i,t,j}y_{i,t-1,k}},$$

where $0 < \pi_{kj}(x_t) < 1$ and $k = 0, 1, 2$.

As there is three levels of past states, the total dynamic model of the time series will be the composite model

$$f(y_{i,t}|x_{i,t}, y_{i,t-1}) = \prod_{k=0}^{2}\prod_{j=0}^{2} \pi_{k,j}(x_{i,t})^{y_{i,t,j}y_{i,t-1,k}}.$$

The conditional covariance of $y_{i,t,j}$ in this total dynamic model is

$$\begin{aligned}\mathrm{Cov}\left\{\left(y_{i,t,j}\middle|x_{i,t}, y_{i,t-1}\right), \left(y_{i,t,j'}\middle|x_{i,t}, y_{i,t-1}\right)\right\}\\ = \begin{cases}\pi_{k,j}(x_{i,t})(\delta_{jj'} - \pi_{k,j'}(x_{i,t})) & \text{if } k = k'\\ 0 & \text{if } k \neq k',\end{cases}\end{aligned}$$

for $k, k', j, j' = 0, 1, 2$. Expressed as a block matrix, we may also write this conditional covariance

$$\begin{aligned}\mathrm{Cov}\left\{\left(y_{i,t}\middle| x_{i,t}, y_{i,t-1,k}=1\right), \left(y_{i,t}\middle| x_{i,t}, y_{i,t-1,k'}=1\right)\right\} \\ = \begin{pmatrix}\Lambda_1 & 0 & 0\\ 0 & \Lambda_2 & 0\\ 0 & 0 & \Lambda_3\end{pmatrix},\end{aligned}$$

where the blocks are

$$\Lambda_k = \begin{pmatrix}\pi_{k,0}(1-\pi_{k,0}) & -\pi_{k,0}\pi_{k,1} & -\pi_{k,0}\pi_{k,2}\\ -\pi_{k,0}\pi_{k,1} & \pi_{k,1}(1-\pi_{k,1}) & -\pi_{k,1}\pi_{k,2}\\ -\pi_{k,0}\pi_{k,2} & -\pi_{k,1}\pi_{k,2} & \pi_{k,2}(1-\pi_{k,2})\end{pmatrix}$$

and the probability $\pi_{k,0}$ is an abbreviation for $\pi_{k,0}(x_{i,t})$.

Now, as the first step towards a general dynamic model, we fit the probability vectors $\pi_k(x_t)$ of each previous level $k = 0, 1, 2$ with a *baseline category logit* model. We define $j = 2$ to be the baseline category. For each previous level k of $y_{i,t-1}$ we then have that

$$\log\left(\frac{\pi_{k,j}(x_{i,t})}{\pi_{k,2}(x_{i,t})}\right) = x_{i,t}^t \beta_{k,j} \qquad k = 0, 1, 2, \quad j = 0, 1,$$

where $\beta_{k,j} = (\beta_{k,j,1}, \ldots, \beta_{k,j,p})^{\mathrm{t}}$ is a $p \times 1$ dimensional vector of parameters.

We set the parameter vector of the baseline category $j = 2$ to zero for each k and rewrite

$$\pi_{k,j}(x_{i,t}) = \frac{\exp(x_{i,t}^t \beta_{k,j})}{1 + \sum_{h=0}^{1} \exp(x_{i,t}^t \beta_{k,h})} \qquad k = 0, 1, 2, \quad j = 0, 1. \tag{2.2}$$

Since the Markov chain $\{y_{i,t}\}$ has three previous levels $k = 0, 1, 2$, the total model of the chain will consist of three such baseline category logit models. The total parameter β of this composite model will be a $3 \cdot 2 \cdot p \times 1$ vector $\beta = (\beta_0^{\mathrm{t}}, \beta_1^{\mathrm{t}}, \beta_2^{\mathrm{t}})^{\mathrm{t}}$ where $\beta_k = (\beta_{k,0}^{\mathrm{t}}, \beta_{k,1}^{\mathrm{t}})^{\mathrm{t}}$ for $k = 0, 1, 2$.

This dynamic logit model includes now only models with full Markov dependency in all effects. To allow for more parsimonious models, we will consider some of the effects to be independent of past level k.

If the effect of the r'th covariate is independent of past level k, it is the case that $\beta_{0,j,r} = \beta_{1,j,r} = \beta_{2,j,r}$ for $j = 0, 1$. Let q be the number of covariates in the model with such Markov independent effects. We denote the subset of the covariate vector $x_{i,t}$ which has Markov dependent effects by $u_{i,t} = (u_{i,t,1}, \ldots u_{i,t,q})^{\mathrm{t}}$. The Markov independent effects of $u_{i,t}$ will be a $2\,q \times 1$ vector $\gamma = (\gamma_0^{\mathrm{t}}, \gamma_1^{\mathrm{t}})^{\mathrm{t}}$, where $\gamma_j = (\gamma_{j1}, \ldots, \gamma_{jq})^{\mathrm{t}}$.

For a dynamic multinomial model which takes p covariates into consideration, there will then be $w = p - q$ covariates with Markov dependent effects. We denote these covariate by $z_{i,t} = (z_{i,t,1}, \ldots z_{i,t,w})^{\mathrm{t}}$. The Markov dependent effects of $z_{i,t}$ will be a $2 \cdot 3\,w \times 1$-vector $b = (b_0^{\mathrm{t}}, b_1^{\mathrm{t}}, b_2^{\mathrm{t}})^{\mathrm{t}}$, where $b_k = (b_{k0}^{\mathrm{t}}, b_{k1}^{\mathrm{t}})$ and $b_{kj} = (b_{kj1}, \ldots, b_{kjw})^{\mathrm{t}}$.

In this extended framework, we may express the total covariate vector as the $q + w \times 1$ vector

$$x_{i,t} = \begin{pmatrix} u_{i,t} \\ z_{i,t} \end{pmatrix}.$$

The total parameter vector β of both Markov dependent and Markov independent effects is the $2q + 2 \cdot 3w \times 1$ vector

$$\beta = \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ b_0 \\ b_1 \\ b_2 \end{pmatrix}.$$

When we also take Models with Markov dependent effects into consideration, the *multinomial logit* model of each row in the transition matrix becomes

$$\log \frac{\pi_{k,j}(x_{i,t})}{\pi_{k,2}(x_{i,t})} = u_{i,t}^t \gamma_j + z_{i,t}^t b_{k,j} \qquad k = 0, 1, 2, \quad j = 0, 1.$$

or equivalently

$$\pi_{k,j}(x_{i,t}) = \frac{\exp(u_{i,t}^t \gamma_j + z_{i,t}^t b_{k,j})}{1 + \sum_{h=0}^{1} \exp(u_{i,t}^t \gamma_j + z_{i,t}^t b_{k,j})} \qquad k = 0, 1, 2, \quad j = 0, 1. \tag{2.3}$$

This class of models is just as flexible as the model defined by Kedem, Fokianos, Kaufmann and Fahrmeir. If $u_{i,t} = \emptyset$ and $z_{i,t} = x_{i,t}$ we have a model with full Markov dependency in all effects. Conversely, if $u_{i,t} = x_{i,t}$ and $z_{i,t} = \emptyset$ we have the standard multinomial model. Between those extremes lies the huge class of models with both Markov dependent and Markov independent effects.

Notice that it is not possible to fit each of the three multinomial submodels separately. As the γ-parameters are common across past levels, all submodels have to be fitted simultaneously.

2.3 Conditional Likelihood, Score Vector and Fisher Information Matrix

We now turn to the question of inference for this dynamic *multinomial logit* model. We derive analytical expressions for the log-likelihood, the score vector and the Fisher information matrix.

We choose to condition the inference on responses $y_{i,t}$ only. We assume that the time series of covariates $\{x_{i,t}\}$ are generated by an underlying unknown covariate distribution. We denote the marginal distribution of each observation

$x_{i,t}$ by $C(x)$. The joint distribution of all $m \cdot n$ covariate vectors x_{tot} we denote $C_{\mathrm{joint}}(x_{\mathrm{tot}})$.

To ensure nice behavior of the covariates, we assume the underlying distribution $C(x)$ implies that covariate vectors x almost surely lie in a non-random compact subset $\Gamma \subset R^p$. We also assume that the marginal covariate distribution $C(x)$ is such that for every continuous and bounded function f on Γ it is the case that

$$\frac{\sum_{t=1}^{n} f(x_t)}{n} \rightarrow_p \int f(x) \mathrm{d}C(x).$$

It should also be that case that for any x in Γ it is the case that xx^{t} is positive definite.

These assumptions would have been sufficient to give asymptotic theory for correctly specified models. They correspond to the assumptions made by Kaufmann (1987) and Kedem and Fokianos (2002). We are however to develop large sample asymptotics also for misspecified models. For this to be possible we need also the assumption that for each finite $N \in \mathbb{N}$ and each $N\,p \times 1$ vector $x_N = ((x^{(1)})^{\mathrm{t}}, \ldots, (x^{(N)})^{\mathrm{t}})$ consisting of N covariate vectors, it is the case that the joint distribution $C_{\mathrm{joint},N}(x_N)$ of x_N is such that for each bounded function f on $\Gamma \times \ldots \times \Gamma$ it follows that

$$\frac{\sum_{t=N}^{n} f(x_{i,t}, \ldots x_{i,t-N})}{n-N} \rightarrow_p \int_{R^{p \times \cdots \times p}} f(x_N) \mathrm{d}C_{\mathrm{joint},N}(x_N).$$

This stronger assumption will allow us to find non-stochastic limits of covariance matrices under misspecification.

Now, with this assumption of an unknown, well-behaving covariate distribution, the joint distribution of all response variables y_{tot} and covariate values x_{tot} may be expressed as

$$f(y_{\mathrm{tot}}, x_{\mathrm{tot}}) = f(y_{\mathrm{tot}}|x_{\mathrm{tot}}) C_{\mathrm{joint}}(x_{\mathrm{tot}}).$$

We base inference on the likelihood conditioned on the given covariate values x_{tot}. This conditional likelihood is given by

$$L(\beta|x_{\mathrm{tot}}) = f(y_{i,0}|x_{\mathrm{tot}}) \prod_{i=1}^{m} f(y_{i,n}, \ldots, y_{i,1}|x_{\mathrm{tot}}).$$

Under the regularity assumptions on the covariate distribution, the loss of information by maximizing the conditional likelihood goes asymptotically to zero in probability. See also Kaufmann (1987) for this point. In addition, the logarithm of $f(y_{i,0}|x_{\mathrm{tot}})$ will be small in comparison to the logarithm of the subsequent joint distribution $f(y_{i,1}, \ldots, y_{i,n}|x_{\mathrm{tot}})$ when the number n of observation in every independent chain grows. We choose therefore to ignore the first observations $y_{i,0}$ and base inference regarding β on the observations $1, \ldots n$ only.

We then get the following expression for the total conditional likelihood of m chains, each with n observations :

$$\begin{aligned} L(\beta|x_{\text{tot}}) &= \prod_{i}^{m} f(y_{i,n}, \ldots, y_{i,1}|x_i) \\ &= \prod_{i=1}^{m}\prod_{t=1}^{n} f(y_{i,t}|y_{i,t-1}, x_{i,t}) \\ &= \prod_{i=1}^{m}\prod_{t=1}^{n}\prod_{k=0}^{2}\prod_{j=0}^{2} \pi_{k,j}(x_{i,t})^{y_{i,t,j}y_{i,t-1,k}}. \end{aligned}$$

Taking logarithms, and inserting (2.3) for $\pi_{k,j}(x_{i,t})$ we get the conditional log-likelihood

$$\begin{aligned} \ell^{(m,n)}(\beta) &= \sum_{i=1}^{m}\sum_{t=1}^{n} \ell_{i,t}^{(m,n)}(\beta) \\ &= \sum_{i=1}^{m}\sum_{t=1}^{n}\sum_{k=0}^{2}\Bigg\{\Bigg(\sum_{j=0}^{2} y_{i,t,j}\log \pi_{k,j}(x_{i,t})\Bigg)y_{i,t-1,k}\Bigg\} \\ &= \sum_{i=1}^{m}\sum_{t=1}^{n}\sum_{k=0}^{2}\Bigg\{\Bigg(\sum_{j=0}^{1} y_{i,t,j}\log\Bigg(\frac{\exp(u_{i,t}^{\mathrm{t}}\gamma_j + z_{i,t}^{\mathrm{t}}b_{k,j})}{1+\sum_{h=0}^{1}\exp(u_{i,t}^{\mathrm{t}}\gamma_h + z_{i,t}^{\mathrm{t}}b_{k,h})}\Bigg) \\ &\qquad\qquad - y_{i,t,2}\log\Bigg(\frac{1}{1+\sum_{h=0}^{1}\exp(u_{i,t}^{\mathrm{t}}\gamma_h + z_{i,t}^{\mathrm{t}}b_{k,h})}\Bigg)\Bigg)y_{i,t-1,k}\Bigg\} \\ &= \sum_{i=1}^{m}\sum_{t=1}^{n}\sum_{k=0}^{2}\Bigg\{\Bigg(\sum_{j=0}^{1}(u_{i,t}^{\mathrm{t}}\gamma_j + z_{i,t}^{\mathrm{t}}b_{k,j})y_{i,t,j} \\ &\qquad\qquad\qquad - \log\Bigg(1+\sum_{h=0}^{1}\exp(u_{i,t}^{\mathrm{t}}\gamma_h + z_{i,t}^{\mathrm{t}}b_{k,h})\Bigg)\Bigg)y_{i,t-1,k}\Bigg\}. \end{aligned}$$

Strictly speaking, we should be writing $\ell^{(m,n)}(\beta|\mathbf{x}_{\text{tot}})$ for the log-likelihood. For readability we will nevertheless write $\ell^{(m,n)}(\beta)$ in the rest of the thesis. The reader should keep in mind that we are talking about the conditional log-likelihood.

The conditional score vector of the model is the $(2\,q + 6\,w) \times 1$- vector

$$\frac{\partial \ell^{(m,n)}(\beta)}{\partial \beta} = \begin{pmatrix} \frac{\partial \ell^{(m,n)}(\beta)}{\partial \gamma} \\ \frac{\partial \ell^{(m,n)}(\beta)}{\partial b} \end{pmatrix}.$$

The first element $\partial \ell^{(m,n)}(\beta)/\partial \gamma$ is here a $2 \cdot q \times 1$-vector. For each $j = 0, 1$ and each $r = 1, \ldots, q$ the elements in this vector are given by

$$\begin{aligned}
\frac{\partial \ell^{(m,n)}(\beta)}{\partial \gamma_{j,r}} &= \sum_{i=1}^{m}\sum_{t=1}^{n}\sum_{k=0}^{2}\bigg\{\bigg(\frac{\partial}{\partial \gamma_{j,r}}\sum_{j=0}^{1}(u_{i,t}^{\mathrm{t}}\gamma_j + z_{i,t}^{\mathrm{t}}b_{k,j})y_{i,t,j} \\
&\qquad\qquad - \frac{\partial}{\partial \gamma_{j,r}}\log\bigg(1 + \sum_{h=0}^{1}\exp(u_{i,t}^{\mathrm{t}}\gamma_h + z_{i,t}^{\mathrm{t}}b_{k,h})\bigg)\bigg)y_{i,t-1,k}\bigg\} \\
&= \sum_{i=1}^{m}\sum_{t=1}^{n}\sum_{k=0}^{2}\bigg\{\bigg(y_{i,t,j}u_{i,t,r} \\
&\qquad\qquad\qquad - \frac{u_{i,t,r}\exp(u_{i,t}^{\mathrm{t}}\gamma_j + z_{i,t}^{\mathrm{t}}b_{k,j})}{1 + \sum_{h=0}^{1}\exp(u_{i,t}^{\mathrm{t}}\gamma_h + z_{i,t}^{\mathrm{t}}b_{k,h})}\bigg)y_{i,t-1,k}\bigg\} \\
&= \sum_{i=1}^{m}\sum_{t=1}^{n}\sum_{k=0}^{2}\bigg\{\bigg(y_{i,t,j} - \pi_{k,j}(x_{i,t})\bigg)u_{i,t,r}y_{i,t-1,k}\bigg\}.
\end{aligned}$$

Similarly $\partial \ell^{(m,n)}(\beta)/\partial b$ is a $2 \cdot 3w \times 1$ vector where for each $j = 0, 1$, $k = 0, 1, 2$ and $r = 1, \ldots, w$ the elements are

$$\frac{\partial \ell^{(m,n)}(\beta)}{\partial b_{k,j,r}} = \sum_{t=1}^{m}\sum_{i=1}^{n}\bigg\{\bigg(y_{i,t,j} - \pi_{k,j}(x_{i,t})\bigg)z_{i,t,r}y_{i,t-1,k}\bigg\}$$

The defined dynamic multinomial logit model (2.3) consists of three submodels on the same form as the standard multinomial logit model. We know that the log-likelihood of the standard multinomial logit model is a *concave* function, see Agresti (2015, p. 206). As the log-likelihood of (2.3) consists of the same functions as the log-likelihood of the standard multinomial function, the log-likelihood $\ell^{(m,n)}(\beta)$ will be a *concave* function too. This implies that $\ell^{(m,n)}(\beta)$ has a unique maximum.

The Hessian of the log-likelihood is given by the $(2\,q + 6\,w) \times (2\,q + 6\,w)$ matrix

$$H(\beta) = \nabla^2 \ell^{(m,n)}(\beta) = \begin{pmatrix} \frac{\partial^2 \ell^{(m,n)}(\beta)}{\partial \gamma \partial \gamma^{\mathrm{t}}} & \frac{\partial^2 \ell^{(m,n)}(\beta)}{\partial \gamma \partial b^{\mathrm{t}}} \\ \frac{\partial^2 \ell^{(m,n)}(\beta)}{\partial\ b \partial \gamma^{\mathrm{t}}} & \frac{\partial^2 \ell^{(m,n)}(\beta)}{\partial b \partial b^{\mathrm{t}}} \end{pmatrix}.$$

The blocks in this matrix are found by partial derivation of the score vectors. For all cases below $i, i' = 1, \ldots m$, $k, k' = 0, 1, 2$, $j, j' = 0, 1$, $r, r' = 1, \ldots q$ and $s, s' = 1, \ldots w$. The upper left $2\,q \times 2\,q$ matrix has elements

$$\frac{\partial^2 \ell^{(m,n)}(\beta)}{\partial \gamma_{j,r} \partial \gamma_{j,r'}} = -\sum_{i=1}^{m} \sum_{t=1}^{n} \sum_{k=0}^{2} \Big\{ u_{i,t,r} u_{i,t,r'} \pi_{k,j}(x_{i,t}) \Big(\delta_{j,j'} - \pi_{k,j'}(x_{i,t}) \Big) y_{i,t-1,k} \Big\},$$

the lower right $6w \times 6w$-matrix $\partial^2 \ell(\beta)/\partial b \partial b^{'}$ has elements

$$\frac{\partial^2 \ell^{(m,n)}(\beta)}{\partial b_{k,j,s} b_{k',j',s'}} = -\sum_{i=1}^{m} \sum_{t=1}^{n} \Big\{ z_{i,t,s} z_{i,t,s'} \pi_{k,j}(x_{i,t}) \Big(\delta_{j,j'} - \pi_{k,j'}(x_{i,t}) \Big) y_{i,t-1,k} \Big\},$$

if $k = k'$, zero else.

The two remaining $2q \times 6w$-matrices are the transposed of each other and has elements

$$\frac{\partial^2 \ell^{(m,n)}(\beta)}{\partial \gamma_{j,r} \partial \beta_{k,j',s'}} = -\sum_{t=1}^{m} \sum_{i=1}^{n} \Big\{ u_{i,t,r} z_{i,t,s'} \pi_{k,j}(x_{i,t}) \Big(\delta_{j,j'} - \pi_{k,j'}(x_{i,t}) \Big) y_{i,t-1,k} \Big\}.$$

The Fisher information matrix is the negative mean of the Hessian matrix. Thus the Fisher information matrix per observation conditioned on the covariates is

$$J_{m \cdot n} = -\frac{1}{m \cdot n} \mathrm{E}\ H(\beta),$$

which is a block matrix

$$J_{m \cdot n} = \frac{1}{m \cdot n} \begin{pmatrix} J_{\gamma} & J_{\gamma b_0} & J_{\gamma b_1} & J_{\gamma b_2} \\ J_{b_0 \gamma} & J_{b_0} & 0 & 0 \\ J_{b_1 \gamma} & 0 & J_{b_1} & 0 \\ J_{b_2 \gamma} & 0 & 0 & J_{b_2} \end{pmatrix},$$

where

$$J_{\gamma} = \sum_{t=1}^{m} \sum_{i=1}^{n} \sum_{k=0}^{2} \begin{pmatrix} u_{i,t} u^{\mathrm{t}}_{i,t} \pi_{k,0}(1 - \pi_{k,0}) & -u_{i,t} u^{\mathrm{t}}_{i,t} \pi_{k,0} \pi_{k,1} \\ -u_{i,t} u^{\mathrm{t}}_{i,t} \pi_{k,0} \pi_{k,1} & u_{i,t} u^{\mathrm{t}}_{i,t} \pi_{k,0}(1 - \pi_{k,0}) \end{pmatrix} \mathrm{E}\ y_{i,t-1,k},$$

$$J_{b_k} = \sum_{t=1}^{m} \sum_{i=1}^{n} \begin{pmatrix} z_{i,t} z^{\mathrm{t}}_{i,t} \pi_{k,0}(1 - \pi_{k,0}) & -z_{i,t} z^{\mathrm{t}}_{i,t} \pi_{k,0} \pi_{k,1} \\ -z_{i,t} z^{\mathrm{t}}_{i,t} \pi_{k,0} \pi_{k,1} & z_{i,t} z^{\mathrm{t}}_{i,t} \pi_{k,0}(1 - \pi_{k,0}) \end{pmatrix} \mathrm{E}\ y_{i,t-1,k},$$

and

$$J_{\gamma b_k} = \sum_{t=1}^{m} \sum_{i=1}^{n} \begin{pmatrix} u_{i,t} z^{\mathrm{t}}_{i,t} \pi_{k,0}(1 - \pi_{k,0}) & -u_{i,t} z^{\mathrm{t}}_{i,t} \pi_{k,0} \pi_{k,1} \\ -u_{i,t} z^{\mathrm{t}}_{i,t} \pi_{k,0} \pi_{k,1} & u_{i,t} z^{\mathrm{t}}_{i,t} \pi_{k,0}(1 - \pi_{k,0}) \end{pmatrix} \mathrm{E}\ y_{i,t-1,k},$$

where the probability π_{kj} is an abbreviation for $\pi_{kj}(x_t)$.

Notice that in the widest class of models with only Markov dependent effects, the structure of the Fisher information matrix will simplify. It will then be given by the block diagonal matrix

$$J_{m\cdot n} = \frac{1}{m \cdot n}\begin{pmatrix} J_{b_0} & 0 & 0 \\ 0 & J_{b_1} & 0 \\ 0 & 0 & J_{b_2}. \end{pmatrix}$$

On the other side of the spectrum, the Fisher information matrix of the model with only Markov independent effects, that is the standard multinomial model, is simply $J_{m\cdot n} = J_\gamma$.

2.4 Ergodicity

Under the regularity assumption of a decent covariate distribution made in Section 2.3 the Markov chain $\{y_{i,t}\}$ generated by the dynamic multinomial logit model with parameter value β will be a *strongly ergodic* Markov chain.

Recall from Section 1.1 that *weak ergodicity* means that in the limit the state of a Markov chain $\{y_{i,t}\}$ is independent of the starting position of the chain. For each $j = 0, 1, 2$ there exists a probability function $\pi_j(t)$ such that for each $k = 0, 1, 2$ it is the case that

$$\lim_{s\to\infty} P_{kj}^{(s)}(t) = \lim_{s\to\infty} P_{kj}^{(s)}(x_{t+s}, \ldots x_{t+1})) = \pi_j(t),$$

To show the *weak ergodicity* of $\{y_{i,t}\}$, notice that the compactness of the space of covariates Γ implies that there for each β exists a constant c for which $0 < c < 1$. For every $x_{i,t} \in \Gamma$, and every $k, j = 0, 1, 2$ it is then the case that

$$c < \pi_{k,j}(x_{i,t}) < 1 - c.$$

This implies that the chain $\{y_{i,t}\}$ at any time $t-1$ and regardless of the state at $t-1$ has a non-diminishing probability of making a transition to all other states at time t.

In accordance with Dobrushin (1956) we define the *ergodic coefficient* of the inhomogeneous Markov chain (See also Sethuraman and Varadhan (2005) on this point). Given a covariate vector $x_{i,t} \in \Gamma$ let

$$\kappa(x_{i,t}) = \sup_{k,j,j'}\left|\pi_{k,j}(x_{i,t}) - \pi_{k,j'}(x_{i,t})\right|, \qquad k, j, j' = 0, 1, 2.$$

We define the *ergodic coefficient* of the Markov chain $\{y_{i,t}\}$ to be

$$\kappa = \sup_{x_{i,t} \in \Gamma} \kappa(x_{i,t})$$

which exists due to the boundedness of Γ assumed in Section 2.3.

Lemma 2.4.1. *Consider an inhomogeneous Markov chain $\{y_{i,t}\}$ with contraction coefficient κ. For all $k, k', j, j' = 0, 1, 2$, for all $s \geq 0$ and for all $t > s$ it is the case that*

$$\left| P_{kj}^{(s)}(t) - P_{k'j}^{(s)}(t) \right| < \kappa^s.$$

Proof. We will prove the lemma inductively. The case for $s = 0$ is trivial and the case for $s = 1$ is equivalent to the definition of the contraction coefficient. Assume therefore that $s > 1$, and that the lemma holds for $s - 1$. Writing $P_{kj}^{(s)} = P_{kj}^{(s)}(t)$ and $\pi_{kj} = \pi_{kj}(x_{t+s})$ for readability, we have for all given covariates $x_0, x_1, \ldots, x_{t+s}$ that

$$\begin{aligned}
\left| P_{kj}^{(s)} - P_{k'j}^{(s)} \right| &= \Big| \pi_{0j} P_{k0}^{(s-1)} + \pi_{1j} P_{k1}^{(s-1)} + \pi_{2j} P_{k2}^{(s-1)} \\
&\qquad\qquad - \pi_{0j} P_{k'0}^{(s-1)} - \pi_{1j} P_{k'1}^{(s-1)} - \pi_{2j} P_{k'2}^{(s-1)} \Big| \\
&= \Big| \pi_{0j} P_{k0}^{(s-1)} + \pi_{1j} P_{k1}^{(s-1)} + \pi_{2j} (1 - P_{k0}^{(s-1)} - P_{k1}^{(s-1)}) \\
&\qquad - \pi_{0j} P_{k'0}^{(s-1)} - \pi_{1j} P_{k'1}^{(s-1)} - \pi_{2j} (1 - P_{k'0}^{(s-1)} - P_{k'1}^{(s-1)}) \Big| \\
&= \Big| (\pi_{0j} - \pi_{2j}) P_{k0}^{(s-1)} + (\pi_{1j} - \pi_{2j}) P_{k1}^{(s-1)} + \pi_{2j} \\
&\qquad\qquad - (\pi_{0j} - \pi_{2j}) P_{k'0}^{(s-1)} - (\pi_{1j} - \pi_{2j}) P_{k'1}^{(s-1)} - \pi_{2j} \Big| \\
&= \left| (\pi_{0j} - \pi_{2j})(P_{k0}^{(s-1)} - P_{k'0}^{(s-1)}) + (\pi_{1j} - \pi_{2j})(P_{k1}^{(s-1)} - P_{k'1}^{(s-1)}) \right| \\
&< \kappa \left| (P_{k0}^{(s-1)} - P_{k'0}^{(s-1)}) + (P_{k1}^{(s-1)} - P_{k'1}^{(s-1)}) \right| \\
&= \kappa \left| P_{k'2}^{(s-1)} - P_{k2}^{(s-1)} \right| < \kappa \kappa^{s-1} = \kappa^s,
\end{aligned}$$

which by induction proves the lemma. ■

Now, we have that that $0 < \kappa < 1 - c$. This implies that there for given covariate values $x_{i,0}, x_{i,1}, \ldots, x_{i,t+s}$ exists a function $\pi_j(t)$ such that

$$\pi_j(t) = \lim_{s\to\infty} P_{kj}^{(s)}(t) = \lim_{s\to\infty} P_{k'j}^{(s)}(t) \qquad k, k' = 0, 1, 2,$$

which means that the *inhomogeneous* Markov chain $\{y_{i,t}\}$ conditioned on the covariate values is a *weakly ergodic* Markov chain.

It follows immediately that the chain also is *strongly ergodic*. For any $s > 0$ and for $k = 0, 1, 2$ write

$$\pi_j(0) = \lim_{t\to\infty} P_{kj}^{t}(0) = \sum_{r=0}^{2} P_{kr}^{s}(0) \lim_{t\to\infty} P_{rj}^{(t-s)}(s) = \pi_j(s) \sum_{r=0}^{2} P_{kr}^{s}(0) = \pi_j(s)$$

which implies that there exists a constant π_j such that $\pi_j = \pi_j(t)$ for all $t \geq 0$.

This ergodic property of the *inhomogeneous* chain will be of importance in the demonstrations of the large sample properties of the dynamic *multinomial logit* model in the next chapter. The probability $P_{kj}^{(s)}(t) - P_{2j}^{(s)}(t)$ will play an important role in these demonstrations. We define therefore the variable

$$\Phi_{kj}^{(s)}(t) = P_{kj}^{(s)}(t) - P_{2j}^{(s)}(t),$$

where $k = 0, 1$.

CHAPTER 3

Large Sample Theory under Misspecification

The *Focused Information Criterion* is based on the asymptotic mean squared errors of focus parameter estimators. Hence, in this chapter I work out the asymptotic properties of the dynamic multinomial logit models described in Chapter 2. This I will do both for correctly specified models as well as for misspecified models.

I assume that the true distribution of responses y_t conditioned on covariates x_t and past observations y_{t-1}, is a dynamic multinomial logit model with only Markov dependent effects. That is, I assume that the true conditional distribution of $y_t|y_{t-1}$ is on the form of (2.2). In concordance with Cunen, Walløe, and Hjort (2019) I call this true model the *wide* model. Operators and model parameters under this *wide* model are denoted with the subscript *wide*, like this: β_{wide} and $\mathrm{E}_{\mathrm{wide}}$. The true value of β_{wide} is denoted β_{true}.

Other dynamic multinomial logit models on the form of (2.3) are called candidate models. These models are different from the true *wide* model and will therefore involve misspecification. Parameters and covariates of these misspecified models are denoted with a subscript M, like this: β_M, $x_{M,i,t}$ and $\pi_M(x_{M,i,t})$.

In Section 3.1 I show that maximum likelihood estimators are consistent when data are Markov chains distributed according to the *wide* model. Consistency of ML estimators is the case both for the correctly specified *wide* model as well as for misspecified candidate models. In section Section 3.2 I define random score vectors of the dynamic multinomial logit model (2.3) and I find expressions for the covariance matrices of these random score vectors. I also show that these covariance matrices have non-stochastic limits. In Section 3.3 I show that the random score vectors of the *wide* model and the random score vector of a candidate model have approximate joint normal distribution. This joint normal distribution will allow me to find in Section 3.4 an expression for the approximate joint distribution of the maximum likelihood estimators of the *wide* model and the candidate model. In Section 3.5 I show that these results also hold when data from multiple independent Markov chains are considered. In

Section 3.6 I show that standard procedures for testing and confidence intervals are applicable to dynamic multinomial logit models on the form (2.3), even in the case of model misspecification under the *wide* model.

To facilitate the discussion in this chapter, I restrict the discussion in crefthreeone to Section 3.4 to the situation where we have observations from only one Markov chain $\{y_{i,t}\}$. In these sections I drop the subscript i. I denote observations $y_{i,t,j}$ by $y_{t,j}$ and covariates $x_{i,t}$ by x_t. After having established large sample theory for data from one Markov chain, I consider the situation with data from multiple independent chains again in Section 3.5.

3.1 Consistency of Maximum Likelihood Estimators

In this section I show that the maximum likelihood estimator $\hat{\beta}$ of the correctly specified *wide* model tends to the true parameter value β_{true}. I also show that the maximum likelihood estimator $\hat{\beta}_M$ of any misspecified candidate model under the *wide* model tends to the least false parameter value $\beta_{M,0}$. This means that maximum likelihood estimators of dynamic multinomial logit models are consistent, both in case of a correctly specified *wide* model and in case of misspecified models under the *wide* model.

To show this important large sample result, we first need to ensure that the sum of correlations between the data in a Markov chain generated by the *wide* model does not grow to infinity as the length n of the chain increases.

Define for any $t \in \mathbb{N}$ and $k, j = 0, 1, 2$ the functions $f_{kj}^{(t)} = f_{kj}(x_t)$,$f_{kj}^{*(t)} = f_{kj}^{*}(x_t)$, $g_{kj}^{(t)} = g_{kj}(x_t)$ and $g_{kj}^{*(t)} = g_{kj}(x_t)$. Let all these functions be uniformly bounded. That is, for any t and $k, j = 0, 1, 2$ there exists an $M > 0$ such that

$$\begin{aligned} |f_{kj}^{(t)}| &< M, & |g_{kj}^{(t)}| &< M, \\ |f_{kj}^{*(t)}| &< M, & |g_{kj}^{*(t)}| &< M. \end{aligned}$$

Define also composite functions

$$\begin{aligned} \psi_{t,k,j} = \psi(x_t, y_t, y_{t-1}) &= \left(f_{kj}^{(t)} y_{tj} + g_{kj}^{(t)} \right) y_{t-1,k}, \qquad (3.1) \\ \psi_{t,k,j}^{*} = \psi^{*}(x_t, y_t, y_{t-1}) &= \left(f_{kj}^{*(t)} y_{tj} + g_{kj}^{*(t)} \right) y_{t-1,k}. \end{aligned}$$

These functions will by the uniform boundedness of the functions $f_{kj}^{(t)}, f_{kj}^{*(t)}, g_{kj}^{(t)}$ and $g_{kj}^{*(t)}$ also be uniformly bounded.

The following theorem will be of central importance to the demonstrations in this chapter:

Lemma 3.1.1. *Let covariate vectors $x_0, x_1, \ldots, x_t$ be generated by some unknown covariate distribution C(x) in accordance with the assumptions in Section 2.3.*

Given these covariate values, let $\{y_{i,t}\}$ be a Markov chain generated by the wide model. Define functions $\psi_{t,k,j}$ and $\psi^*_{t,k,j}$ as above. It is then the case for all $t > 0$ that

$$\lim_{t\to\infty} \sum_{s=0}^{t-1} \mathrm{Cov}_{wide} \Big\{ \psi_{t,k,j}, \psi_{t-s,k',j'} \Big\} < \infty.$$

The proof of this important lemma is given in Appendix A.1.

Express the conditional distribution of the true *wide* model by

$$f_{\mathrm{wide}}(y_t \mid \beta_{\mathrm{true}}, x, y_{t-1}),$$

and express the conditional distribution of a candidate model by

$$f_M(y_t \mid \beta_M, x, y_{t-1}).$$

The candidate model involves misspecification, it will therefore be at a distance from the true *wide* model. Distance between parametric models may be measured by the Kullback-Leibler distance (See Claeskens and Hjort (2008b) for an introduction). Conditioned on covariate vector $x \in \Gamma$, the Kullback-Leibler distance from the candidate model to the true *wide* model is

$$\begin{aligned}
&\mathrm{KL}_x\big(f_{\mathrm{wide}}(\cdot \mid \beta_{\mathrm{true}}, x, \cdot), f_M(\cdot|\beta_M, x, \cdot)\big) \\
&\qquad = \sum_{j=0}^{2}\sum_{k=0}^{2} f_{\mathrm{wide}}(y_t|\beta_{\mathrm{true}}, x, y_{t-1}) \log \frac{f_{\mathrm{wide}}(y_t|\beta_{\mathrm{true}}, x, y_{t-1})}{f_M(y_t|\beta_M, x, y_{t-1})} \\
&\qquad = \mathrm{E}_{\mathrm{wide}} \log f_{\mathrm{wide}}(y_t|\beta_{\mathrm{true}}, x, y_{t-1}) - \mathrm{E}_{\mathrm{wide}} \log f_M(y_t|\beta_M, x, y_{t-1}).
\end{aligned}$$

The subscript x here denotes that the distance is conditioned on a covariate value x.

An overall Kullback-Leibler distance may be obtained by integration over the compact space Γ of covariates such that

$$\begin{aligned}
&\mathrm{KL}\big(f_{\mathrm{wide}}(\cdot \mid \beta_{\mathrm{true}}, \cdot), f_M(\cdot|\beta_M, \cdot)\big) \\
&\qquad = \int \sum_{j=0}^{2}\sum_{k=0}^{2} f_{\mathrm{wide}}(y_t|\beta_{\mathrm{true}}, x, y_{t-1}) \log \frac{f_{\mathrm{wide}}(y_t|\beta_{\mathrm{true}}, x, y_{t-1})}{f_M(y_t|\beta_M, x, y_{t-1})} \,\mathrm{d}C(x) \\
&\qquad = \int \mathrm{E}_{\mathrm{wide}} \log f_{\mathrm{wide}}(y_t|\beta_{\mathrm{true}}, x, y_{t-1}) \,\mathrm{d}C(x) \\
&\qquad\qquad\qquad - \int \mathrm{E}_{\mathrm{wide}} \log f_M(y_t|\beta_M, x, y_{t-1}) \,\mathrm{d}C(x). \qquad (3.2)
\end{aligned}$$

The *least false* parameter value $\beta_{M,0}$ of the candidate model will be the value of β_M that minimizes the Kullback-Leibler distance to the true *wide* model.

As the *true* model parameter β_{true} is fixed, close inspection of the expression of the Kullback-Leibler distance reveals that this least false parameter value will be the value of β_M that maximizes the last term integral the expression (3.2). Consequently, the least false parameter of candidate model M will be the parameter value

$$\beta_{M,0} = \arg\max_{\beta_M}\biggl\{\int \mathrm{E}_{\text{wide}}\, \log f_M(y|\beta_M, x)\, \mathrm{d}C(x)\biggr\}. \tag{3.3}$$

As the log-likelihood function of the dynamic multinomial logit model is *concave* this least false parameter value will be unique.

In the following theorem we show that the maximum likelihood estimator $\hat{\beta}_{M,n}$ of the candidate model tends to the least false parameter value $\beta_{M,0}$

Theorem 3.1.2. *Let covariate vectors $x_0, x_1, \ldots, x_t$ be generated by some unknown covariate distribution C(x) in accordance with the assumptions in Section 2.3. Given these covariate values, let $\{y_{i,t}\}$ be a Markov chain generated by the wide model. Fit a candidate model M on the form of* (2.3) *to the data. Denote the maximum likelihood estimate of this candidate model by $\hat{\beta}_M$. Let $\beta_{M,0}$ be the least false parameter value of the candidate model. It is then the case that*

$$\hat{\beta}_{M,n} \rightarrow_p \beta_{M,0},$$

Proof. The variance of the log-likelihood function conditioned on covariates $x_0, x_1, \ldots, x_n$ is given by

$$\begin{aligned}
\mathrm{Var}_{\text{wide}}\; \ell^{(n)}(\beta_M) &= \mathrm{Cov}_{\text{wide}} \sum_{t=1}^{n} \ell_t^{(n)}(\beta_M) \\
&= 2\sum_{t=1}^{n}\sum_{s=0}^{t-1} \mathrm{Cov}_{\text{wide}}\left\{\ell_t^{(n)}(\beta_M), \ell_{t-s}^{(n)}(\beta_M)\right\} \\
&\qquad\qquad - \sum_{t=1}^{n} \mathrm{Var}_{\text{wide}}\; \ell_t^{(n)}(\beta_M) \\
&< 2\sum_{t=1}^{n}\sum_{s=0}^{t-1} \mathrm{Cov}_{\text{wide}}\left\{\ell_t^{(n)}(\beta_M), \ell_{t-s}^{(n)}(\beta_M)\right\}.
\end{aligned}$$

From Section 3.2 we have that

$$\begin{aligned}
\ell_t^{(n)}(\beta_M) = \sum_{k=0}^{2}\biggl\{\biggl(&\sum_{j=0}^{1} y_{t,j}\bigl(u_t^{\mathrm{t}}\gamma_j + z_t^{\mathrm{t}} b_{k,j}\bigr) \\
&- \log\biggl(1 + \sum_{h=0}^{1}\exp\bigl(u_t^{\mathrm{t}}\gamma_h + z_t^{\mathrm{t}} b_{k,h}\bigr)\biggr)\biggr) y_{t-1,k}\biggr\}.
\end{aligned}$$

Each term in this function is on the form of the general function ψ_{tkj} defined in Equation (3.1). Define

$$\begin{aligned}
\psi_{tk0} &= y_{t,0}(u_t^{\mathrm{t}}\gamma_0 + z_t^{\mathrm{t}} b_{k,0}) y_{t-1,k},\\
\psi_{tk1} &= y_{t,1}(u_t^{\mathrm{t}}\gamma_1 + z_t^{\mathrm{t}} b_{k,1}) y_{t-1,k},\\
\psi_{tk2} &= -\log\biggl(1 + \sum_{h=0}^{1} \exp\bigl(u_t^{\mathrm{t}}\gamma_h + z_t^{\mathrm{t}} b_{k,h}\bigr)\biggr) y_{t-1,k},
\end{aligned}$$

and write

$$\mathrm{Cov}_{\mathrm{wide}}\Bigl\{\ell_t(\beta_M), \ell_{t-s}(\beta_M)\Bigr\} = \sum_{r,r'=0}^{2}\sum_{k,k'=0}^{2} \mathrm{Cov}_{\mathrm{wide}}\Bigl\{\psi_{t,k,r}, \psi_{t-s,k',r'}\Bigr\}.$$

It follows from Lemma 3.1.1 that for all $t > 0$

$$\begin{aligned}
&\sum_{s=0}^{t-1} \mathrm{Cov}_{\mathrm{wide}}\Bigl\{\ell_t^{(n)}(\beta_M), \ell_{t-s}^{(n)}(\beta_M)\Bigr\}\\
&\qquad = \sum_{r,r'=0}^{2}\sum_{k,k'=0}^{2}\sum_{s=0}^{t-1} \mathrm{Cov}_{\mathrm{wide}}\Bigl\{\psi_{t,k,r}, \psi_{t-s,k',r'}\Bigr\} < \infty.
\end{aligned}$$

As $\psi_{t,k,r}$ is uniformly bounded, there exists then a constant $G > 0$ such that for all $t > 0$ it is the case that

$$\sum_{s=0}^{t-1} \mathrm{Cov}_{\mathrm{wide}}\Bigl\{\ell_t^{(n)}(\beta_M), \ell_{t-s}^{(n)}(\beta_M)\Bigr\} < G.$$

It follows then that

$$\mathrm{Var}_{\mathrm{wide}}\ \ell^{(n)}(\beta_M) < 2\sum_{t=1}^{n}\sum_{s=0}^{t-1} \mathrm{Cov}_{\mathrm{wide}}\Bigl\{\ell_t^{(n)}(\beta_M), \ell_{t-s}^{(n)}(\beta_M)\Bigr\} < 2nG.$$

Looking now at the variance of the likelihood per observation conditioned on the covariates $x_0, x_1, \ldots, x_n$, it is easy to see that this goes to zero as the number of observations n from the Markov chain grows:

$$\mathrm{Var}_{\mathrm{wide}}\Bigl\{\frac{1}{n}\ell^{(n)}(\beta_M)\Bigr\} = \frac{1}{n^2}\,\mathrm{Var}_{\mathrm{wide}}\ \ell^{(n)}(\beta_M) < \frac{2G}{n} \to 0.$$

This implies that we may write

$$\begin{aligned}\frac{1}{n}\ell^{(n)}(\beta_M) &= \frac{1}{n}\,\mathrm{E}_{\text{wide}}\;\ell^{(n)}(\beta_M) + o_p(1)\\ &= \frac{1}{n}\sum_{t=1}^{n}\mathrm{E}_{\text{wide}}\;\ell_t^{(n)}(\beta_M) + o_p(1).\end{aligned}$$

where $o_p(1)$ here denotes a random term that converges to zero in probability.

Remembering that $\ell_t^{(n)}(\beta_M)$ is an abbreviation for $\ell_t^{(n)}(\beta_M|x_t)$, we get from the regularity assumptions on the covariate distribution $C(x)$ in Section 2.3 that

$$\frac{1}{n}\sum_{t=1}^{n}\mathrm{E}_{\text{wide}}\;\ell_t^{(n)}(\beta_M) = \int \mathrm{E}_{\text{wide}}\;\log f_M(y_t|\beta_M, x_{t-1})\,\mathrm{d}C(x) + o_p(1)$$

which again implies that

$$\frac{1}{n}\ell^{(n)}(\beta_M) \to_p \int \mathrm{E}_{\text{wide}}\;\log f_M(y_t|\beta_M, x, y_{t-1})\,\mathrm{d}C(x)$$

From Equation (3.3) we know that the least false parameter value $\beta_{M,0}$ of the candidate model is defined as the parameter value of β_M that maximizes this limit. Thus, the maximum likelihood estimator tends to the least false parameter value $\beta_{M,0}$, which is what we set out to prove. ■

From Section 2.2 we know that the *wide* model is a special case of the more general model (2.3). The result, therefore, holds also for the maximum likelihood estimator $\hat{\beta}_{\text{wide}}$ of the *wide* model. As the *wide* model is correctly specified, the least false parameter of this model will be the true parameter value β_{true}. It follows then from Theorem 3.1.2 that

$$\hat{\beta}_{\text{wide}} \to_p \beta_{\text{true}}.$$

The maximum likelihood estimator of the true *wide* model tend to the true parameter value of the data-generating mechanism.

Thus maximum likelihood estimators of the dynamic multinomial logit models are *consistent* both in the case of the correctly specified *wide* model and misspecified candidate models.

3.2 Variance of Random Score Vectors

The conditional score function of the *wide* model is given by the $6\,p \times 1$ vector

$$u(y_t|\beta, x, y_{t-1}) = \frac{\partial}{\partial \beta} \log f_{\text{wide}}(y_t|\beta, x_t, y_{t-1})$$

Examining the partial derivative of the log-likelihood function given in Section 2.2, we see that under the true model

$$\mathrm{E}_{\text{wide}}\ u(y_t|\beta_{\text{true}}, x_t, y_{t-1}) = 0$$

The conditional score functions of a candidate model is given by the $(2\,q+6\,w)\times 1$ vector

$$u_M(y_t|\beta_M, x_t, y_{t-1}) = \frac{\partial}{\partial \beta_M} \log f_M(y_t|\beta_M, x_t, y_{t-1})$$

Let $\beta_{M,0,n}$ be the least false parameter of the candidate model conditioned on the covariates $x_0, x_1, \dots x_n$. Since the candidate model is misspecified, it is not the case that $\mathrm{E}_{\text{wide}}\ u_M(y_t|\beta_{M,0,n}, x_t, y_{t-1}) = 0$. We have instead that

$$\begin{aligned}\frac{1}{n}\sum_{t=1}^{n} &\mathrm{E}_{\text{wide}}\ u(y|\beta_{M,0,n}, x_t, y_{t-1}) \\ &= \frac{1}{n}\sum_{t=1}^{n} \mathrm{E}_{\text{wide}}\ \frac{\partial}{\partial \beta_M} \log f(y_t|\beta_{M,0,n}, x_t, y_{t-1}) \to_p 0. \qquad (3.4)\end{aligned}$$

This is the case because we from the assumptions in Section 3.3 have that

$$\frac{1}{n}\sum_{t=1}^{n} \mathrm{E}_{\text{wide}}\ \log f(y_t|\beta_{M,0,n}, x_t, y_{t-1}) \to_p \int \mathrm{E}_{\text{wide}}\ \log f(y_t|\beta_{M,0})\, \mathrm{d}C(x),$$

and the definition of the least false parameter $\beta_{M,0}$ in (3.3) this limit is global maximum point.

Define now random score vectors U_n and $U_{M,n}$, where U_n is the $6\,p \times 1$ vector function

$$U_n = \frac{1}{\sqrt{n}} \frac{\partial \ell(\beta_{\text{true}})}{\partial \beta} = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial \ell_t(\beta_{\text{true}})}{\partial \beta} = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} u(y_t|\beta_{\text{true}}, x_t, y_{t-1}),$$

and where $U_{M,n}$ is the $(2\,q + 6\,w) \times 1$ vector function

$$\begin{aligned}U_{M,n} = \frac{1}{\sqrt{n}} \frac{\partial \ell^{(n)}(\beta_{M,0,n})}{\partial \beta_{M,n}} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \ell_i^{(n)}(\beta_{M,0,n})}{\partial \beta_{M,n}} \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_M(y_t|\beta_{M,0,n}, x_t, y_{t-1}).\end{aligned}$$

Define further matrices

$$\begin{aligned}
J_n &= \mathrm{Var}_{\mathrm{wide}}\ U_n, \\
K_{M,n} &= \mathrm{Var}_{\mathrm{wide}}\ U_{M,n}, \\
C_{M,n} &= \mathrm{Cov}_{\mathrm{wide}} \Big\{ U_n, U_{M,n} \Big\}, \\
J_{M,n} &= -\frac{1}{n} \sum_{t=1}^{n} \mathrm{E}_{\mathrm{wide}}\ H_M(\beta_{M,0,n}).
\end{aligned}$$

The last matrix is the Fisher information matrix of the candidate model. The expression of this matrix is given in Section 2.3.

Let us find an expression for J_n. The *wide* model random score vector U_n may be written as

$$U_n = \begin{pmatrix} \upsilon_{00} \\ \upsilon_{01} \\ \upsilon_{10} \\ \upsilon_{11} \\ \upsilon_{20} \\ \upsilon_{21} \end{pmatrix},$$

where $\upsilon_{kj} = \partial \ell^{(n)}(\beta_{\mathrm{true}})/\partial \beta_{kj}$.

Write

$$\upsilon_{kj} = \sum_{t=1}^{n} \upsilon_{kjt}, \qquad k = 0, 1, 2, \quad j = 0, 1,$$

where

$$\upsilon_{kjt} = \frac{1}{\sqrt{n}} \Big(y_t - \pi_{kj}(x_t) \Big) x_t y_{t-1,k}.$$

Define for each $t \le n$ and each $k, j = 0, 1, 2$, the functions $f_{kj}^{(t)} = \frac{1}{\sqrt{n}} x_t$ and $g_{kj}^{(t)} = \frac{1}{\sqrt{n}} \pi_{kj}(x_t)$. These functions are uniformly bounded by 1. Thus υ_{kjt} is on the form of the general function ψ_{tkj} defined in (3.1), and υ_{kj} is a sum of such functions.

Define for all $t \le n$ the quantity

$$\lambda_{kj}^{(t)} = \mathrm{E}_{\mathrm{wide}} \left\{ \frac{1}{\sqrt{n}} \Big(y_t - \pi_{kj}(x_t) \Big) x_t y_{t-1,k} \Big| y_{t-1} \right\} = 0.$$

It then follows from (A.1) in Appendix A.1 that

$$\mathrm{Cov}_{\mathrm{wide}}\left\{\upsilon_{kj}, \upsilon_{k'j}\right\} = \sum_{t=1}^{n} \mathrm{Cov}_{\mathrm{wide}}\left\{\upsilon_{kj,t}, \upsilon_{k'j,t}\right\},$$

and that for $t \leq n$

$$\begin{aligned}
&\mathrm{Cov}_{\mathrm{wide}}\left\{\upsilon_{kj,t}, \upsilon_{k'j,t}\right\} \\
&\quad = \frac{1}{n}\sum_{t=1}^{n} x_t x_t^{\mathrm{t}}\, \mathrm{Cov}_{\mathrm{wide}}\left\{y_{t,j}y_{t,j'} \mid y_{t-1,k'=1}\right\}\delta_{kk'}\, \mathrm{E}_{\mathrm{wide}}\ y_{t-1,k'} \\
&\quad = \begin{cases} \frac{1}{n}\sum_{t=1}^{n} x_t x_t^{\mathrm{t}} \pi_{kj}(x_t)\Big(\delta_{jj'} - \pi_{kj}(x_t)\Big)\, \mathrm{E}_{\mathrm{wide}}\ y_{t-1,k} & \text{if } k = k' \\ 0 & \text{if } k \neq k'. \end{cases}
\end{aligned}$$

Hence

$$\mathrm{Var}_{\mathrm{wide}}\ U_n = \frac{1}{n}\sum_{t=1}^{n} \begin{pmatrix} J_{\beta_0}(x_t) & 0 & 0 \\ 0 & J_{\beta_1}(x_t) & 0 \\ 0 & 0 & J_{\beta_2}(x_t) \end{pmatrix},$$

where for $k = 0, 1, 2$

$$J_{\beta_k}(x_t) = \begin{pmatrix} x_t x_t^{\mathrm{t}} \pi_{k,0}(1-\pi_{k,0}) & -x_t x_t^{\mathrm{t}} \pi_{k,0}\pi_{k,1} \\ -x_t x_t^{\mathrm{t}} \pi_{k,0}\pi_{k,1} & x_t x_t^{\mathrm{t}} \pi_{k,1}(1-\pi_{k,1}) \end{pmatrix} \mathrm{E}_{\mathrm{wide}}\ y_{t-1,k}$$

abbreviating $\pi_{k,j}$ for $\pi_{k,j}(x_t)$.

Comparing this expression with the Fisher Information matrix J_n described in Section 2.3, we see that

$$\mathrm{Var}_{\mathrm{wide}}\ U_n = -\frac{1}{n}\, \mathrm{E}_{\mathrm{wide}}\ \ H(\beta_{\mathrm{true}})$$

which is the result we would expect from Bartlett's second identity.

With reference again to the proof of Lemma 3.1.1 in Appendix A.1, we may show in a parallel manner to the derivation of the expression of J_n that $C_{M,n}$ is a $6\,p \times (2\,q + 6\,w)$-matrix on the form

$$C_{M,n} = \frac{1}{n}\sum_{t=1}^{n} \begin{pmatrix} C_{M,\gamma,0}(x_t) & C_{M,b,0}(x_t) & 0 & 0 \\ C_{M,\gamma,1}(x_t) & 0 & C_{M,b,1}(x_t) & 0 \\ C_{M,\gamma,2}(x_t) & 0 & 0 & C_{M,b,2}(x_t) \end{pmatrix},$$

where for $k = 0, 1, 2$ the left blocks are $2\,p \times 2\,q$-matrices

$$C_{M,\gamma,k}(x_t) = \begin{pmatrix} x_t u_{M,t}^{\mathrm{t}} \pi_{k,0}(1-\pi_{k,0}) & -x_t u_{M,t}^{\mathrm{t}} \pi_{k,0}\pi_{k,1} \\ -x_t u_{M,t}^{\mathrm{t}} \pi_{k,0}\pi_{k,1} & x_t u_{M,t}^{\mathrm{t}} \pi_{k,0}(1-\pi_{k,0}) \end{pmatrix} \mathrm{E}_{\mathrm{wide}}\ y_{t-1,k},$$

and the other blocks are $2\,p \times 2\,w$-matrices

$$C_{M,b,k}(x_t) = \begin{pmatrix} x_t z_{M,t}^{\mathrm{t}} \pi_{k,0}(1-\pi_{k,0}) & -x_t z_{M,t}^{\mathrm{t}} \pi_{k,0}\pi_{k,1} \\ -x_t z_{M,t}^{\mathrm{t}} \pi_{k,0}\pi_{k,1} & x_t z_{M,t}^{\mathrm{t}} \pi_{k,0}(1-\pi_{k,0}) \end{pmatrix} \mathrm{E}_{\mathrm{wide}}\ y_{t-1,k}.$$

For readability, we have abbreviated $\pi_{kj}(x_t) = \pi_{kj}$.

Finally, we find an expression for the variance of the candidate random score function, $\mathrm{Var}_{\mathrm{wide}}\ U_{M,n}$. The expression for this variance matrix will be much more complicated than the rather orderly matrices J_n, $J_{M,n}$ and $C_{M,n}$. The reason for this is the misspecification of the candidate model. Because the candidate model is not the correct model, we need to account for correlation between every pair of observations in the chain. This will complicate matters considerably.

The candidate random score vector $U_{M,n}$ may be written as

$$U_{M,n} = \begin{pmatrix} \upsilon_{\gamma 0} \\ \upsilon_{\gamma 1} \\ \upsilon_{b00} \\ \upsilon_{b01} \\ \upsilon_{b10} \\ \upsilon_{b11} \\ \upsilon_{b20} \\ \upsilon_{b21} \end{pmatrix},$$

where

$$\upsilon_{\gamma j} = \frac{\partial \ell^{(n)}(\beta_{M,0,n})}{\partial \gamma_j},$$

and

$$\upsilon_{bkj} = \frac{\partial \ell^{(n)}(\beta_{M,0,n})}{\partial b_{kj}}.$$

Write for $j = 0, 1$

$$\upsilon_{\gamma j} = \sum_{t=1}^{n} \upsilon_{\gamma j t},$$

where

$$\upsilon_{\gamma jt} = \sum_{k=0}^{2} \frac{1}{\sqrt{n}} \Big(y_{t,j} - \pi_{Mkj}(x_{M,t}) \Big) u_{M,t} y_{t-1,k} \qquad j = 0, 1.$$

Similarily, write for $k = 0, 1, 2$ and $j = 0, 1$

$$\upsilon_{bkj} = \sum_{t=1}^{n} \upsilon_{bkjt}$$

where

$$\upsilon_{bkjt} = \frac{1}{\sqrt{n}} \Big(y_{t,j} - \pi_{Mkj}(x_{M,t}) \Big) z_{M,t} y_{t-1,k} \qquad k = 0, 1, 2, \quad j = 0, 1.$$

Now, define $f_{kj}^{(t)} = \frac{1}{\sqrt{n}} d_{M,t}$ and $g_{kj}^{(t)} = \frac{1}{\sqrt{n}} \pi_{Mkj}(x_{M,t}) d_{M,t}$, where $d_{M,t}$ is either $u_{M,t}$ or $z_{M,t}$. From this we see that the elements in $U_{M,n}$ may be expressed as sums where the terms are on the form of the general function ψ_{tkj} in (3.1).

Let now

$$\phi_{kj}(x_t) = \pi_{k,j}(x_t) - \pi_{M,k,j}(x_{M,t}), \qquad k = 0, 1, 2 \quad j = 0, 1,$$

and define

$$\lambda_{kj}^{(t)} = \mathrm{E}_{\text{wide}} \left\{ \frac{1}{\sqrt{n}} \Big(y_t - \pi_{Mkj}(x_{M,t}) \Big) d_t y_{t-1,k} \middle| y_{t-1} \right\} = \frac{1}{\sqrt{n}} \phi_{kj}(x_{M,t}) d_t y_{t-1,k}.$$

We may then use results from Appendix A to find an expression for the $K_{M,n}$ matrix. Define

$$U_{M,n,t} = \frac{1}{\sqrt{n}} \frac{\partial \ell_{M,t}(\beta_{M,0,n})}{\partial \beta_M}$$

such that

$$U_{M,n} = \sum_{t=1}^{n} U_{M,n,t}$$

We then have that

$$
\begin{aligned}
K_{M,n} &= \text{Var}_{\text{wide}}\ U_{M,n} \\
&= \sum_{t=1}^{n} \text{Var}_{\text{wide}}\ U_{M,n,t} + \sum_{t=2}^{n}\sum_{s=1}^{t-1} \text{Cov}_{\text{wide}} \Big\{ U_{M,n,t}, U_{M,n,t-s} \Big\} \\
&\qquad\qquad + \sum_{t=2}^{n}\sum_{s=1}^{t-1} \text{Cov}_{\text{wide}} \Big\{ U_{M,n,t-s}, U_{M,n,t} \Big\}. \qquad (3.5)
\end{aligned}
$$

As all random vectors $U_{M,n,t}$ are on the general form of ψ_{tkj} in (3.1) we may use (A.4) in (A.1) to find expressions for each of these sums. Inserting the appropriate functions $f_{kj}^{(t)}$, $g_{kj}^{(t)}$ and $\lambda_{kj}^{(t)}$ in (A.4), we may write the first sum in (3.5) as

$$
\sum_{t=1}^{n} \text{Var}_{\text{wide}}\ U_{M,n,t} = \frac{1}{n} J^*_{M,n} + \frac{1}{n} V_n
$$

Expressions for the elements of these matrices are given in Appendix A.4. Here we emphasize only that J^*_M is *not* the same matrix as the Fisher Information matrix $J_{M,n}$ of the candidate model, although they look very similar. The $J^*_{M,n}$ uses the true probabilities of the correct *wide* model, whereas $J_{M,n}$ uses the probabilities of the candidate model.

By inserting appropriate functions $f_{kj}^{(t)}$, $g_{kj}^{(t)}$ and $\lambda_{kj}^{(t)}$ in (A.4) in Appendix A.1, we may write the second term in (3.5) as

$$
\frac{1}{n} \sum_{t=2}^{n}\sum_{s=1}^{t-1} \text{Cov}_{\text{wide}} \Big\{ U_{M,n,t}, U_{M,n,t-s} \Big\} = \frac{1}{n} Q_{M,n} + \frac{1}{n} W_{M,n},
$$

Expressions for the rather complicated matrices $Q_{M,n}$ and $W_{M,n}$ are given in Appendix A.4. Here we emphasize only that neither $Q_{M,n}$ or $W_{M,n}$ are symmetric. Considerable care must therefore be shown by the calculation of these matrices.

The last term of sums in (3.5) is the transposed of the middle term. Consequently, we get the following expression for the variance matrix $K_{M,n}$ of the candidate score function:

$$
K_{M,n} = \frac{1}{n} \Big\{ J^*_{M,n} + V_{M,n} + (W_{M,n} + W^t_{M,n}) + (Q_{M,n} + Q^t_{M,n}) \Big\} \qquad (3.6)
$$

Under the assumption of a proper covariate distribution in Section 2.4, the matrices J_n, $J_{M,n}$, $C_{M,n}$ and $K_{M,n}$ all converge asymptotically to well-defined limit matrices.

Theorem 3.2.1. *Let covariate vectors $x_0, x_1, \dots, x_t$ be generated by some unknown covariate distribution C(x) in accordance with the assumptions in Section 2.3. Given these covariate values, let $\{y_{i,t}\}$ be a Markov chain generated by the wide model. It is then the case that*

$$
\begin{aligned}
J_n &\to_p J, \\
J_{M,n} &\to_p J_{M,n}, \\
C_{M,n} &\to_p C_M, \\
K_{M,n} &\to_p K_M,
\end{aligned}
$$

where J, $J_{M,n}$, C_M and K_M are well-defined non-stochastic matrices.

Proof. We prove convergence in probability of $K_{M,n}$. The proofs of convergence in probability of J_M, $J_{M,n}$ and C_M will be equivalent.

From the above discussion, we know that the elements in $U_{M,n,t}$ are on the general form of the function $\psi_{t,k,j}$ in (3.1). Write therefore $\psi^{(a)}_{t,k,j}$ and $\psi^{(b)}_{t,k,j}$ for any two elements in the vector $U_{M,n,t}$. Let $k_{M,n}$ be any element in $K_{M,n}$. From (3.5) we may express $k_{M,n}$ as

$$
\begin{aligned}
k_{M,n} = \frac{1}{n}\sum_{t=1}^{n} \mathrm{Cov}_{\mathrm{wide}}\left\{\psi^{(a)}_{t,k,j}, \psi^{(b)}_{t,k',j'}\right\} + \frac{1}{n}\sum_{t=2}^{n}\sum_{s=1}^{t-1} \mathrm{Cov}_{\mathrm{wide}}\left\{\psi^{(a)}_{t,k,j}, \psi^{(b)}_{t-s,k',j'}\right\} \\
+ \frac{1}{n}\sum_{t=2}^{n}\sum_{s=1}^{t-1} \mathrm{Cov}_{\mathrm{wide}}\left\{\psi^{(a)}_{t-s,k,j}, \psi^{(b)}_{t,k',j'}\right\}.
\end{aligned}
$$

As $\psi^{(a)}_{t,k,j}$ and $\psi^{(b)}_{t,k,j}$ are functions on the general form of $\psi_{t,k,j}$, we may use the result from Lemma A.3.1. Thus for any $\epsilon > 0$ we may find a finite $N \in \mathbb{N}$ such that there are functions $f_i(x_t, \dots x_{t-N})$, $i = 1, 2, 3$, for which it is the case that for all series of covariates $x_0, x_1, \dots, x_t$ we have for all $t \geq 1$ that

$$
\lim_{t\to\infty} P\left(\left|\mathrm{Cov}_{\mathrm{wide}}\left\{\psi^{(a)}_{t,k,j}, \psi^{(b)}_{t,k',j'}\right\} - f_1(x_t, \dots, x_{t-N})\right| > \frac{\epsilon}{6}\right) = 0,
$$

and for all $t \geq 2$ that

$$
\begin{aligned}
\lim_{t\to\infty} P\left(\left|\sum_{s=1}^{t-1} \mathrm{Cov}_{\mathrm{wide}}\left\{\psi^{(a)}_{t,k,j}, \psi^{(b)}_{t-s,k',j'}\right\} - f_2(x_t, \dots, x_{t-N})\right| > \frac{\epsilon}{6}\right) = 0, \\
\lim_{t\to\infty} P\left(\left|\sum_{s=1}^{t-1} \mathrm{Cov}_{\mathrm{wide}}\left\{\psi^{(b)}_{t,k,j}, \psi^{(a)}_{t-s,k',j'}\right\} - f_3(x_t, \dots, x_{t-N})\right| > \frac{\epsilon}{6}\right) = 0.
\end{aligned}
$$

Defining the composite function $f(x_1, \dots x_{t-N}) = \sum_{i=1}^{3} f_i(x_t, \dots, x_{t-N})$, we get by repeated use of the triangle equality that

$$\lim_{n\to\infty} P\left(\left|k_{M,n} - \frac{1}{n}\sum_{t=2}^{n} f(x_1, \dots x_{t-N})\right| > \frac{1}{n}\sum_{t=1}^{n}\frac{\epsilon}{2} = \frac{\epsilon}{2}\right) = 0. \tag{3.7}$$

From the assumption on the covariate distribution $C(x)$ in Section 2.3 we have that

$$\frac{1}{n}\sum_{t=N}^{n} f(x_1, \dots x_{t-N}) \to \int f(x_t, \dots, x_{t-N})\,\mathrm{d}C_{\text{joint},,N}(x_N)$$

Writing for this limit $\kappa_{M,n} = \int f(x_t, \dots, x_{t-N})\,\mathrm{d}C_{\text{joint},,N}(x_N)$, we have then that

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{t=2}^{n} f(x_1, \dots x_{t-N}) - \kappa_{M,N}\right| > \frac{1}{n}\sum_{t=1}^{n}\frac{\epsilon}{2} = \frac{\epsilon}{2}\right) = 0. \tag{3.8}$$

Consequently, using the triangle equality results (3.7) and (3.8), we get that for each ϵ there exists a finite N and a constant $\kappa_{M,n}$ such that

$$\lim_{n\to\infty} P\left(\left|k_{M,n} - \kappa_{M,n}\right| > \epsilon\right) = 0.$$

For any two integers m, n such that $m > n$, it is now the case that for each ϵ there is an $N^* \in \mathbb{N}$ such that for all $m, n > N^*$

$$\begin{aligned} &P\left(|\kappa_{M,m} - \kappa_{M,n}| > \epsilon\right) \\ &\qquad \le \lim_{t\to\infty} P\left(\left|\kappa_{M,n} - k_{M,t}\right| > \frac{\epsilon}{2}\right) + \lim_{t\to\infty} P\left(\left|\kappa_{M,m} - k_{M,t}\right| > \frac{\epsilon}{2}\right) = 0. \end{aligned}$$

Thus, the series $\{\kappa_{M,n}\}$ is a Cauchy-series in $\mathbb{R}$. As $\kappa_{M,n} \in \mathbb{R}$ and $\mathbb{R}$ is a complete space, this implies that $\kappa_{M,n}$ is a convergent series converging to a constant κ_M. Thus

$$k_{M,n} \to_p \kappa_M.$$

As this result holds for all elements in $K_{M,n}$, the matrix $K_{M,n}$ converges to a non-stochastic matrix K_M, as is what we set out to prove.

■

3.3 Asymptotic Normality

In this section, I prove that the random score vector $U_{M,n}$ of the candidate model is normally distributed in the limit. As the *wide* model is a special case of the more general model (2.3), the result will follow also for the random score function U_n of the *wide* model. I will prove the normality of $U_{M,n}$ with the Martingale Central Limit theorem of Hall and Heyde (1980). The first task in this section is therefore to show that $U_{M,n}$ constitutes a martingale asymptotically.

Let $y_{0:k}$ denote every observation in the Markov chain up to time k inclusive. With Sethuraman and Varadhan (2005), define for $0 \leq k \leq n$ the random variables

$$Z_k^{(n)} = \sum_{t=k}^{n} \mathrm{E}_{\text{wide}} \left\{ \frac{\partial \ell_{t,M}^{(n)}(\beta_{M,0,n})}{\partial \beta_{M,n}} \middle| y_{0:k} \right\}.$$

A few values of these random variable are

$$\begin{aligned} Z_n^{(n)} &= \frac{\partial \ell_n^{(n)}(\beta_{M,0,n})}{\partial \beta_M} \\ Z_{n-1}^{(n)} &= \frac{\partial \ell_{n-1}^{(n)}(\beta_{M,0,n})}{\partial \beta_M} + \mathrm{E}_{\text{wide}} \left\{ \frac{\partial \ell_n^{(n)}(\beta_{M,0,n})}{\partial \beta_M} \middle| y_{0:n-1} \right\} \\ Z_{n-2}^{(n)} &= \frac{\partial \ell_{n-2}^{(n)}(\beta_{M,0,n})}{\partial \beta_M} + \mathrm{E}_{\text{wide}} \left\{ \frac{\partial \ell_{n-1}^{(n)}(\beta_{M,0,n})}{\partial \beta_M} \middle| y_{0:n-2} \right\} \\ &\qquad\qquad + \mathrm{E}_{\text{wide}} \left\{ \frac{\partial \ell_n^{(n)}(\beta_{M,0,n})}{\partial \beta_M} \middle| y_{0:n-1} \right\}. \end{aligned}$$

We see from the first few values that $Z_k^{(n)}$ may be expressed alternatively as

$$Z_k^{(n)} = \begin{cases} \frac{\partial \ell_n^{(n)}(\beta_{M,0,n})}{\partial \beta_M} & \text{if } k = n \\ \frac{\partial \ell_k^{(n)}(\beta_{M,0,n})}{\partial \beta_M} + \sum_{r=k+1}^{n} \mathrm{E}_{\text{wide}} \left\{ \frac{\partial \ell_r^{(n)}(\beta_{M,0,n})}{\partial \beta_M} \middle| y_{0:r} \right\} & \text{else.} \end{cases} \tag{3.9}$$

Taking expectations under the *wide* model, we may write further

$$\begin{aligned} \mathrm{E}_{\text{wide}} \left\{ Z_{k+1}^{(n)} \middle| y_k \right\} &= \mathrm{E}_{\text{wide}} \left\{ \frac{\partial \ell_{k+1}^{(n)}(\beta_{M,0,n})}{\partial \beta_M} \middle| y_{0:k} \right\} \\ &\qquad + \sum_{r=k+2}^{n} \mathrm{E}_{\text{wide}} \left\{ \frac{\partial \ell_r^{(n)}(\beta_{M,0,n})}{\partial \beta_M} \middle| y_{0:r-1} \right\} \\ &= \sum_{r=k+1}^{n} \mathrm{E} \left\{ \frac{\partial \ell_r^{(n)}(\beta_{M,0,n})}{\partial \beta_M} \middle| y_{0:r-1} \right\}. \end{aligned} \tag{3.10}$$

This result will allow us to give a decomposition of the partial derivative of the log-likelihood. Inserting (3.10) in (3.9), we get

$$\frac{\partial \ell_k(\beta_{M,0,n})}{\partial \beta_{M,n}} = Z_k^{(n)} - \mathrm{E}_{\text{wide}}\left\{ Z_{k+1}^{(n)} \middle| y_{0:k} \right\} \tag{3.11}$$

Writing

$$\begin{aligned}\frac{\partial \ell_k(\beta_{M,0,n})}{\partial \beta_{M,n}} &= Z_k^{(n)} - \mathrm{E}_{\text{wide}}\left\{ Z_k^{(n)} \middle| y_{0:k} \right\} \\ + \mathrm{E}_{\text{wide}}\left\{ Z_k^{(n)} \middle| y_{0:k} \right\} &- \mathrm{E}_{\text{wide}}\left\{ Z_{k+1}^{(n)} \middle| y_{0:k} \right\}\end{aligned}$$

we get by telescoping the alternative expression for the log-likelihood

$$\begin{aligned}\frac{\partial \ell^{(n)}(\beta_{M,0,n})}{\partial \beta_M} &= \sum_{k=1}^{n} \frac{\partial \ell_k(\beta_{M,0,n})}{\partial \beta_M} \\ &= \sum_{k=2}^{n} \left\{ Z_k^{(n)} - \mathrm{E}_{\text{wide}}\left\{ Z_k^{(n)} \middle| y_{0:k-1} \right\} \right\} + Z_1^{(n)}.\end{aligned}$$

For $0 \leq k \leq n$, define further with Sethuraman and Varadhan (2005) the random variable

$$M_k = \sum_{l=2}^{k} \left\{ Z_l^{(n)} - \mathrm{E}_{\text{wide}}\left\{ Z_l^{(n)} \middle| y_{0:l-1} \right\} \right\}.$$

Taking expectation we get

$$\mathrm{E}_{\text{wide}}\left\{ M_k \middle| y_{0:k-1} \right\} = M_{k-1},$$

which implies that M_k is a martingale with regards to the history $y_{0:k-1}$ at time k.

This means that the candidate score function $U_{M,n}$ can be decomposed into a Martingale plus a term $Z_1^{(n)}$.

$$U_{M,n} = \frac{1}{\sqrt{n}} \frac{\partial \ell^{(n)}(\beta_{M,0,n})}{\partial \beta_{M,n}} = \frac{1}{\sqrt{n}} M_n + \frac{1}{\sqrt{n}} Z_1^{(n)}.$$

Until now we have followed the general proceedings in Sethuraman and Varadhan (2005). To show that the $U_{M,n}$ has the same limit as $n^{-\frac{1}{2}} M_n$ and therefore constitutes a martingale asymptotically, we need to show that the term $Z_1^{(n)}$ is bounded for the case of a misspecified model on the form of (2.3)

Define for a finite $k < n$ the $(2\,q + 6\,w) \times 1$ vector

$$\Omega_k = \sum_{t=k+1}^{n} \mathrm{E}_{\text{wide}} \left\{ \frac{\partial \ell_t^{(n)}(\beta_{M,0,n})}{\partial \beta_M} \middle| y_{0:k} \right\}$$

and write (3.9) as

$$Z_k^{(n)} = \frac{\partial \ell_k^{(n)}(\beta_{M,0,n})}{\partial \beta_M} + \Omega_k$$

We know from Section 2.3 that $\partial \ell_k^{(n)}(\beta_{M,0,n})/\partial \beta_M$ is bounded for each k. We need to show that Ω_k is bounded for each k too.

For $l = 1, \dots, 2q + 6w$ let $\omega_{k,l}$ denote the elements in Ω_k. For a candidate model M with $q > 0$, the first element in Ω_k is

$$\begin{aligned}
\omega_{k,1} &= \sum_{t=k+1}^{n} \mathrm{E}_{\text{wide}} \left\{ \sum_{r=0}^{2} \Big(y_{t,0} - \pi_{Mr0}(x_t) \Big) u_t y_{t-1,r} \middle| y_{0:k} \right\} \\
&= \sum_{t=k+1}^{n} \sum_{r=0}^{2} \bigg\{ \phi_{r0}(x_t) u_t \, \mathrm{E}_{\text{wide}} \left\{ y_{t-1,r} \middle| y_{0:k} \right\} \bigg\} \\
&= \sum_{t=k+1}^{n} \sum_{r=0}^{2} \bigg\{ \phi_{r0}(x_t) u_t \sum_{r'=0}^{2} \Big\{ P_{r'r}^{(t-k-1)}(k) y_{k,r'} \Big\} \bigg\}.
\end{aligned}$$

As the *inhomogeneous* Markov chain $\{y_t\}$ is *strongly ergodic* (see Section 2.4), we have that $\lim_{t\to\infty} P_{kj}^{(t)}(k) = \mathrm{E}_{\text{wide}}\; y_{t+k,j}$. Thus for each finite k there exists a constant $\epsilon \in (0,1)$ such that for $t > k$ and for $r, r' = 0, 1, 2$ it is the case that

$$P_{r'r}^{(t-k-1)}(k) < \mathrm{E}_{\text{wide}}\,(y_{t-1,r}) + \epsilon^{t-k-1}.$$

We may therefore write

$$\begin{aligned}
\omega_{k,1} &< \sum_{t=k+1}^{n} \sum_{r=0}^{2} \bigg\{ \phi_{r0}(x_t) u_t \sum_{r'=0}^{2} \Big(\mathrm{E}_{\text{wide}}\,(y_{t-1,r}) + \epsilon^{t-k-1} \Big) y_{k,r'} \bigg\} \\
&= \sum_{t=k+1}^{n} \sum_{r=0}^{2} \bigg\{ \phi_{r0}(x_t) u_t \, \mathrm{E}_{\text{wide}}\; y_{t-1,r} \bigg\} + \sum_{t=k+1}^{n} \sum_{r=0}^{2} \bigg\{ \phi_{r0}(x_t) u_t \epsilon^{t-k-1} \bigg\},
\end{aligned}$$

where we in the last line have used that $\sum_{r'=0}^{2} y_{k,r'} = 1$.

The second double sum in this expression is a convergent series, thus bounded. For the first double sum, write

$$\begin{aligned}\sum_{t=k+1}^{n}\sum_{r=0}^{2}&\Big\{\phi_{r0}(x_t)u_t\,\mathrm{E}_{\text{wide}}\ y_{t-1,r}\Big\}\\ &=\sum_{t=1}^{n}\sum_{r=0}^{2}\Big\{\phi_{r0}(x_t)u_t\,\mathrm{E}_{\text{wide}}\ y_{t-1,r}\Big\}-\sum_{t=1}^{k}\sum_{r=0}^{2}\Big\{\phi_{r0}(x_t)u_t\,\mathrm{E}_{\text{wide}}\ y_{t-1,r}\Big\}.\end{aligned} \tag{3.12}$$

Recall that $\gamma_{0,0}$ is the first element in $\beta_{M,n}$. We have therefore that

$$\begin{aligned}\mathrm{E}_{\text{wide}}&\left\{\frac{\partial\ell^{(n)}(\beta_{M,0,n})}{\partial\gamma_{0,0}}\right\}\\ &=\sum_{t=1}^{n}\sum_{r=0}^{2}\mathrm{E}_{\text{wide}}\left\{\Big(y_{t,0}-\pi_{Mr0}(x_t)\Big)u_ty_{t-1,r}\right\}\\ &=\sum_{t=1}^{n}\sum_{r=0}^{2}\mathrm{E}_{\text{wide}}\left\{\mathrm{E}_{\text{wide}}\left\{\Big(y_{t,0}-\pi_{Mr0}(x_t)\Big)u_ty_{t-1,r}\Big|y_{0:t-1}\right\}\right\}\\ &=\sum_{t=1}^{n}\sum_{r=0}^{2}\phi_{r0}(x_t)u_t\,\mathrm{E}_{\text{wide}}\ y_{t-1,r}\end{aligned}$$

We may therefore write (3.13) as

$$\begin{aligned}\sum_{t=k+1}^{n}\sum_{r=0}^{2}&\Big\{\phi_{r0}(x_t)u_t\,\mathrm{E}_{\text{wide}}\ y_{t-1,r}\Big\}\\ &=\mathrm{E}_{\text{wide}}\ \frac{\partial\ell^{(n)}(\beta_{M,0,n})}{\partial\gamma_{0,0}}-\mathrm{E}_{\text{wide}}\sum_{t=1}^{k}\frac{\partial\ell_t^{(n)}(\beta_{M,0,n})}{\partial\gamma_{0,0}}.\end{aligned} \tag{3.13}$$

From (3.4) we may show that the expected score vector of the candidate model is bounded asymptotically at the least false parameter $\beta_{M,0,n}$. Thus the first term in (3.13) is bounded for all n. Since $\partial\ell_t^{(n)}(\beta_{M,0,n})/\partial\gamma_{0,0}$ is bounded for any $t\leq n$ and since k is a finite integer, we have that the second sum in (3.13) is bounded too. Thus $\omega_{k,1}$ is bounded.

The situation is parallel for every element in Ω_k since each of these elements are on the same general form as $\omega_{k,1}$. The same argument holds therefore for these other elements in Ω_k too. Thus every element in Ω_k is bounded for a finite k, which again implies that $Z_k^{(n)}$ is bounded for any finite k.

We therefore have that $U_{M,n}$ and $\frac{1}{\sqrt{n}}M_n$ have the same limit, which implies that the candidate score function $U_{M,n}$ is a Martingale asymptotically.

To investigate this limit we use the Martingale Central Limit Theorem stated in Hall and Heyde (1980):

Lemma 3.3.1. *For each $n \geq 1$ let $\left\{(W_i^{(n)}, \mathcal{F}_i^{(n)}) \mid 0 \leq i \leq n\right\}$ be a martingale relative to the nested family $\mathcal{F}_i^{(n)} \subset \mathcal{F}_{i+1}^{(n)}$ with $W_0^{(n)} = 0$. Let $\xi_i^{(n)} = W_i^{(n)} - W_{i-1}^{(n)}$ be their differences. Suppose that*

$$\begin{aligned} \underset{1 \leq i \leq n}{max} \mid \xi_i^{(n)} \mid \quad &\to 0 \qquad \textit{and} \\ \sum_{i=1}^{n} \mathrm{E}\left\{(\xi_i^{(n)})^2 \mid \mathcal{F}_{i-1}^{(n)}\right\} &\to \eta \end{aligned}$$

Then

$$W_n^{(n)} \to_d N(0, \eta)$$

Following Sethuraman and Varadhan (2005), we define $W_t^{(n)} = \frac{1}{\sqrt{n}} M_t$ for $t \geq 2$. To be in full accordance with the theorem, we define $W_1^{(n)} = W_0^{(n)} = 0$. The history $\mathcal{F}_t^{(n)}$ in our situation is $y_{0:t}$. Obviously $y_{0:t} \subset y_{0:t+1}$. We define also

$$\xi_t^{(n)} = \frac{\Delta M_t}{\sqrt{n}} = \frac{1}{\sqrt{n}} \left\{ Z_t^{(n)} - \mathrm{E}\left\{ Z_t^{(n)} \middle| y_{0:t-1} \right\} \right\}, \tag{3.14}$$

where $\Delta M_t = M_t - M_{t-1}$.

To show that the first condition in the Martingale Central Limit Theorem is fulfilled, notice that we may write by inserting (3.11) in (3.14)

$$\begin{aligned} &Z_t^{(n)} - \mathrm{E}\left\{ Z_t^{(n)} \middle| y_{0:t-1} \right\} \\ = \quad & \frac{\partial \ell_t^{(n)}(\beta_{M,0,n})}{\partial \beta_M} + \mathrm{E}_{\mathrm{wide}}\left\{ Z_{t+1} \middle| y_{0:t} \right\} - \mathrm{E}_{\mathrm{wide}}\left\{ Z_t \middle| y_{0:t-1} \right\} \\ = \quad & \frac{\partial \ell_t^{(n)}(\beta_{M,0,n})}{\partial \beta_M} + \sum_{r=t+1}^{n} \mathrm{E}_{\mathrm{wide}}\left\{ \frac{\partial \ell_r^{(n)}(\beta_{M,0,n})}{\partial \beta_M} \middle| y_{0:r-1} \right\} \\ & \qquad - \sum_{r=t}^{n} \mathrm{E}_{\mathrm{wide}}\left\{ \frac{\partial \ell_r^{(n)}(\beta_{M,0,n})}{\partial \beta_M} \middle| y_{0:r-1} \right\} \\ = \quad & \frac{\partial \ell_t(\beta_{M,0,n})}{\partial \beta_M} - \mathrm{E}_{\mathrm{wide}}\left\{ \frac{\partial \ell_t^{(n)}(\beta_{M,0,n})}{\partial \beta_M} \middle| y_{0:t-1} \right\}, \end{aligned} \tag{3.15}$$

which clearly is bounded. Thus we have that $\xi_t^{(n)} \to_p 0$ for each $t \leq n$.

To show that the second condition in the Martingale central limit theorem is fulfilled, first notice that

$$\sum_{t=1}^{n} \mathrm{E}_{\mathrm{wide}}\left\{ (\xi_t^{(n)})^2 \middle| y_{0:t-1} \right\} = \sum_{t=1}^{n} \mathrm{Var}_{\mathrm{wide}}\left\{ \xi_t^{(n)} \middle| y_{0:t-1} \right\},$$

since $\mathrm{E}_{\mathrm{wide}}\{\xi_t^{(n)}|y_{0:t-1}\} = 0$. Further we have by inserting (3.15) in (3.14)

$$\begin{aligned}
&\sum_{t=1}^{n} \mathrm{Var}_{\mathrm{wide}}\left\{\xi_t^{(n)}\middle|y_{0:t-1}\right\} \\
&\qquad = \frac{1}{n}\sum_{t=1}^{n} \mathrm{Var}_{\mathrm{wide}}\left\{\frac{\partial\ell_t^{(n)}(\beta_{M,0,n})}{\partial\beta_M} - \mathrm{E}_{\mathrm{wide}}\left\{\frac{\partial\ell_t^{(n)}(\beta_{M,0,n})}{\partial\beta_M}\middle|y_{0:t-1}\right\}\middle|y_{0:t-1}\right\} \\
&\qquad\quad = \frac{1}{n}\sum_{t=1}^{n} \mathrm{Var}_{\mathrm{wide}}\left\{\frac{\partial\ell_t^{(n)}(\beta_{M,0,n})}{\partial\beta_M}\middle|y_{0:t-1}\right\}
\end{aligned}$$

We know that $\partial\ell_t^{(n)}(\beta_{M,0,n})/\partial\beta_{M,n}$ is a bounded function of random variables y_t and y_{t-1}. For each element in $\sum_{t=1}^{n} \mathrm{Var}_{\mathrm{wide}}\{\xi_t^{(n)}|y_{0:t-1}\}$ we may therefore define uniformly bounded functions $f_k^{(t)} = f_k^{(t)}(x)$, $k = 0, 1, 2$ such that the any element ω_{xi} in this matrix may be written as

$$\omega_{\mathrm{xi}} = \frac{1}{n}\sum_{t=1}^{n}\sum_{k=0}^{2} f_k^{(t)} y_{t-1,k}.$$

Again, $f_k^{(t)} y_{t-1,k}$ is on form of the general function $\psi_{t,k,j}$ in (3.1). We may thus use Lemma 3.1.1 in the same manner as in the proof of Theorem 3.1.2 to show that there for each element in $\sum_{t=1}^{n} \mathrm{Var}_{\mathrm{wide}}\{\xi_t^{(n)}|y_{0:t-1}\}$ exist a constant G such that

$$\mathrm{Var}_{\mathrm{wide}}\ \frac{1}{n}\sum_{t=1}^{n}\sum_{k=0}^{2} f_k^{(t)} y_{t-1,k} < \frac{G}{n},$$

which implies that

$$\mathrm{Var}_{\mathrm{wide}}\ \frac{1}{n}\sum_{t=1}^{n}\sum_{k=0}^{2} f_k^{(t)} y_{t-1,k} \to_p 0.$$

Since this is the case for each element in $\sum_{t=1}^{n} \mathrm{Var}_{\mathrm{wide}}\{\xi_t^{(n)}|y_{0:t-1}\}$, it follows that there exists a constant vector η such that

$$\sum_{t=1}^{n} \mathrm{Var}_{\mathrm{wide}}\left\{\xi_t^{(n)}\middle|y_{0:t-1}\right\} \to_p \eta.$$

Thus the second condition in the Martingale Central Limit theorem is fulfilled. It follows that there exists a non-random vector U_M such that

$$\frac{M_n}{\sqrt{n}} \rightarrow_p U_M \sim N(0, \eta) \tag{3.16}$$

Now, we want to show that $\eta = K_M$. We know that

$$K_{M,n} = \mathrm{Var}_{\mathrm{wide}}\ U_{M,n} = \mathrm{Var}_{\mathrm{wide}} \left\{ \sum_{t=1}^{n} \xi_t^{(n)} + \frac{Z_1^{(n)}}{\sqrt{n}} \right\}.$$

As the limit K_M of $K_{M,n}$ is well-defined and $Z_1^{(n)}$ is bounded, it must be the case that

$$\lim_{n\to\infty} \mathrm{Var}_{\mathrm{wide}} \left\{ \sum_{t=1}^{n} \xi_t^{(n)} \right\} = K_M$$

Martingales have uncorrelated increments, we may therefore write

$$\mathrm{Var}_{\mathrm{wide}} \left\{ \sum_{t=1}^{n} \xi_t^{(n)} \right\} = \sum_{t=1}^{n} \mathrm{Var}_{\mathrm{wide}}\ \xi_t^{(n)}.$$

Also, since $\mathrm{E}_{\mathrm{wide}}\, \{\xi_t^{(n)} | y_{0:t-1}\} = 0$, we get from the law of total covariance

$$\mathrm{Var}_{\mathrm{wide}}\ \xi_t^{(n)} = \mathrm{E}_{\mathrm{wide}}\ \mathrm{Var}_{\mathrm{wide}} \left\{ \xi_t^{(n)} \middle| y_{0:t-1} \right\}.$$

From this, it follows that

$$\eta = \lim_{n\to\infty} \sum_{t=1}^{n} \mathrm{Var}_{\mathrm{wide}} \left\{ \xi_t^{(n)} \middle| y_{0:t-1} \right\} = \lim_{n\to\infty} \sum_{t=1}^{n} \mathrm{Var}_{\mathrm{wide}}\ \xi_t^{(n)} = K_M. \tag{3.17}$$

Combining (3.17) and (3.16), we get then from the Martingale central limit theorem that

$$\frac{M_n}{\sqrt{n}} \rightarrow_{\mathrm{p}} U_M \sim N\Big(0, K_M\Big).$$

Since the candidate random score vector $U_{M,n}$ has the same limit as $\frac{M_n}{\sqrt{n}}$, it follows that the random score vector $U_{M,n}$ of the candidate model is normally in the limit:

$$U_{M,n} \rightarrow_{\mathrm{p}} U_M \sim N\Big(0, K_M\Big).$$

As the *wide* model is a special case of the general model (2.3), it follows from the same arguments, that $U_n \to_d N(0, J)$. The joint limit distribution of the random score vectors will therefore be

$$\begin{pmatrix} U_n \\ U_{M,n} \end{pmatrix} \to_d N\left(0, \begin{pmatrix} J & C_M \\ C_M & K_M \end{pmatrix}\right).$$

This is an important result. It will make us able to find the joint asymptotic distribution of maximum likelihood estimators, which is what I will do in the next section.

3.4 Approximate Normality of MLEs

With the asymptotic normality of the random score vectors assured, we are now in the position to show that maximum likelihood estimators $\hat{\beta}$ of the wide model and $\hat{\beta}_M$ of the candidate model are approximately normally distributed. The following theorem, proved in N. L. Hjort and Pollard (1993), will be of central importance to the demonstration:

Lemma 3.4.1. *Suppose $A_n(s)$ is convex and can be represented as $\frac{1}{2}s^tVs+U_n^ts+C_n+r_n(s)$ where V is symmetric and positive definite, U_n is stochastically bounded, C_n is arbitrary, and $r_n(s)$ goes to zero in probability for each s. Then α_n, the argmin of A_n, is only $o_p(1)$ away from $\beta_n = -V^{-1}U_n$, the argmin of $\frac{1}{2}s^tVs+U_n^ts+C_n$. If also $U_n \to_d U$ then $\alpha_n \to_d -V^{-1}U$.*

Define first the function

$$A_n(h) = \ell_M^{(n)}(\beta_{M,0,n}+h) - \ell_M^{(n)}(\beta_{M,0,n})$$

where $\beta_{M,0,n}$ is the least false parameter of the parametric class of the candidate model M given covariates $x_0, x_1, \dots x_n$. We know from Section 2.3 that the likelihood of the dynamic multinomial logit function is a concave function. Thus the function $A_n(h)$ is a concave function, with the maximizer $\hat{\beta}_{M,n} - \beta_{M,0,n}$.

By exact Taylor expansion, $A_n(h)$ may be reformulated as

$$\begin{aligned} A_n(h) &= \ell_M^{(n)}(\beta_{M,0,n}+h) - \ell_M^{(n)}(\beta_{M,0,n}) \\ &= \frac{\partial \ell_M^{(n)}(\beta_{M,0,n})}{\partial \beta_M}^{\mathrm{t}} h + \frac{1}{2} h^{\mathrm{t}} \frac{\partial^2 \ell_M^{(n)}(\beta_{M,0,n})}{\partial \beta_M \partial \beta_M^{\mathrm{t}}} h \\ &\qquad\qquad + \frac{1}{6} \sum_q \sum_r \sum_w \frac{\partial^3 \ell_M^{(n)}(\beta_{M,0,n}+h^*)}{\partial \beta_{M,q} \partial \beta_{M,r} \partial \beta_{M,w}} h_q h_r h_w, \end{aligned}$$

where h^* is a $(2\,q+6\,w)\times 1$ vector such that $0 < |h^*| < |h|$. The partial derivate operator $\partial/\partial\beta_{M,r}$ here denotes derivation with respect to the r'th element of $\beta_{M,n}$. The variable h_r is the corresponding r'th element of h, and $\sum_r$ is an abbreviation for summing over all $2q+6w$ elements in β_M.

Define for every $1 \leq t \leq n$ the entity

$$\begin{aligned} R_t(y_t, y_{t-1}, h) = &\frac{1}{2} h^{\mathrm{t}} \frac{\partial^2 \ell^{(n)}_{M,t}(\beta_{M,0,n})}{\partial\beta\partial\beta^{\mathrm{t}}_M} h \\ &+ \frac{1}{6} \sum_q \sum_r \sum_w \frac{\partial^3 \ell^{(n)}_{M,t}(\beta_{M,0,n} + h^*)}{\partial\beta_{M,q}\partial\beta_{M,r}\partial\beta_{M,w}} h_q h_r h_w. \end{aligned}$$

and rewrite

$$A_n(h) = \frac{\partial \ell^{(n)}_M(\beta_{M,0,n})}{\partial\beta_M}^{\mathrm{t}} h + \sum_{t=1}^{n} R_t(y_t, y_{t-1}, h)$$

Now, let

$$\begin{aligned} v_{t,0}(h) &= \mathrm{E}_{\mathrm{wide}}\ \frac{1}{6} \sum_q \sum_r \sum_w \frac{\partial^3 \ell_{M,t}(\beta_{M,0,n} + h^*)}{\partial\beta_{M,r}\partial\beta_{M,s}\partial\beta_{M,v}} h_q h_r h_w, \\ r_n(h) &= \sum_{t=1}^{n} v_{t,0}(h) \end{aligned}$$

and

$$r_{n,0}(h) = \sum_{t=1}^{n} \Big\{ R_t(y_t, y_{t-1}, h) - \mathrm{E}_{\mathrm{wide}}\ R_t(y_t, y_{t-1}, h) \Big\}.$$

Consider for $h = \frac{s}{\sqrt{n}}$

$$\begin{aligned} A_n(s) &= \frac{\partial \ell^{(n)}_M(\beta_{M,0,n})}{\partial\beta_M}^{\mathrm{t}} \frac{s}{\sqrt{n}} + \sum_{t=1}^{n} R_t(y_t, y_{t-1}, \frac{s}{\sqrt{n}}) \\ &= \frac{\partial \ell^{(n)}_M(\beta_{M,0,n})}{\partial\beta_M}^{\mathrm{t}} \frac{s}{\sqrt{n}} + \sum_{t=1}^{n} \mathrm{E}_{\mathrm{wide}}\ R_t(y_t, y_{t-1}, \frac{s}{\sqrt{n}}) \\ &\qquad + \sum_{t=1}^{n} \Big\{ R_t(y_t, y_{t-1}, \frac{s}{\sqrt{n}}) - \mathrm{E}_{\mathrm{wide}}\ R_t(y_t, y_{t-1}, \frac{s}{\sqrt{n}}) \Big\} \\ &= \frac{\partial \ell^{(n)}(\beta_{M,0,n})}{\partial\beta_M}^{\mathrm{t}} \frac{s}{\sqrt{n}} + \mathrm{E}_{\mathrm{wide}} \Big\{ \sum_{t=1}^{n} \frac{1}{2} \frac{1}{n} s^{\mathrm{t}} \frac{\partial^2 \ell^{(n)}_{M,t}(\beta_{M,0,n})}{\partial\beta_M\partial\beta^{\mathrm{t}}_M} s \Big\} \\ &+ \sum_{t=1}^{n} v_{t,0}(\frac{s}{\sqrt{n}}) + \sum_{t=1}^{n} \Big\{ R_t(y_t, y_{t-1}, \frac{s}{\sqrt{n}}) - \mathrm{E}_{\mathrm{wide}}\ R_t(y_t, y_{t-1}, \frac{s}{\sqrt{n}}) \Big\} \\ &= U^{\mathrm{t}}_{M,n} s - \frac{1}{2} s^{\mathrm{t}} J_{M,n} s + r_n(s) + r_{n,0}(s), \end{aligned}$$

where we in the last line have used that the Fisher information matrix $J_{M,n}$ of the candidate model is the negative mean of the candidate Hessian matrix.

We know that $J_{M,n}$ is negative definite due to the concavity of $\ell^{(n)}(\beta_M)$. We know also that $U_{M,n}$ is stochastically bounded. Thus $J_{M,n}$ and $U_{M,n}$ fulfill the conditions in Lemma 3.4.1. To show that we are in full accordance with the conditions in lemma we need to show that $r_n(s) + r_{n,0}(s)$ goes to zero in probability for each s.

Consider the case for $r_n(s)$ first. We have that

$$v_{t,0}(s) = \frac{1}{6}\sum_q \sum_r \sum_w \mathrm{E}_{\text{wide}} \left\{ \frac{\partial^3 \ell^{(n)}_{M,t}(\beta_{M,0,n} + \frac{s*}{\sqrt{n}})}{\partial \beta_{M,q} \partial \beta_{M,r} \partial \beta_{M,w}} \frac{s_q s_r s_w}{n\sqrt{n}} \right\}.$$

We may show that for each third partial derivative of $\ell^{(n)}_{M,t}(\beta_{M,n})$ there exists uniformly bounded function $f_{q,r,w,k}(x_t)$ such that $|f_{q,r,w,k}(x_t)| < 1$. We may write

$$\frac{\partial^3 \ell_t(\beta_{M,0,n} + \frac{s}{\sqrt{n}})}{\partial \beta_q \partial \beta_r \partial \beta_w} = d_{t,q} d_{t,r} d_{t,w} (\sum_{k=0}^{2} f_{q,r,w,k}(x_t) y_{t-1,k}), \tag{3.18}$$

where d_t denotes either u_t or z_t, depending on the candidate model and the value of the indexes q, r, w.

Since the space of covariates is bounded and $\mathrm{E}_{\text{wide}}\ y_{t,k} < 1$ for all t and k, there exists a constant G such that for all s and all q, r, w it is the case that

$$\mathrm{E}_{\text{wide}}\ \frac{\partial^3 \ell_{M,t}(\beta_{M,0,n} + \frac{s*}{\sqrt{n}})}{\partial \beta_{M,q} \partial \beta_{M,r} \partial \beta_{M,w}} < G.$$

From the boundedness of the covariate space Γ, assumed in Section 2.3, it follows that there exists a constant G' such that for each t we have that

$$v_{t,0}(s) < \frac{G}{6} \sum_{q=1}^{p} \sum_{r=1}^{p} \sum_{w=1}^{p} \frac{s_q s_r s_w}{n\sqrt{n}} < \frac{G'}{n\sqrt{n}}.$$

From this it follows that

$$r_n(s) = \sum_{t=1}^{n} v_{t,0}(s) < \sum_{t=1}^{n} \frac{G'}{n\sqrt{n}} = \frac{G'}{\sqrt{n}} = o_p(1)$$

which means that $r_n(s)$ goes to zero in probability.

Consider then $r_{n,0}(s)$. Clearly $\mathrm{E}\, r_{n,0}(s) = 0$, so if $\mathrm{Var}\, r_{n,0}(s) \to 0$, it follows that $r_{n,0}(s) \to_p 0$ too.

Write

$$
\begin{aligned}
&\mathrm{Var}_{\mathrm{wide}}\ r_{n,0}(s) = \mathrm{Var}_{\mathrm{wide}}\left\{\sum_{t=1}^{n}\left\{R_t(y_t, y_{t-1}, \frac{s}{\sqrt{n}}) - \mathrm{E}_{\mathrm{wide}}\ R_t(y_t, y_{t-1}, \frac{s}{\sqrt{n}})\right\}\right\}\\
&= \mathrm{Var}_{\mathrm{wide}} \sum_{t=1}^{n} R_t(y_t, y_{t-1}, \frac{s}{\sqrt{n}})\\
&= \mathrm{Var}_{\mathrm{wide}}\left\{\sum_{t=1}^{n}\left\{\frac{1}{2}\frac{1}{n}s^{\mathrm{t}}\frac{\partial^2 \ell_{M,t}^{(n)}(\beta_{M,0,n})}{\partial\beta_M\partial\beta_M^{\mathrm{t}}}s\right.\right.\\
&\qquad\qquad \left.\left.+\frac{1}{6}\sum_{q=1}\sum_{r=1}\sum_{w=1}\frac{\partial^3 \ell_{M,t}^{(n)}(\beta_{M,0,n}+\frac{s*}{\sqrt{n}})}{\partial\beta_{M,q}\partial\beta_{M,r}\partial\beta_{M,w}}\frac{s_q}{\sqrt{n}}\frac{s_r}{\sqrt{n}}\frac{s_w}{\sqrt{n}}\right\}\right\}\\
&= \mathrm{Var}_{\mathrm{wide}}\left\{\sum_{t=1}^{n}\frac{1}{2}\frac{1}{n}s^{\mathrm{t}}\frac{\partial^2 \ell_{M,t}^{(n)}(\beta_{M,0,n})}{\partial\beta_M\partial\beta_M^{\mathrm{t}}}s\right\}\\
&\qquad\qquad + \mathrm{Cov}_{\mathrm{wide}}\left\{\sum_{t=1}^{n}\frac{1}{2}\frac{1}{n}s^{\mathrm{t}}\frac{\partial^2 \ell_{M,t}^{(n)}(\beta_{M,0,n})}{\partial\beta_M\partial\beta_M^{\mathrm{t}}}s, \star\right\}\\
&\qquad\qquad + \mathrm{Cov}_{\mathrm{wide}}\left\{\star, \sum_{t=1}^{n}\frac{1}{2}\frac{1}{n}s^{\mathrm{t}}\frac{\partial^2 \ell_{M,t}^{(n)}(\beta_{M,0,n})}{\partial\beta_M\partial\beta_M^{\mathrm{t}}}s\right\} + \mathrm{Var}_{\mathrm{wide}}\left\{\star\right\}
\end{aligned}
$$

where

$$
\star = \frac{1}{6}\sum_{t=1}^{n}\sum_{q,r,w=1}^{p}\frac{\partial^3 \ell_{M,t}^{(n)}(\beta_{M,0,n}+\frac{s*}{\sqrt{n}})}{\partial\beta_{M,q}\partial\beta_{M,r}\partial\beta_{M,w}}\frac{s_q}{\sqrt{n}}\frac{s_r}{\sqrt{n}}\frac{s_w}{\sqrt{n}}
$$

We may show with the same techniques as in Theorem 3.1.2 and Theorem 3.2.1, that

$$
\sum_{t=1}^{n}\frac{1}{n}\frac{\partial^2 \ell_t^{(n)}(\beta_{M,0,n})}{\partial\beta_M\partial\beta_{M,n}^{\mathrm{t}}} \to J_M.
$$

The three first terms in the expression of $\mathrm{Var}_{\mathrm{wide}}\ r_{n,0}(s)$ will therefore go to zero in probability. We need to show that this is the case also for the last term.

Define

$$
\omega_t^{(n)}(s) = \sum_{q,r,w=1}^{p}\frac{\partial^3 \ell_{M,t}^{(n)}(\beta_{M,n}+\frac{s*}{\sqrt{n}})}{\partial\beta_{M,q}\partial\beta_{M,r}\partial\beta_{M,w}}s_q s_r s_w,
$$

Rewriting the third partial derivative in the same manner as in (3.18), we see that $\omega_t(s)$ is a finite sum of functions on the general form ψ_{tkj}. Proceeding in the same manner as in the proof of Theorem 3.1.2, we may show that there exists a constant G such that

$$\sum_{t=1}^{n}\sum_{t'=1}^{n}\mathrm{Cov}_{\mathrm{wide}}\left\{\omega_t^{(n)}(s),\omega_{t'}^{(n)}(s)\right\} < nG$$

Consequently, we have that

$$\mathrm{Var}_{\mathrm{wide}}\left\{\sum_{t=1}^{n}\sum_{k=0}^{2}\sum_{q,r,w=1}^{p}\frac{\partial^3\ell_{M,t}^{(n)}(\beta_{M,0,n}+\frac{s*}{\sqrt{n}})}{\partial\beta_{M,q}\partial\beta_{M,r}\partial\beta_{M,w}}\frac{s_q s_r s_w}{n\sqrt{n}}y_{i-1,k}\right\}$$
$$=\frac{1}{n^3}\sum_{t=1}^{n}\sum_{t'=1}^{n}\mathrm{Cov}_{\mathrm{wide}}\left\{\omega_t(s),\omega_{t'}(s)\right\} < \frac{nG}{n^3}$$

which goes to zero asymptotically. Thus $\mathrm{E}\, r_{n,0}(s)=0$ and $\mathrm{Var}_{\mathrm{wide}}\ r_{n,0}(s)\to 0$, which means that $r_{n,0}(s)=o_p(1)$.

We have then that both $r_n \to_p 0$ and $r_{n,0}\to_p 0$ which means that we may write

$$A_n(s) = -\frac{1}{2}s^t J_M s + U_n^{\mathrm{t}} s + o_p(1)$$

in line with the conditions stated in Lemma 3.4.1.

As $\sqrt{n}(\hat{\beta}_{M,n}-\beta_{M,0,n})$ is the maximizer of the concave function $A_n(s)$, it follows from Lemma 3.4.1 that

$$\sqrt{n}(\hat{\beta}_M-\beta_{M,0,n}) = J_{M,n}^{-1}U_{M,n}+o_p(1). \tag{3.19}$$

The *wide* model is a special case of the general candidate model (2.3). It will therefore also be the case that

$$\sqrt{n}(\hat{\beta}-\beta_{\mathrm{true}}) = J_n^{-1}U_n+o_p(1). \tag{3.20}$$

We know from Theorem 3.2.1 that $J_n\to J$ and that $J_{M,n}\to J_M$. From Section 3.3 we know that $U_n\to_p U$ and that $U_{M,n}\to_p U_M$. Using Slutsky's theorem we may then write

$$J_n^{-1}U_n+o_p(1)\to_p J^{-1}U$$
$$J_{M,n}^{-1}U_{M,n}+o_p(1)\to_p J_M^{-1}U_M$$

We have from Section 3.3 the limiting distribution

$$\begin{pmatrix} J^{-1}U \\ J_M^{-1}U_M\end{pmatrix} \sim \mathrm{N}\left(0,\begin{pmatrix} J^{-1} & J^{-1}C_M J_M^{-1} \\ J_M^{-1}C_M J^{-1} & J_M^{-1}K_M J_M^{-1}\end{pmatrix}\right),$$

which means that

$$\begin{pmatrix} J_n^{-1}U_n \\ J_{M,n}^{-1}U_{M,n} \end{pmatrix} \approx_d \mathrm{N}\left(0, \begin{pmatrix} J_n^{-1} & J_n^{-1}C_{M,n}J_{M,n}^{-1} \\ J_{M,n}^{-1}C_{M,n}J_n^{-1} & J_{M,n}^{-1}K_{M,n}J_{M,n}^{-1} \end{pmatrix}\right),$$

From (3.19) and (3.20) it then follows that the maximum likelihood estimators $\hat{\beta}$ of the wide model and $\hat{\beta}_M$ of the candidate model have the approximate joint distribution

$$\begin{pmatrix} \sqrt{n}(\hat{\beta} - \beta_{\text{true}}) \\ \sqrt{n}(\hat{\beta}_M - \beta_{M,0,n}) \end{pmatrix} \approx_d \mathrm{N}\left(0, \begin{pmatrix} J_n^{-1} & J_n^{-1}C_{M,n}J_{M,n}^{-1} \\ J_{M,n}^{-1}C_{M,n}J_n^{-1} & J_{M,n}^{-1}K_{M,n}J_{M,n}^{-1} \end{pmatrix}\right). \quad (3.21)$$

This result constitutes the foundation upon which the FIC-procedure will be built. We have thus established the crucial result that will allow us to do *focused model selection* for our dynamic multinomial Markov Chain models.

3.5 Extension to multiple dyads

The developed theory may be extended to observations $y_{i,t}$ from more than one Markov chain. Assume that there are m independent Markov chains where the observations $y_{i,t}$ conditioned on covariates $x_{i,t}$ are generated by the same true *wide* model in every chain.

Define now the random score vector of the *wide* model as

$$U_n^{(m)} = \frac{1}{\sqrt{m \cdot n}} \frac{\partial \ell^{(m,n)}(\beta_{\text{true}})}{\partial \beta}.$$

Define also the random score vector of the candidate model

$$U_{M,n}^{(m)} = \frac{1}{\sqrt{m \cdot n}} \frac{\partial \ell_M^{(m,n)}(\beta_{M,0,n})}{\partial \beta_M}.$$

Denote the random score vectors of the i'th dyad only as $U_{i,n}$ and $U_{M,i,n}$ for the *wide* model and the candidate model respectively. Denote also the corresponding log-likelihood function for the i'th chain only as $\ell_i^{(m,n)}(\beta_{\text{true}})$ and $\ell_{M,i}^{(m,n)}(\beta_{M,0,n})$. We may then write

$$U_n^{(m)} = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \frac{1}{\sqrt{n}} \frac{\partial \ell_i^{(m,n)}(\beta_{\text{true}})}{\partial \beta} = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} U_{i,n}$$

$$U_{M,n}^{(m)} = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \frac{1}{\sqrt{n}} \frac{\partial \ell_i^{(m,n)}(\beta_{M,0,n})}{\partial \beta_M} = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} U_{M,i,n}$$

In a similar manner, let $J_{n,i}$,$K_{M,i,n}$ and $C_{M,i,n}$ denote the random score variance and covariance matrices of the data from the i'th dyad only. We then have that

$$\begin{aligned} J_n^{(m)} &= \mathrm{Var}_{\mathrm{wide}}\ U_n^{(m)} = \mathrm{Var}_{\mathrm{wide}}\,\{\frac{1}{\sqrt{m}}\sum_{i=1}^{m} U_{i,n}\} \\ &= \frac{1}{m}\sum_{i=1}^{m} \mathrm{Var}_{\mathrm{wide}}\ U_{i,n} = \frac{1}{m}\sum_{i=1}^{m} J_{i,n} \end{aligned}$$

and similarily

$$\begin{aligned} K_{M,n}^{(m)} &= \frac{1}{m}\sum_{i=1}^{m} K_{M,i,n}, \\ C_{M,n}^{(m)} &= \frac{1}{m}\sum_{i=1}^{m} C_{M,i,n}. \end{aligned}$$

From Section 3.3 we know that there for each Markov chain i exists limits J_i, $K_{M,i}$ and $C_{M,i}$, such that for each $i \le m$

$$\begin{aligned} J_{i,n} &\to J_i, \\ K_{i,n} &\to K_{M,i}, \\ C_{M,n} &\to C_{M,i}. \end{aligned}$$

As the covariate distribution $C(x)$ is the same for every chain i, we have that $J_i = J$, $K_{M,i} = K_M$ and that $C_{M,i} = C_M$. Thus

$$\begin{aligned} J_n^{(m)} &\to \frac{1}{m}\sum_{i=1}^{m} J_i = J, \\ K_{M,n}^{(m)} &\to \frac{1}{m}\sum_{i=1}^{m} K_{M,i} = K_M, \\ C_{M,n}^{(m)} &\to \frac{1}{m}\sum_{i=1}^{m} C_{M,i} = C_M. \end{aligned}$$

From Section 3.3 we know that $U_{i,n}$ and $U_{M,i,n}$ are normally distributed in the limit for each $i \le m$. Using the well-known result for the sum of independent normally distributed random variables, we then get the limit distribution

$$\begin{aligned} U_n^{(m)} &\to_d N\big(0, J\big) \\ U_{M,n}^{(m)} &\to_d N\big(0, K_M\big) \end{aligned}$$

With just minor adjustments to the arguments in Section 3.4, we may show that the maximum likelihood estimators $\hat{\beta}$ and β_M of *wide* model and candidate models respectively have approximately joint distribution

$$\begin{pmatrix} \sqrt{m \cdot n}(\hat{\beta} - \beta_{\text{true}}) \\ \sqrt{m \cdot n}(\hat{\beta}_M - \beta_{M,0,n}) \end{pmatrix} \sim N\left(0, \begin{pmatrix} J^{-1} & J^{-1}C_M J_M^{-1} \\ J_M^{-1} C_M J^{-1} & J_M^{-1} K_M J_M^{-1} \end{pmatrix}\right).$$

Thus the developed theory also holds when we have data from multiple Markov chains.

3.6 Testing and Confidence intervals

The approximate large sample normal distributions of $\hat{\mu}$ and $\hat{\mu}_M$ allow testing of $H_{\text{null}} : \mu = \mu_{\text{null}}$ against $H_{\text{a}} : \mu \neq \mu_{\text{null}}$ using the *Wald statistic*.

Define

$$SE_{\text{wide}} = (\hat{c}^{\text{t}} \hat{J}_n^{-1} \hat{c})^{\frac{1}{2}}$$

and

$$SE_M = (\hat{c}_M^{\text{t}} \hat{J}_{M,n}^{-1} \hat{K}_{M,n} \hat{J}_{M,n}^{-1} \hat{c}_M)^{\frac{1}{2}}$$

For the correctly specified wide model, the Wald test statistic for testing $H_0 : \mu_{\text{true}} = \mu_{\text{null}}$ is

$$Z = \frac{(\hat{\mu} - \mu_{\text{null}})}{SE_{\text{wide}}}$$

which is approximately standard normally distributed.

For a candidate model misspecified under the *wide* model, the Wald statistic for testing $H_0 : \mu_{M,0} = \mu_{\text{null}}$ is

$$Z_M = \frac{\hat{\mu}_M - \mu_{\text{null}}}{SE_M}$$

which is approximately standard normally distributed. See White (1982) for this misspecification result.

The approximate standard normal distributions of the Wald statistic allows construction of approximate pointwise confidence intervals for the focus parameters. Let z_α denote the $(1 - \alpha)$ quantile of the standard normal distribution. A $(100 - \alpha)\%$ confidence interval based on the Wald statistic is then for the *wide* model

$$\hat{\mu} \pm z_{\frac{\alpha}{2}}\, SE_{\text{wide}}.$$

For the misspecified candidate model, a $(100 - \alpha)\%$ confidence interval based on the Wald statistic is

$$\hat{\mu}_M \pm z_{\frac{\alpha}{2}}\, SE_M$$

These results for pointwise confidence intervals enable plots of focus parameters with pointwise confidence bands.

CHAPTER 4

The Focused Information Criterion for the Dynamic Multinomial Logit Model

The *Focused Information Criterion* aims at selecting the model which best estimates a parameter μ of interest, the *focus parameter.* This will be the model with the lowest *mean squared error* of the estimator $\hat{\mu}$. The FIC score of a model is defined as

$$\mathrm{fic} = \widehat{\mathrm{mse}}\,\hat{\mu} = \widehat{\mathrm{Var}}\,\hat{\mu} + \widehat{\mathrm{bsq}}$$

where bsq stands for squared bias.

To be able to calculate the fic score, a true distribution needs to be assumed. This true distribution is chosen in different ways in the two versions of the FIC. The first version of the FIC, formulated in Claeskens and Hjort (2008b), is situated in a local misspecification context. Here, the true model is such that it is only $O(n^{-\frac{1}{2}})$ away from a narrow model. In the second version of the FIC, developed in Jullum and Hjort (2017) and Cunen, Walløe, and Hjort (2019), the true model is considered to be fixed and can be at any distance from all candidate models.

In this chapter I develop a FIC for dynamic multinomial logit models. This FIC will be of the second version, with a fixed true model. The proceedings in this chapter follow Cunen, Walløe, and Hjort (2019) closely.

The model that in this chapter will play the role as the fixed true model is the *wide* model as defined in (2.2). This model will be used to estimate true parameters. Reflection on which which covariates to include is therefore important. The performance of the FIC will depend on how close this model is to the actual data-generating mechanism, so any covariates that are likely to have an effect on the response variable should be included. Ideally, this should be done before analyzing the data. The reflection requires therefore some prior knowledge of the phenomenon to be analyzed. Often, this is subjectspecific knowledge the statistician do not possess. The choice of *wide* model should therefore be done with guidance from researchers in the field of analysis. In the case of conflict modeling, these are the researchers of international relations.

Candidate models are listed $M_1 \dots M_r$. In the FIC with a fixed *wide* model, these can be at any distance from the *wide* model. The candidate models do not need to be submodels of the *wide* model, but can be model on the form of (2.3). They may include all, less or even more covariates than the *wide* model. For computational feasability, the list of candidate models should not be too long. It should only include relevant models, or models that are of special interest in some other respect.

In this chapter, I describe the FIC procedure of the dynamic multinomial logit model in more detail. In section 4.1 I describe the distribution of the maximum likelihood estimators of the *focus parameter* μ. In section 4.2 I develop the FIC for a single focus parameter. In section 4.3 I extend the selection mechanism to a range of focus parameters, a selection mechanism called the Average-FIC, or the AFIC. In section 4.4 I present an estimation strategy for the FIC and the AFIC.

4.1 The Focus Parameter

The *Focused Information Criterion* aims at selecting the model which gives the most precise estimate of the *focus parameter* μ. Before launching the FIC machinery therefore, the focus parameter needs to be chosen.

There is a wide range of parameters that can be chosen to be focus parameters. The focus parameter can be a mean, a quantile, a model parameter β or something very different. It needs however to be a parameter that has the same interpretation across all candidate models. If this is not the case, one is doing the famous comparison of apples and pears. The focus parameter must also be a smooth function $\mu_{\text{wide}} = \mu_{\text{wide}}(\beta)$ in the *wide* model, and $\mu_M = \mu(\beta_M)$ in the candidate model.

The focus parameter chosen should be the parameter that reflects the research question at best. Practitioners should spend some time thinking about which parameter this is. After all, the FIC is only focused with regard to the purpose of the analysis. If the purpose of the analysis is unclear, or if the focus parameter is wrongly chosen, the selection procedure will be out of focus from the beginning.

In accordance with Cunen, Walløe, and Hjort (2019), we introduce the column vectors

$$\begin{aligned} c &= \frac{\partial \mu(\beta_{\text{true}})}{\partial \beta}, \\ c_M &= \frac{\partial \mu_M(\beta_{\text{M,n,0}})}{\partial \beta_M}. \end{aligned}$$

We also let $\mu_{\text{true}} = \mu_{\text{wide}}(\beta_{\text{true}})$ denote the *true* focus parameter value and we let $\mu_{M,n,0} = \mu_M(\beta_{M,0,n})$ the least false parameter value of the candidate model based on given covariates $x_0, x_1, \dots x_n$.

As we have from (3.21) that the model parameters β and β_M have an approximate joint normal distribution, we get via delta method arguments, as in Cunen, Walløe, and Hjort (2019), that the maximum likelihood estimators $\hat{\mu}_{\text{wide}} = \mu(\hat{\beta})$ of the *wide* model and $\hat{\mu}_M = \mu_M(\hat{\beta}_M)$ of the candidate model have approximate joint normal distribution too:

$$\begin{pmatrix} \sqrt{n}(\hat{\mu}_{\text{wide}} - \mu_{\text{true}}) \\ \sqrt{n}(\hat{\mu}_M - \mu_{M,0,n}) \end{pmatrix} \approx_{\text{d}} \text{N}\left(0, \begin{pmatrix} \nu_{wide} & \nu_{M,c} \\ \nu_{M,c} & \nu_M \end{pmatrix}\right), \tag{4.1}$$

where

$$\begin{aligned} \nu_{\text{wide}} &= c^{\text{t}} J_n^{-1} c, \\ \nu_{M,c} &= c^{\text{t}} J_n^{-1} C_{M,n} J_{M,n}^{-1} c_M, \\ \nu_M &= c_M^{\text{t}} J_{M,n}^{-1} K_{M,n} J_{M,n}^{-1} c_M. \end{aligned} \tag{4.2}$$

From this joint distribution, we see that estimators $\hat{\mu}_{\text{wide}}$ and $\hat{\mu_M}$ are consistent, both tending to the least false parameter, which in the case of the *wide* model is the true parameter value μ_{true}. Thus the estimator $\hat{\mu}_{\text{wide}}$ is unbiased. The least false parameter $\beta_{M,0,n}$ may be different from β_{true}, thus the estimator $\hat{\mu}_M$ of the candidate model is *biased*. The variance $\nu_M/\sqrt{n}$ of $\hat{\mu}_M$ may however be smaller than the variance $\nu_{\text{wide}}/\sqrt{n}$ of $\hat{\mu}$. Thus, even if candidate models are biased it may nevertheless be the case that this candidate model gives the most precise estimate of the focus parameter μ.

4.2 FIC

The *Focused Information Criterion* ranks models according to the estimated mean squared error of the focus parameter estimator. As described above, the fic score is defined to be the estimated mean squared error of the focus parameter estimator.

The mean squared error of an estimator $\hat{\theta}$ is given by

$$\begin{aligned} \text{mse} &= \text{Var}_{\text{wide}}\ \hat{\theta} + (\text{E}_{\text{wide}}\ \hat{\theta} - \theta)^2 \\ &= \text{Var}_{\text{wide}}\ \hat{\theta} + \text{bsq} \end{aligned}$$

The fic score of the unbiased wide model is therefore

$$\text{fic}_{\text{wide}} = \widehat{\text{Var}}_{\text{wide}} \hat{\mu} = \frac{\hat{\nu}_{\text{wide}}}{n}.$$

The candidate estimator of a biased candidate model is

$$\begin{aligned}\mathrm{fic}_M &= \widehat{\mathrm{Var}}_{\mathrm{wide}}\hat{\mu}_M + \max\Big\{\widehat{\mathrm{bsq}}_M, 0\Big\}\\ &= \frac{\hat{\nu}_M}{n} + \max\Big\{\widehat{\mathrm{bsq}}_M, 0\Big\}\end{aligned}$$

It may seem strange that we truncate the squared bias term to zero. The squared bias must necessarily be positive. The estimate $\widehat{\mathrm{bsq}}$ on the other hand, may take negative values. To avoid negative estimates of a positive quantity we choose to truncate the estimates to zero. There is no agreement on this procedure in the literature, however. Jullum and Hjort (2017) truncate $\widehat{bsq}$ to zero as we do. Cunen, Walløe, and Hjort (2019)argue that one should allow the estimated squared bias to take on negative values, as this may enhance the practical performance of the FIC. This really depends on the model class, however. Simulations of the FIC selection with the dynamic multinomial model show that the performance of the FIC is slightly better when the estimated bias squared terms are truncated to zero.

The model with the lowest fic score is the model deemed to give the most precise estimate of the focus parameter and is therefore the model selected by the FIC selection mechanism. At a first glance, it may seem that this should be the *wide* model, as the *wide* model does not have a squared bias term. As the *wide* model has a high number of parameters however, the variance of $\hat{\mu}$ may be too high. There may be simpler models that are somewhat biased, but that have so much lower variance of $\hat{\mu}_M$ that the resulting mean squared error of $\hat{\mu}_M$ will be lower than the mean squared error of $\hat{\mu}$ in the *wide* model. The FIC chooses the model with the most precise parameter estimator and this is the model which is deemed to strike the best balance between bias and variance in the estimation of the focus parameter.

To visualize the selection procedure, one may plot the results in a *FIC-plot* in the same manner as Cunen, Walløe, and Hjort (2019). Such a plot consists of the points

$$(\mathrm{fic}_M^{1/2}, \hat{\mu}_M)$$

The model with the lowest fic value will be the model which is the furthest to the left in the plot. The root FIC score is the preferred value on the x-axis, as this value is on the same scale as the estimates $\hat{\mu}_M$.

4.3 AFIC

The FIC procedure is valid for any focus parameter that fulfills the conditions stated discussed in Section 4.1. The criterion thus allows for model selection with regards to many different purposes. Often, however, what we want to estimate is not a single focus parameter, but a set of more or less different focus parameters. An extension of the FIC procedure to such situations is the *Average Weighted Focused Information Criterion*, or AFIC for short, as introduced in Claeskens and Hjort (2008b) and Jullum and Hjort (2017).

Consider a list of focus parameters $\{\mu_l\}$, where l is in some index set. Assume that each focus parameter μ_l in the set qualifies as a focus parameter in its own right. For each l there is an estimtor $\hat{\mu}_{\text{wide},l}$ of the *wide* model that aims at the true value μ_l. There is also for each l an estimator $\hat{\mu}_{M,l}$ of candidate models that aims at $\hat{\mu}_{M,l,0,n}$.

To quantify the overall risk of all parameters μ_l in the set, consider the risk function

$$L = \int (\hat{\mu}_l(\beta) - \mu_l)^2 \mathrm{d}W(l).$$

Here the function $W(l)$ is a cumulative weight function which reflects the relative importance of the different focus parameters in the set. Possible choices of this cumulative weight function is discussed in Claeskens and Hjort (2008b).

The expected aggregated risk under the wide model is the weighted mean squared error

$$\mathrm{E}_{\text{wide}}\,(L) = \int \mathrm{E}_{\text{wide}}\,(\hat{\mu}_l(\beta) - \mu_l)^2 \mathrm{d}W(l) = \int \mathrm{mse}_l \mathrm{d}W(l).$$

The AFIC-method selects the model with the lowest weighted mean squared error. In parallel with the fic score Claeskens and Hjort (2008b) and Jullum and Hjort (2017) define the afic-score. The afic score of the *wide* model is

$$\mathrm{afic}_{\text{wide}} = \int \frac{\hat{\nu}_{l,\text{wide}}}{n} \mathrm{d}W(t).$$

The afic score of the candidate model is

$$\mathrm{afic}_{\mathrm{M}} = \int \left(\widehat{\mathrm{bsq}}_{l,M} + \max\left\{ \frac{\hat{\nu}_{l,M}}{n}, 0 \right\} \right) \mathrm{d}W(l).$$

In the case of the dynamic multinomial logit model, which is the model discussed in this thesis, the AFIC procedure will often be the preferred selection mechanism. In cases where the focus parameter is a function of covariates, such that $\mu = \mu(x, \beta)$, what is of interest is typically not to find the best estimate of μ for a single covariate value x^*. Rather, what is of interest is typically to find the best estimate of μ value over a subset $X \subseteq \Gamma$ of covariate values. The focus parameter is in this case rather the aggregated focus parameter

$$\mu_X = \int_X \mu(x) \mathrm{d}C(x)$$

As we know nothing about the covariate distribution $C(x)$, the way to select models with regards to this aggregated focus parameter is to use the AFIC with the empirical distribution of covariates as weight function, $W(t)$. In the case where $X = \Gamma$, afic-scores are in this case given by

$$\begin{aligned} \mathrm{afic}_{\mathrm{wide}} &= \sum_{i=1}^{m}\sum_{t=1}^{n}\Big\{\frac{\hat{\nu}_{\mathrm{wide}(x_{i,t})}}{n}\Big\} \\ \mathrm{afic}_{\mathrm{M}} &= \sum_{i=1}^{m}\sum_{t=1}^{n}\Big\{\frac{\hat{\nu}_{M}(x_{i,t})}{n} + \max\Big\{\widehat{\mathrm{bsq}}_{M}(x_{i,t}), 0\Big\}\Big\}. \end{aligned}$$

for the *wide* model and candidate models respectively. In the cases where X is a proper subset of Γ we sum only over covariate values that are elements in this subset X.

To visualize the AFIC selection procedure, I invent the *AFIC-plot*. This plot parallels the FIC plot in that it consists of the points

$$(\mathrm{afic}_M^{1/2}, \hat{\mu}_{l,M})$$

where l is a particular element in the index set. Only one estimate $\hat{mu}_{l,M}$ can be chosen among the list of focus parameters in the set. The AFIC plot will allow visualization of the AFIC ordering of the models, but the reader should keep in mind that the vertical ordering of estimates may be different if another focus parameter in the list were chosen for plotting. The important point nevertheless that the horizontal ordering is the same for all parameters in the list. Regardless of which l is choosen, the model furthest to the left in the AFIC-plot is the model selected by the AFIC.

4.4 FIC score and Estimation

To calculate FIC and AFIC scores, the mean squared errors of the focus parameters in the wide model and in candidate models need to be estimated. This involves estimating squared bias as well as quantities $\nu_{\mathrm{wide,n}}$, $\nu_{M,n}$ and $\nu_{M,c,n}$. From the expressions in (4.2) we see that these quantities are functions of random score variances J_n, $J_{M,n}$, $K_{M,n}$ and $C_{M,n}$. Calculation of FIC and AFIC scores therefore first requires estimation of these matrices. Formulas for J_n, $J_{M,n}$, $C_{M,n}$ and $K_{M,n}$ are given in Section 3.2 and in Appendix A.4. A glance at these formulas reveals quickly that this estimation task will not be easy. In this section, I propose an estimation strategy that will make this estimation possible.

Notice first that J_n is a function of $\beta_{\mathrm{true},n}$ and that $J_{M,n}$, $C_{M,n}$ and $K_{M,n}$ are functions of both $\beta_{\mathrm{true},n}$ and $\beta_{M,n,0}$. ML estimates of these matrices are then

$$\begin{aligned}
\hat{J}_n &= J_n(\hat{\beta}), \\
\hat{J}_{M,n} &= J_{M,n}(\hat{\beta}, \hat{\beta}_M), \\
\hat{C}_{M,n} &= C_{M,n}(\hat{\beta}, \hat{\beta}_M), \\
\hat{K}_{M,n} &= K_{M,n}(\hat{\beta}, \hat{\beta}_M),
\end{aligned}$$

which imply that estimation of these matrices amounts to plugging in

$$\hat{\pi}_{kj}(x_{i,t}) = \pi_{kj}(x_{i,t}, \hat{\beta}) \text{ for } \pi_{kj}(x_{i,t}),$$

and

$$\hat{\pi}_{Mkj}(x_{i,t}) = \pi_{Mkj}(x_{i,t}, \hat{\beta}_M) \text{ for } \pi_{Mkj}(x_{i,t}),$$

and

$$\hat{\mathbf{P}}(t) = \hat{\mathbf{P}}(t|\hat{\beta}, \hat{\beta}_M) \text{ for } \mathbf{P}(t).$$

Under the mild assumption that the chain moderately fast finds it equilibrium distribution, a moderately high integer S may be chosen so that $\mathrm{E}_{\text{wide}}\, y_{itj}$ may be estimated with

$$\hat{\mathrm{E}}_{\text{wide}}(y_{itj}) = \begin{cases} \hat{P}^{(S)}_{kj}(t-S) & \text{if } t \geq S \\ \sum_{k=0}^{2} \hat{P}^{(t)}_{kj}(0) & \text{if } 1 \leq t < S \end{cases}$$

These estimators may be plugged in in the formulas given in Section 3.2 for the orderly matrices $\hat{J}_n$, $\hat{J}_{M,n}$ and $C_{M,n}$. With this plug-in procedure, these matrices may be cumbersome, but not too difficult to calculate.

The calculation of the $\hat{K}_{M,n}$ matrix is a different story. This matrix will consist of a huge number of terms as it takes into account the correlation between every observation in the chain due to misspecification. Recall from (3.6) that the $K_{M,n}$ matrix consists of four matrices $J^*_{M,n}, V_n, W_n$ and Q_n. Each of these matrices may be estimated in turn. The $J^*_{M,n}$ matrix is the easiest one. This matrix has the same structure as J_n and $J_{M,n}$ and should not cause any problems. The matrices V_n, Q_n and W_n are the troublemakers. If the Markov chain is moderately long, these matrices have an astronomical number of terms. Taking them all into account will be computationally impossible.

Fortunately, there is a way around this brutal problem. By close inspection of the formulas for these matrices in Appendix A.4, we see that elements of these matrices are given by

$$v_n = \sum_{i=1}^{m}\sum_{t=1}^{n}\sum_{w=0}^{t-2}\sum_{q,r=0}^{1} \star\, \Phi_{rk}^{(w)}(t-1)\Phi_{qk'}^{(w)}(t-1),$$

$$q_n = \sum_{i=1}^{m}\sum_{t=1}^{n}\sum_{s=0}^{t-1}\sum_{r=0}^{1} \star\, \Phi_{rk}^{(s-1)}(t-1),$$

$$w_n = \sum_{i=1}^{m}\sum_{t=1}^{n}\sum_{s=0}^{t-2}\sum_{w=0}^{t-s-2}\sum_{r,q=0}^{1} \star\, \Phi_{rk}^{(s+w)}(t-1)\Phi_{qk'}^{(w)}(t-s),$$

where v_n denotes any element in V_n, q_n denotes any element in Q_n and w_n denotes any element in W_n. The parameter $\Phi_{kj}(s)(t)$ denotes $|P_{kj}^{(s)}(t) - P_{2j}^{(s)}(t)|$ as described in Section 2.4, and $\star$ denotes the remaining parts of the expressions, which is not important for detecting the convergence structure.

Under the additional mild assumptions that the Markov chain moderately fast finds its equilibrium distribution, we may choose a moderately large integer S, such that for each $s \geq S$ it is the case that

$$P_{0j}^{(s)}(t) \approx P_{1j}^{(s)}(t) \approx P_{2j}^{(s)}(t) \approx P_j(t),$$

which is equivalent to

$$\Phi_{rj}^{(s)}(t) = \left| P_{r_j}^{(s)}(t) - P_{2j}^{(s)}(t) \right| \approx 0.$$

Under this assumption, we may find also a small integer L such that V_n, Q_n and W_n may be estimated by only calculating the terms $w < L$ and $s < S$. This will in no way make the estimation process *easy*, but at least it will make it computationally feasible. The strategy is so much the better as the contributions to $K_{M,n}$ from V_n, Q_n and W_n will be small compared to the contribution from $J_{M,n}^*$. Testing with different integers L and S shows that this strategy is satisfying, as the difference between $\hat{K}_{M,n}$ estimated with L and S terms and $\hat{K}_{M,n}$ estimated with $2L$ and $2S$ terms is microscopic when L and S are properly chosen.

The ML estimators of vectors c and c_M are

$$\hat{c} = \partial\mu(\hat{\beta})/\partial\beta,$$

$$\hat{c}_M = \partial\mu_M(\hat{\beta}_M)/\partial\beta_M.$$

If we have expressions for the partial derivative of the focus parameter μ, we may calculate the estimates by the plug-in of ML-estimates in these expressions. A quicker method is to calculate these estimates numerically. We may, for example, use the `numDeriv()`-function in R to do this. This is also a reliable method in the cases where we have no analytical expression of the partial derivatives.

With estimates of variance matrices of random score vectors and partial derivatives on board we then find ML estimates of $\nu_{\text{wide},n}$, $\hat{\nu}_{M,n}$ and $\hat{\nu}_{C,M,n}$ by plugging in the estimated values described above in (4.2) such that

$$\begin{aligned}
\hat{\nu}_{\text{wide}} &= \hat{c}^{'}\hat{J}_n^{-1}\hat{c}, \\
\hat{\nu}_{M,c} &= \hat{c}^{'}\hat{J}_n^{-1}\hat{C}_{M,n}\hat{J}_{M,n}^{-1}\hat{c}_M, \\
\hat{\nu}_M &= \hat{c}_M^{'}\hat{J}_{M,n}^{-1}\hat{K}_{M,n}\hat{J}_{M,n}^{-1}\hat{c}_M.
\end{aligned}$$

Finally, the squared bias needs to be estimated. We have that a consistent estimator of the bias of a candidate model is $\hat{b}_n = \hat{\mu}_{M,n} - \hat{\mu}$. We could then easily have been led to estimate squared bias by $\hat{b}_n^2$. This will not constitute a consistent estimator, however. We see from

$$\mathrm{E}\ \hat{b}_{M,n}^2 = (\mathrm{E}\ \hat{b}_{M,n})^2 + \mathrm{Var}\ \hat{b}_{M,n}$$

that this estimator tends to overestimate the true squared bias. From Cunen, Walløe, and Hjort (2019) we have that a consistent estimator of the squared bias is

$$\begin{aligned}
\widehat{\text{bsq}}_M &= (\hat{\mu}_M - \hat{\mu})^2 - \widehat{\text{Var}}\,\hat{b}_M, \\
&= (\hat{\mu}_M - \hat{\mu})^2 - \frac{1}{n}(\hat{\nu}_{\text{wide}} + \hat{\nu}_M - 2\hat{\nu}_{M,c}).
\end{aligned}$$

From this expression, we see that the estimated squared bias may be zero, as discussed above.

CHAPTER 5

Simulations

In this chapter, I illustrate the developed methodology by simulation studies. It will be shown that the simulations confirm the developed methodology.

I start out in Section 5.1 by simulating Markov chains from a *wide* dynamic multinomial logit model. For each simulation round, I fit a *wide* model and a candidate model to the simulated data and I retrieve maximum likelihood estimators $\hat{\beta}$ and $\hat{\beta}_M$ for each round. These simulated maximum likelihood estimators are shown to be approximately normally distributed with the appropriate standard deviations.

In the subsequent sections, I do focused model selection on simulated Markov chains. In Section 5.2 I describe the general setup of these FIC simulations. I show then that the FIC indeed aims at selecting the model with the lowest *true* mean squared error of the focus parameter. I illustrate this result for three different classes of focus parameters. In Section 5.3 I simulate the FIC procedure for the transition matrix $\pi_{12}(x)$. In Section 5.4 I simulate the FIC procedure for the more involved parameter f_{12}, a parameter that I describe more closely in this section. In Section 5.5 I illustrate the AFIC procedure by choosing as focus parameter the effect of the r'th covariate when the past level of the chain is 1. This aggregated focus parameter is constituted by the parameters $\beta_{1,0,r}$ and $\beta_{1,1,r}$.

5.1 Normality of Simulated Maximum Likelihood Estimators

The first simulation study aims to illustrate the normal distribution of estimators $\hat{\beta}_n$ and $\hat{\beta}_{M,n}$ of the wide model and candidate models respectively.

I consider the case where given covariate vectors are on the simple form $x_{i,t} = (1, x_{i,t,1})^{\mathrm{t}}$. For each $i \leq m$ we draw a start value $x_{1,i,0}$ from a uniform distribution on $[0, 1]$. For each $0 < t \leq n$ I then draw $x_{i,t,1}$ from a normal distribution with mean $x_{1,i,t-1}$.

Conditioned on these simulated covariate values, I let the true distribution of response variables $y_{i,t}$ be a *wide* model on the form (2.2). I let the true parameter values be

$$
\begin{aligned}
\beta_{00} &= (0.3, 0.2)^{\mathrm{t}},\\
\beta_{01} &= (0.25, 0.1)^{\mathrm{t}},\\
\beta_{10} &= (0.11, 0.01)^{\mathrm{t}},\\
\beta_{11} &= (0.09, 0.03)^{\mathrm{t}},\\
\beta_{20} &= (0.1, 0.14)^{\mathrm{t}},\\
\beta_{21} &= (0.08, 0.05)^{\mathrm{t}}.
\end{aligned}
$$

I consider also a candidate model. This will be a dynamic multinomial logit model on the form of (2.3) with $u_{i,t} = x_{i,t,1}$ and $z_{i,t} = 1$. From this we see that it is a model where the effect of the covariate value $x_{i,t,1}$ is independent of past observation level $y_{i,t-1}$ in the chain, but where the intercepts of the model are dependent on past observation level.

I set the number of simulated rounds to $\texttt{sim} = 10^4$. To reduce the running time of each simulation round, I make an assumption cruder than the one described in Section 4.4. Recall that the integers L and S decide the number of terms we take into consideration when calculating the $\hat{K}_{M,n}$ matrix. To reduce the running time of each round I let $L = 0$ and $S = 40$. A few simulation rounds where L and S were chosen considerably higher show that the crude approximation is without any practical consequences in this simulation setup.

I consider first the situations with observation y_t from a single, long Markov chain. I set the number of chains to $m = 1$ and the number of subsequent observations in the chain to $n = 5000$. Simulating 10^4 such chains from the *true* distribution, I fit for each chain the *wide* model and the candidate model using the `nlm`-algorithm in R. The simulation procedure then results in a list of ten thousand vectors $\hat{\beta}$ and a list of ten thousand vectors $\hat{\beta}_M$. For each round of simulations, I calculate estimates $\hat{J}_n$ and $K_{M,n}$ as well as the estimated candidate Information matrix $\hat{J}_{M,n}$. To avoid computation overload I hold off on the calculation of $\hat{K}_{M,n}$.

I first present the results for the simulated $\hat{\beta}$ parameter of the *wide* model. According to (3.21), these estimated parameter values are approximately normally distributed with standard deviations $n^{-\frac{1}{2}}\mathrm{diag}(J_n^{-1})$, where J_n is the true information matrix as defined in Section 2.3. For each $k = 0, 1, 2$, each $j = 0, 1$ and each $r = 1, \ldots p$, it is then the case that the standardized parameter estimates are approximately normally distributed such that

$$\frac{\sqrt{n}(\hat{\beta}_{k,j,r} - \beta_{k,j,r})}{\sigma_{k,j,r}} \approx_d N(0,1).$$

Here $\beta_{k,j,r}$ is an element in β_{true}, $\hat{\beta}_{k,j,r}$ is the corresponding element in $\hat{\beta}$ and $\sigma_{k,j,r}$ is the root of the corresponding element in $\mathrm{diag}\{J_n^{-1}\}$.

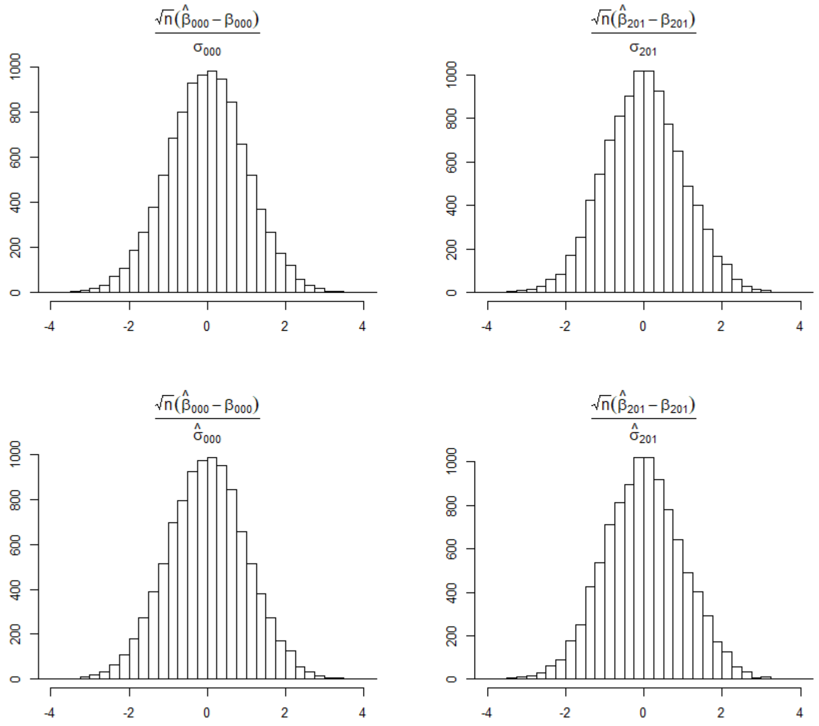

Figure 5.1: *Histograms of standardized simulated maximum likelihood estimates of the wide model in the case of data from one Markov chain of length 5000. The two upper panels use true standard deviations for standardization. The two lover panels use estimated standard deviations.*

In the upper panel of Figure 5.1 are presented histograms of standardized versions of the simulated $\hat{\beta}_{000}$ and $\hat{\beta}_{201}$. We see that both histograms have the form of normal distributions. The situation for the remaining ten parameter estimates is similar. This is a clear indication that the maximum likelihood estimates really are normally distributed.

We need to examine that the estimates $\hat{\beta}$ are distributed with the right variances. As each simulation round is independent, it is the case that the means of the simulated parameter values are distributed according to

$$Z = \sqrt{\text{sim}}\frac{\sqrt{n}(\sum_{i=1}^{\text{sim}} \hat{\beta}_{i,k,j,r} - \beta_{k,j,r})}{\sigma_{k,j,r}} \approx_d N(0,1),$$

In the left plot of Figure 5.2 are plotted the z-values of the mean of the ten thousand simulated maximum likelihood estimators. The dotted lines are

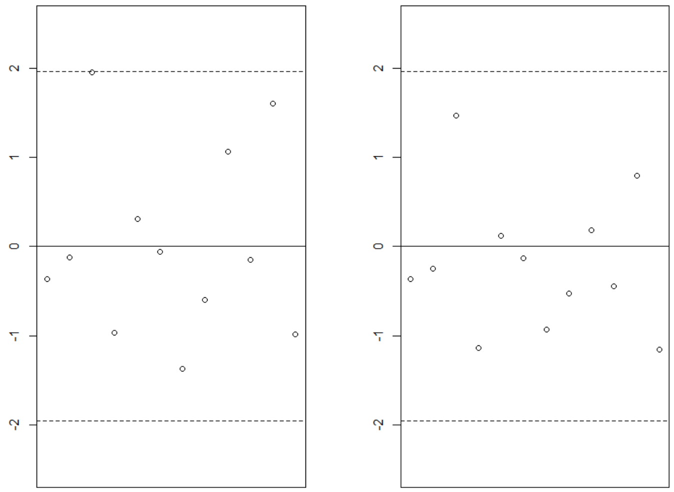

Figure 5.2: *Standardized values of the mean of simulated maximum likelihood estimates of each of the twelve parameters in the wide model. Simulated data are here from one Markov chain at length 5000. In the left panel true standard deviations are used for standardization. In the right panel estimated standard deviations are used. The lower dotted line is the 0.025 quantile. The upper plotted line is the 0.975 quantile.*

the 0.025 quantile and the 0.975 quantile of the standard normal distribution. We see that the z-values of all simulated means are within the 95% confidence bands. We conclude that the simulations show that the maximum likelihood estimates indeed are approximately normally distributed.

It should also be the case that the approximate normal distribution holds with estimated standard deviations. When $\hat{\sigma}_{k,j,r}$ is the root of the corresponding element in diag$\{\hat{J}_n^{-1}\}$, it should for the corresponding parameter in $\hat{\beta}$ and β_{true} be the case that

$$\frac{\sqrt{n}(\hat{\beta}_{k,j,r} - \beta_{k,j,r})}{\hat{\sigma}_{k,j,r}} \sim N(0,1).$$

In the lower panels of Figure 5.1 we see that this indeed is the case for $\hat{\beta}_{000}$ and $\hat{\beta}_{201}$. The situation is parallel for the other ten parameters. The right panel of Figure 5.2 shows the z-values of the mean of the maximum likelihood

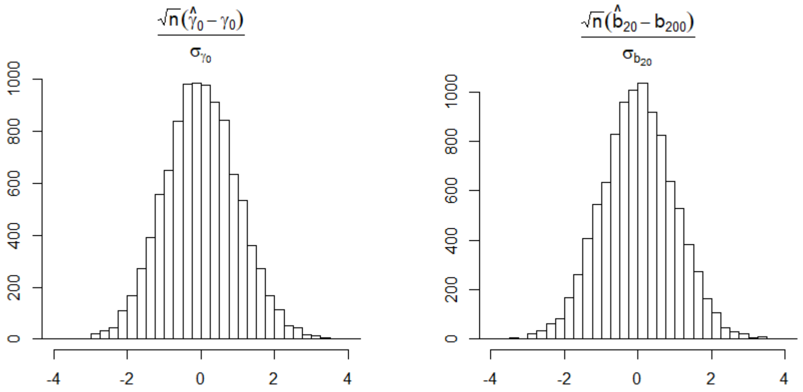

Figure 5.3: *Histograms of standardized simulated maximum likelihood estimates of the candidate model in the case of data from one Markov chain of length 5000*

estimates calculated with estimates $\hat{\sigma}_{k,j,r}$. We see that all standardized values are within the 95% confidence band, as expected.

I consider next the simulated maximum likelihood estimates of the candidate model. In this case, we do not know the 'true' *least false* parameter value $\beta_{M,0}$, so we have used the mean values of the MLE's from another 10^4 simulated chains. We use this estimated least false value to calculate the 'true' least false $K_{M,n}$-matrix. This procedure should be permissible as we know that $\hat{\beta}_M$ is a consistent estimator of $\beta_{M,0}$.

As was the case for the maximum likelihood estimates of the *wide* model, it should be the case that the true standardized values of the simulated candidate parameter estimates are approximately normally distributed, both in the case of Markov independent parameters γ as well as the Markov dependent parameters b. It should therefore be the case that

$$\frac{\sqrt{n}(\hat{\gamma}_{M,j,r} - \gamma_{M,j,r})}{\sigma_{\gamma,j,r}} \approx_d N(0,1),$$
$$\frac{\sqrt{n}(\hat{b}_{M,k,j,r} - b_{M,k,j,r})}{\sigma_{b,k,j,r}} \approx_d N(0,1).$$

Histograms of standardized simulated maximum likelihood estimates $\hat{\gamma}_0$ and $\hat{b}_{20}$ are plotted in Figure 5.3. We see that the form of the histograms in this plot is close to a normal distribution. The situation is parallel for the maximum likelihood estimates of the remaning elements in β_M. The standardized value of the simulated estimates is for an element r now given by

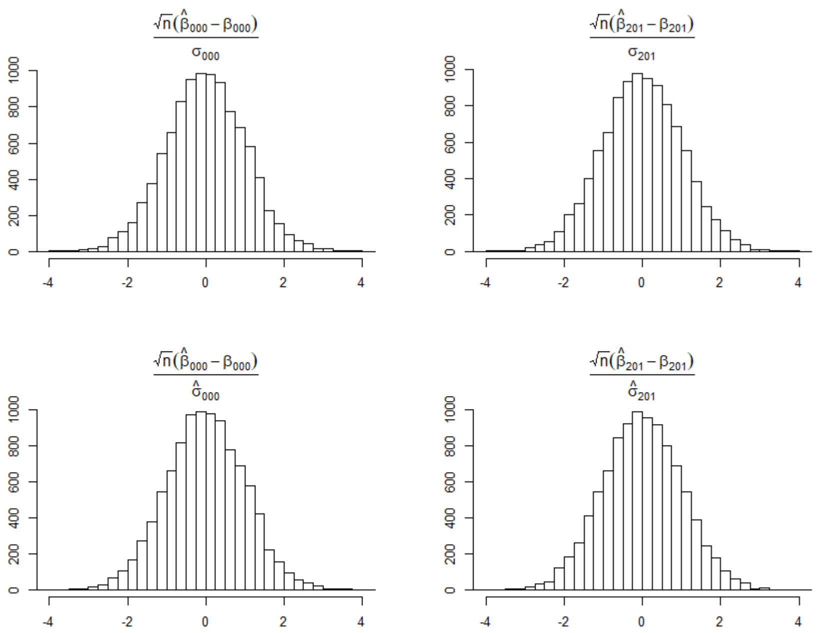

Figure 5.4: *Histograms of standardized simulated maximum likelihood estimates of the wide model in the case of data from 100 Markov chain of length 50. The two upper panels use true standard deviations for standardization. The two lover panels use estimated standard deviations.*

$$\frac{\hat{\beta}_{M,r}-\beta_{M,0,r}}{\sigma_{M,r}},$$

where $\sigma_{M,r}$ is the root of the corresponding element in $\text{diag}\{K_{M,n}\}$.

The empirical standard deviations for these standardized simulated variables are for each of the elements r

$$(0.999, 1.002, 1.014, 1.006, 1.000, 1.011, 0.995, 0.998)^{\text{t}}.$$

All estimated standard deviations are very close to one, indicating that the true variance matrix of random score vector of the candidate model is in fact the $K_{M,n}$ matrix developed in Chapter 3.

The results should be similar for simulated models fitted to data from multiple independent chains, as explained in Section 3.5. To verify that this is the case, we simulate `sim` $= 10^5$ rounds where the number of independent

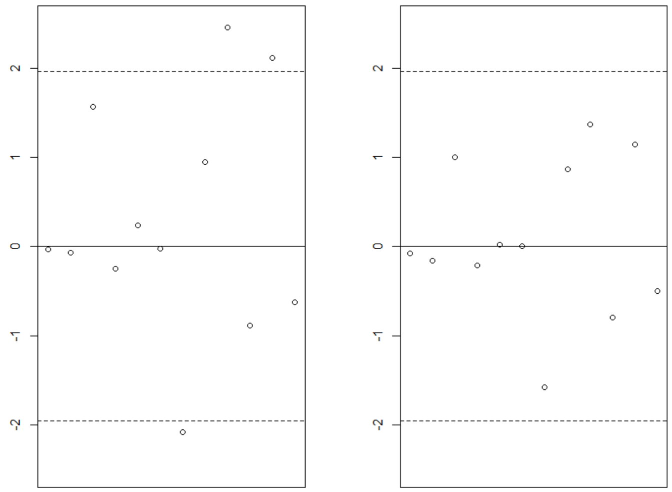

Figure 5.5: *Standardized values of the mean of simulated maximum likelihood estimates of each of the twelve parameters in the wide model. Simulated data are here from 100 Markov chains at length 50. In the left panel true standard deviations are used for standardization. In the right panel estimated standard deviations are used. The lower dotted line is the 0.025 quantile. The upper plotted line is the 0.975 quantile.*

chains is $m = 100$ and the length of each chain is $n = 50$. We plot histograms of the resulting standardized $\hat{\beta}$ values of the *wide* model in Figure 5.4. From these plots, we see that the simulated maximum likelihood estimates of the *wide* model are approximately normally distributed.

In Figure 5.5 we have plotted the z-values of the mean value of simulated maximum likelihood estimates of the *wide* model. In the left panel are plotted the case of true standard deviations. We see in this plot that there are three values that lie somewhat outside the 95%-confidence band. However none of these three z-values are extreme, so I conclude that the simulations verify that the $\hat{\beta}$ parameters are approximately normally distributed with the right variance also in the case of multiple chains.

We also calculate the standardized maximum likelihood estimates of the candidate model in a similar manner as above. Figure 5.6 shows histograms for two of these standardized estimates. We see also in this case that the estimates are approximately normally distributed. The situation is similar for the other elements in $\hat{\beta}_M$. The empirical standard deviations of the standardized

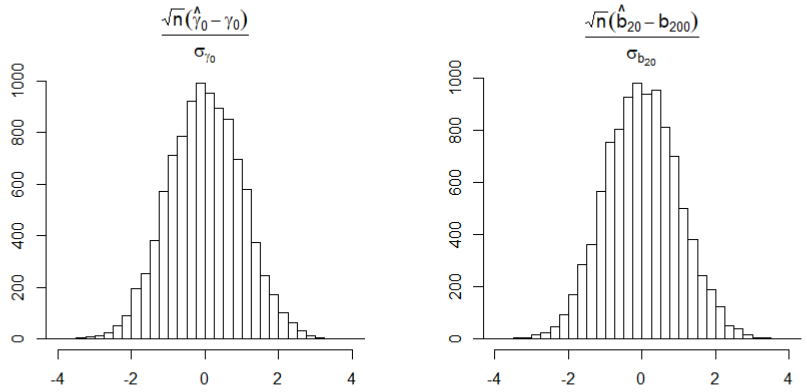

Figure 5.6: *Histograms of standardized simulated maximum likelihood estimates of the candidate model in the case of data from 100 Markov chain of length 50*

simulated parameters are now

$$(0.996, 1.008, 1.002, 1.012, 0.999, 1.005, 0.998, 1.007)^{\mathrm{t}},$$

confirming in this case too, that the variance matrix $K_{M,n}$ of the candidate model is correctly identified in chapter 3.

For this simulation setup, we also consider the case of estimated standard deviations for the candidate model. We should have

$$\frac{\sqrt{n}(\hat{\gamma}_{M,j,r} - \gamma_{M,j,r})}{\hat{\sigma}_{\gamma,j,r}} \sim N(0,1),$$

$$\frac{\sqrt{n}(\hat{b}_{M,k,j,r} - b_{M,k,j,r})}{\hat{\sigma}_{b,k,j,r}} \sim N(0,1).$$

Due to the complexity of the $K_{M,n}$-matrix, the running time of each simulation round is high when also $\hat{K}_{M,n}$ is calculated. I therefore restrict the number of simulations to sim = 1000 in this case. The number of chains is $m = 100$, but the length only $n = 30$. In figure Section 5.1 I have plotted histograms of the ensuing standardized maximum likelihood estimates. There is too few simulations to expect that the histograms have the forms of proper normal distributions. Nevertheless, we see the contours of normal distributions for both of the two selected elements. The situation is parallel for the other elements in $\hat{\beta}_M$. The standard deviations of the normalized variables are in this case

$$(0.996, 0.997, 0.985, 1.016, 0.996, 0.990, 0.974, 1.016)^{\mathrm{t}},$$

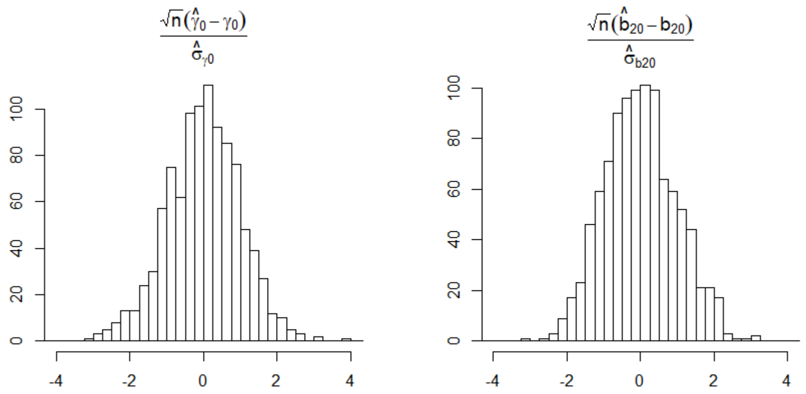

Figure 5.7: *Histograms of standardized simulated maximum likelihood estimates of the candidate model in the case of data from 100 Markov chain of length 30. The standardization is here calculated with estimated standard deviations in each round.*

which is close to the theoretical values of 1, considering that we have only thousand simulation rounds.

I conclude that the simulation study in this section confirms the theory developed in Chapter 3.

5.2 Setup for FIC Simulation

In the rest of this chapter, I simulate to illustrate the FIC machinery developed in Chapter 4. I now let there be $m = 30$ independent dyads, each with $n = 100$ observations. The covariate vectors I now extend to be on the form $x_{i,t} = (1, x_{i,t,1}, x_{i,t,2})^{\mathrm{t}}$. I let $x_{i,t,1}$ be a quantitative variable with values between zero and one, and I let $x_{i,t,2}$ be a qualitative variable which may take values zero or one.

For each $i \leq m$ I draw values $x_{i,t,1}$ from a normal distribution with mean $x_{i,t-1,1}$ in the same manner as in Section 5.1. The values $x_{i,t,2}$ I draw from a Bernoulli distribution, where the probabilities depend on the past value $x_{i,t-1,1}$.

Conditioned on these simulated covariates, I let the *true* distribution of dependent observations $y_{i,t}$ be the *wide* model with parameters

$$
\begin{aligned}
\beta_{00} &= (0.4, 0.2, 1.8)^{\mathrm{t}},\\
\beta_{01} &= (0.4, 0.3, -0.5)^{\mathrm{t}},\\
\beta_{10} &= (0.1, 0.01, -0.41)^{\mathrm{t}},\\
\beta_{11} &= (0.1, 0.03, 0.02)^{\mathrm{t}},\\
\beta_{20} &= (0.2, -0.14, 0.008)^{\mathrm{t}},\\
\beta_{21} &= (-0.2, -0.01, -0.002)^{\mathrm{t}}.
\end{aligned}
$$

This true *wide* model has thus $p = 6 \cdot 3 = 18$ parameters.

I define covariate vector $x_{\text{high}} = (1, 0.70, 1)^{\mathrm{t}}$. At this specific covariate value the transition matrix is

$$
\mathbf{P}(x_{\text{high}}) = \begin{pmatrix} 0.831 & 0.089 & 0.080 \\ 0.256 & 0.398 & 0.346 \\ 0.381 & 0.278 & 0.342 \end{pmatrix}.
$$

Comparing each row, we see that the Markov dependency at this covariate value is considerable: The transition probability from level 0 to level 2 is close to zero, whereas the transition to level 2 from levels 0 or 1, is about 0.34.

I also define $x_{\text{low}} = (1, 0.30, 0)^{\mathrm{t}}$. The transition matrix is for this specific covariate value

$$
\mathbf{P}(x_{\text{low}}) = \begin{pmatrix} 0.376 & 0.387 & 0.237 \\ 0.344 & 0.346 & 0.310 \\ 0.390 & 0.273 & 0.335 \end{pmatrix}
$$

We see that transition probabilities now are more even at different rows. At x_{low} the Markov dependency is therefore smaller. Thus for the *inhomogeneous* Markov chain generated by the *wide* model, the degree of Markov dependency is a function of where one is situated in the space of covariates Γ.

For data simulated from this true *wide* model, I fit the following list of models:

- M_0: $u_{i,t} = 1$ $z_{i,t} = (x_{1,i,t}, x_{2,i,t})^{\mathrm{t}}$
- M_1: $u_{i,t} = 1$ $z_{i,t} = (x_{1,i,t}, x_{2,i,t})^{\mathrm{t}}$
- M_2: $u_{i,t} = 1$ $z_{i,t} = x_{1,i,t}$
- M_3: $u_{i,t} = (x_{1,i,t}, x_{2,i,t})^{\mathrm{t}}$ $z_{i,t} = 1$
- M_4: $u_{i,t} = (1, x_{1,i,t}, x_{2,i,t})^{\mathrm{t}}$ $z_{i,t} = \emptyset$

The vector $u_{i,t}$ here denotes covariates corresponding to effects that are Markov independent, whereas $z_{i,t}$ denotes covariates corresponding to effects that are Markov dependent, as explained in Chapter 2. The first model M_0 is the *wide* model, the last model M_4 is the standard multinomial model. The remaining three candidate models are intermediate models with different combinations of Markov-dependent and Markov-independent effects.

For a specific focus parameter μ the true value $\mu_{\text{true}} = \mu_{\text{wide}}(\beta_{\text{true}})$ and the true fic value of the *wide* model, are found by directly inserting the known value β_{true} into the relevant expressions.

To find the corresponding values $\mu_M = \mu_M(\beta_{M,0})$ and true fic values for each of the candidate models, I simulate 750 rounds chains $\{y_{i,t}\}$ from the true, *wide* model. For each simulation round, I fit candidate models $M_1, \dots M_4$ and get a list of maximum likelihood estimates $\hat{\beta}_M$ for each of the candidate models. The 'true' $\beta_{M,0}$ values of each candidate model I estimate by taking the mean of the simulated maximum likelihood estimates $\hat{\beta}_M$ of the model. I then find 'true' values of μ_M and the true fic value of each candidate model by insertion of this 'true' least false parameter value in the relevant expressions.

Having found for each candidate model the least false values of the focus parameter and the fic value, I then simulate $m = 100$ chains of length $n = 30$ for $\text{sim} = 100$ rounds. For each round, I fit the *wide* model and candidate models to the simulated data and calculate corresponding $\text{fic} = \widehat{\text{mse}}$ values. In the end I get a list of hundred fic values for each candidate model.

5.3 First Focus Parameter: A Transition Probability

In this section, I illustrate the FIC apparatus for the focus parameter $\pi_{12}(x)$. I simulate $\text{sim} = 100$ rounds and proceed as described above in Section 5.2.

I first take $\pi_{12}(x_{\text{high}})$ as focus parameter. As we saw above, this is a covariate value for which the *true* matrix $\mathbf{P}(x_{\text{high}})$ shows considerable Markov dependency. Because of this Markov dependency, we would expect the multinomial model to be a bad model for this focus parameter.

The results of the simulations for this focus parameter are shown in Figure 5.8. Inspired by (Cunen, Walløe, & Hjort, 2019) I show the variance part of the mean squared error in the left panel of this figure and the squared bias part of the mean squared error in the right panel of the figure. The red lines show the 'true' values for the models $M_0, \dots, M_4$, where the 'true' values are calculated from β_{true} in the case of the *wide* model, and from the least false values estimated from 750 simulation rounds in the case of the candidate models. The grey crosses are the estimated $\hat{\nu}_M$ and $\widehat{\text{bsq}}_M$ in each of the 100 simulation rounds (the crosses are plotted without truncation to zero). The black line shows the mean values of these 100 estimates, where negative values of $\hat{bsq}_M$ are truncated to zero.

The true root mean squared errors of $\hat{\pi}_{12}(x_{\text{high}})$ are 0.0328, 0.0329, 0.0281, 0.0803, 0.1632 for the models M_0, M_1, M_2, M_3, M_4 respectively. We see that the second candidate model M_2 gives the lowest mean squared error of this estimate.

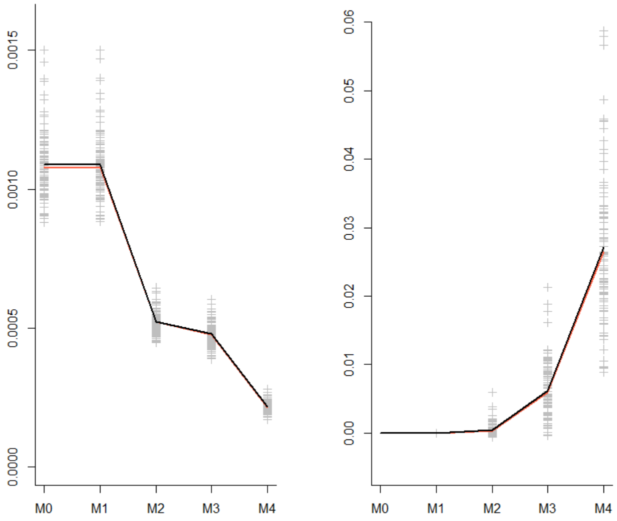

Figure 5.8: *Simulation results for $\mu = \pi_{12}(x_{high})$. The variance part of the mse is given in the left panel. The squared bias part is given in the right panel. The red lines are the true values, the grey crosses are the variance and squared bias part in each of the 100 simulation rounds. The black lines are the average scores of the variance part and the squared bias part.*

The multinomial model M_4 is a particularly bad model as we would expect. The *wide* model M_0 and M_1 are equally good, but somewhat behind the winner. The mean values of the estimated $\sqrt{\text{fic}}$ values are 0.0329,0.0329,0.0307,0.0776,0.1621. The best model M_2 is selected in 70% of the runs. The wide model and $M1$, which are not that far behind, are selected in 11% of the runs, and 17% of the runs respectively. Reassuringly, the multinomial model is never selected. The FIC thus tends to identify the correct model for the focus parameter $\pi_{12}(x_{\text{high}})$.

Next, I consider $\pi_{12}(x_{\text{low}})$ as focus parameter. That is: the same transition probability, but at another covariate value. Recall that at x_{low} the transition matrix $\mathbf{P}(x)$ showed less Markov dependency. We would therefore expect the multinomial model to give a more precise estimate of the focus parameter at this covariate value.

The simulation results for focus parameter $\pi_{12}(x_{\text{low}})$ are shown in Figure 5.9. We see on the scales of the plots in this figure that both variance and squared bias is less for this focus parameter. The true root mean squared errors of $\hat{\pi}_{12}(x_{\text{low}})$ are $0.0183, 0.0165, 0.0128, 0.0315, 0.0232$. The second candidate model M_2 is still the best model. The multinomial model M_4 is somewhat behind,

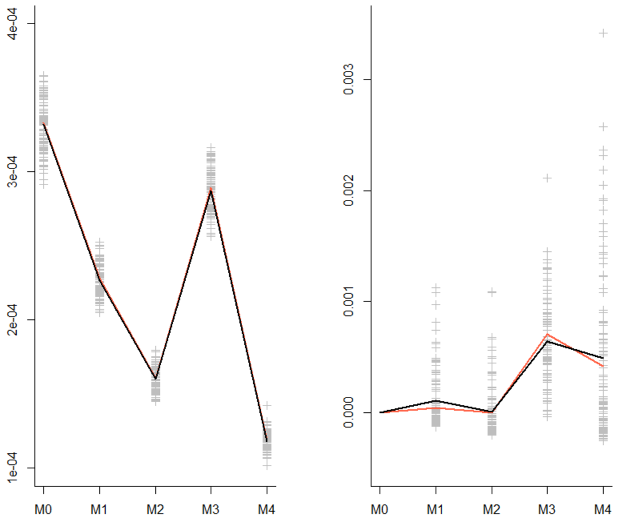

Figure 5.9: *Simulation results for* $\mu = \pi_{12}(x_{low})$*. The variance part of the mse is given in the left panel. The squared bias part is given in the right panel. The red lines are the true values, the grey crosses are the variance and squared bias part in each of the 100 simulation rounds. The black lines are the average scores of the variance part and the squared bias part.*

but not that much worse. The average $\sqrt{\text{fic}}$-values of the hundred simulations are $0.0182, 0.0186, 0.0152, 0.0298, 0.0224$. The second candidate model M_2 is correctly selected in 51% of the runs. The multinomial model is however selected in 36% of the runs. This is due to the very low variance of this model. In the simulation rounds where the multinomial model has zero estimated bias, it will be considered to be the winner. Still, the correct model is selected as the FIC winner in the majority of runs also for this focus parameter.

5.4 Second Focus Parameter: Long Term Probability

I then change the focus parameter. We will find the best estimate of the probability of the chain entering state two before entering state zero when it starts in state one at time t. We denote this probability $f_{12}(x)$.

This parameter needs some further explanation. In accordance with assumptions in Section 2.3 we do not know anything about the covariance distribution $C(x)$. To calculate $\hat{f}_{12}$ we need to assume that the covariate vector is constant at x_t in the future. The parameter f_{12} is then more correctly interpreted as the probability of the chain entering state two before entering state zero when it start in state one at time t, and the world stays the same in every other respect for all time after t. The probability $f_{12}(x)$ is the long term probability of entering state two before zero at a specific covariate value x.

As constant covariate value x is assumed, we may use standard theory of *homogenous* Markov chains to calculate $f_{12}(x)$. The first state $j = 0$ is to be considered as an absorbing state. In accordance with Ross (2014) we define the probability matrix of only transient states $j = 1, 2$ as

$$P_T(x_{i,t}) = \begin{pmatrix} \pi_{11}(x_{i,t}) & \pi_{12}(x_{i,t}) \\ \pi_{21}(x_{i,t}) & \pi_{22}(x_{i,t}) \end{pmatrix}.$$

For $k, j = 1, 2$, let s_{kj} denote the expected number of time periods that the homogenous Markov chain is in state j, given that it starts in state k. We define the matrix of the expected number of time periods in each transient state

$$S(x_{i,t}) = \begin{pmatrix} s_{11}(x_{i,t}) & s_{12}(x_{i,t}) \\ s_{21}(x_{i,t}) & s_{22}(x_{i,t}) \end{pmatrix}$$

This matrix may be found by

$$S(x_{i,t}) = \Big(I - P_T(x_{i,t})\Big)^{-1}$$

The probability $f_{12}(x)$ of the chain entering state 2 before entering state 0 when starting in state 1 is then given by

$$f_{12}(x_{i,t}) = \frac{s_{12}(x_{i,t})}{s_{22}(x_{i,t})}.$$

In the simulation study, I first consider $f_{12}(x_{\text{high}})$ as focus parameter. The results of the FIC simulations for this focus parameter are given in Figure 5.10. The true root mean squared errors of $\hat{f}_{12}(x_{\text{high}})$ are 0.0493, 0.0493, 0.0592, 0.2183, 0.3501. The *wide* model M_0 and the first candidate model M_1 are equally good models. In fact, M_1 is a tiny bit better than M_0 in that its root mean squared error is 0.00005 lower. The worst model is again the multinomial model. The average $\sqrt{\text{fic}}$-values of the hundred simulation rounds are 0.0439, 0.0439, 0.0565, 0.2154, 0.3490, which reflect nicely the true values. The model chosen the most often is however the model M_2! This model is chosen in 43% of the runs, due to lower bias and a moderately low squared bias. In the runs where it is considered unbiased, it will be the winner. The *wide* model M_0 and M_1 are chosen in 19% and 38% of the runs respectively. As these two models are practically identical for f_{12}, the FIC chooses one of the winner models in 57% of the runs so we may conclude that also in this simulation setup, the FIC

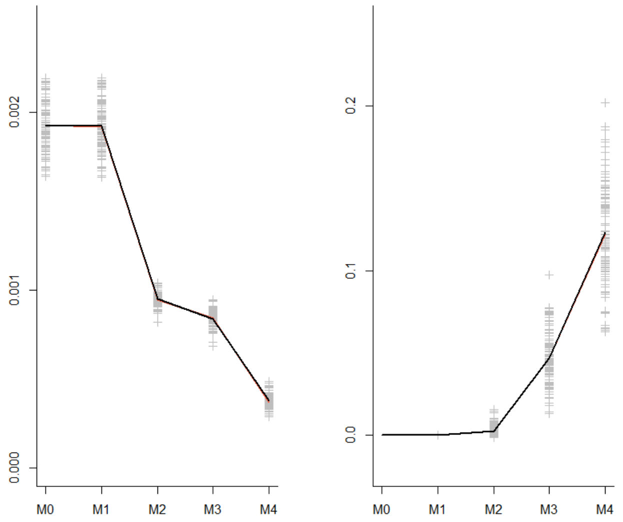

Figure 5.10: *Simulation results for $\mu = f_{12}(x_{high})$. The variance part of the mse is given in the left panel. The squared bias part is given in the right panel. The red lines are the true values, the grey crosses are the variance and squared bias part in each of the 100 simulation rounds. The black lines are the average scores of the variance part and the squared bias part.*

tends to select the correct model. We remark that although M_1 is just a tiny amount better than the M_0, the FIC still selects this model twice as often.

I then consider $f_{12}(x_{\text{low}})$ as the focus parameter. The results of the FIC simulations for this focus parameter are given in Figure 5.11. The true root mean squared errors of $\hat{f}_{12}(x_{\text{low}})$ are 0.0244, 0.0246, 0.0196, 0.0837, 0.0373. The second candidate model M_2 is now the best model. The average values of the simulation runs are $0.0243, 0.0266, 0.0211, 0.0819, 0.0365$. The best model M_2 is selected in 62% of the runs. The multinomial is selected in 30% of the runs. Again, this is due to the low variance of the multinomial model, and that it in some runs will have an estimated bias of zero.

5.5 Third Focus Parameter: Model Parameters

Finally, I choose as focus parameter the effect of covariate $x_{i,t,1}$ when past level $y_{i,t-1}$ is 1. Strictly speaking, this will be an example of the AFIC procedure. As the effect of $x_{i,t,1}$ at past level $k = 1$ consists of β_{M101} and β_{M111} we have a set of focus parameter for which we need to calculate the aggregated estimated

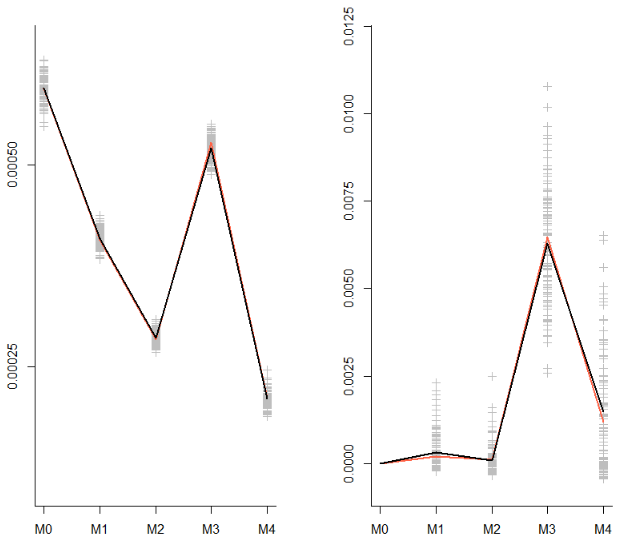

Figure 5.11: *Simulation results for $\mu = f_{12}(x_{low})$. The variance part of the mse is given in the left panel. The squared bias part is given in the right panel. The red lines are the true values, the grey crosses are the variance and squared bias part in each of the 100 simulation rounds. The black lines are the average scores of the variance part and the squared bias part.*

mean squared error as described in section Section 4.3.

The AFIC results for this composite focus parameter are given in Figure 5.12. The true aggregated root mean squared errors are $0.458, 0.469, 0.523, 0.304, 0.358$. Surprisingly, the model M_3, which deemed amongt the worst models in all the above settings, is now considered to be the best model. The multinomial model is considered the second best, only somewhat behind. The average values of the simulation runs are $0.463, 0.500, 0.551, 0.399, 0.420$. The correct model M_3 is chosen in 62% of the runs.

This last simulation round illustrates nicely the point that a model may be give precise estimates of a focus parameter at the same as it is not very precise in the estimates of a different focus parameter. The FIC is an excellent method for detecting such varying performance of a model. The simulations studies in this chapter illustrate that the FIC for dynamic multinomial logit models tends to select the model which best estimates the focus parameter.

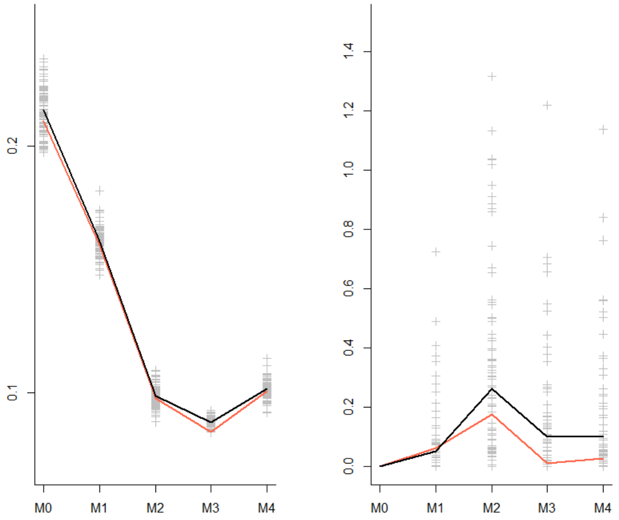

Figure 5.12: *Simulation results for focus parameter set* $\{\beta_{M101}, \beta_{M111}\}$*. The variance part of the mse is given in the left panel. The squared bias part is given in the right panel. The red lines are the true values, the grey crosses are the variance and squared bias part in each of the 100 simulation rounds. The black lines are the average scores of the variance part and the squared bias part.*

CHAPTER 6

Applications to conflict data

As an application of the developed methodology, I analyze the *Militarized Interstate Dispute* (MID) data set from the Correlates of War project (Maoz et al., 2018). This data set contains information on every militarized conflict between states in the period from 1816 to 2010. The interstate conflicts in MID are categorized according to the level of severity, thus the data set constitutes dyadic time series of conflict in concordance with the setup in Chapter 2. I restrict the analysis in this section to the years 1950 to 2010.

The *focus* of the analysis will be the effect of democracy on conflict escalation. As mentioned in the introduction, it is a firmly established result in the study of international relations that democracies rarely fight each other. Among the researchers in the field, there is considerable agreement that the 'absence of war between democratic states comes as close as anything we have to an empirical law in international relations' (Levy, 1989). But are democracies also less prone to let conflicts escalate? It may be that democracies rarely enter minor conflicts in the first place, but when states already are in a minor conflict does the level of democracy have any lowering effect on the probability of escalation into full-scale war?

I construct ten multinomial Markov models on the form of (2.3) to model the conflict dynamics between states. With the purpose of finding the model that best estimates the effect of democracy on escalation, I do model selection with the *Focused Information Criterion*.

I use two *focus parameters*. The first focus parameter is the probability $\pi_{12}(x)$, which in this setting means the probability of escalation from minor conflict in year t to war in year $t+1$. The second focus parameter is the probability $f_{12}(x)$ described in Chapter 5, which in this setting means the probability of a minor conflict developing into war, before a state of peace is entered. We learn from this FIC analysis, that democracy does in fact seem to have a lowering effect on escalation probability. However, the results are not significant, as would be expected.

In section 6.1 I describe the MID data set more in detail. We describe here also covariates to be used in the analysis, as well as the data sources of these. In section 6.2 I argue that the model setup of chapter 2 is appropriate to analyze the data. We describe the *wide* model together with nine candidate models. In section 6.3 I do 'traditional' model selection with the AIC. In section 6.4 I take on the question of conflict escalation specifically and we do *focused* model selection for $\pi_{12}(x)$. In section 6.5 I do focused model selection for the more involved parameter $f_{12}(x)$. A summary of the analysis is given in section 6.6.

6.1 The Data

The MID data set contains information on all instances when one state threatened, displayed, or used force against another state in the period from 1816 to 2010. The conflict instances in the data set are categorized according to hostility levels on a scale from one to five. On this scale level 1 represents *no militarized conflict*, level 2 represents a *threat*, the levels 3, 4 and 5 represent a *display of force*, *use of force* and *interstate war*, respectively.

The MID data set contains only instances of militarized conflict *between* states. Civil wars, however violent, are not included. Neither are state-led operations on foreign territory, if the operations are not directed against the foreign state itself. It may be for example, that a country use force in foreign countries by invitation, as France does in Mali since 2013. Another example is the American-led invasion of Afghanistan in 2010. This invasion was an interstate war with a registry of level 5 in 2001 in the MID data set. The Taliban state was quickly conquered and it was replaced by a US friendly regime that led the country in the following years. The militarized incidents in Afghanistan after 2001 were therefore not between states and are not included in the MID data set.

We will restrict the analysis to the years 1950 to 2010. In this period there is $m = 1089$ politically relevant pairs of countries, or dyads. Lemke and Reed (2001) define a politically relevant dyad to be a dyad consisting of two neighboring countries or at least one major power. We define neighboring countries according to the COW data set of contiguous countries (Stinnett, Tir, Schafer, Diehl, & Gochman, 2002). Thus to take Switzerland an example, this country constitutes politically relevant dyads with each of its neighboring countries France, Italy, Austria and Germany. The major powers in the period are considerd to be the USA, England, France, Russia/USSR, and China. (Correlates of War Project, 2017) As major powers, they have the capability to engage in military operations all over the globe and will constitute politically relevant dyads with every country in the world.

Countries too small, like Monaco, St.Kitts or Solomon Islands are omitted from the analysis. Countries with missing covariate values, like Surinam, Fiji or South Vietnam are also omitted from the analysis. Generally, these omitted countries are countries whose status typically is unclear. However, we have omitted West Germany and East Germany due to missing GDP. values in the period.

For each politically relevant pair of countries in the period, we construct a time series of conflict $\{y_{i,t}\}$ from the MID data set. Each politically relevant

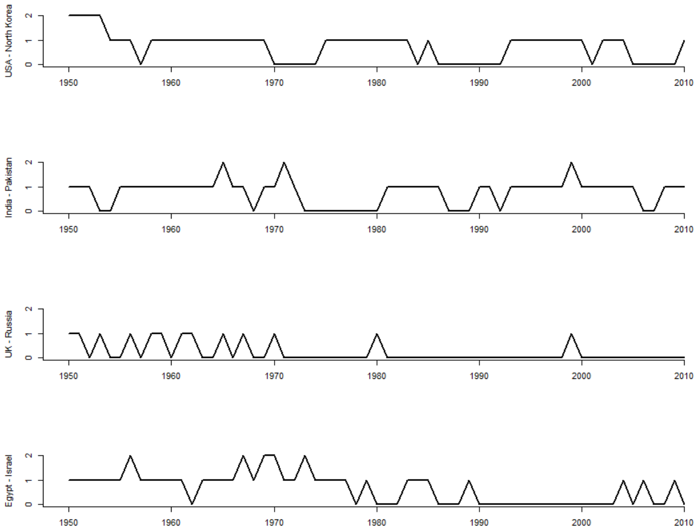

Figure 6.1: *Conflict levels in four dyads. From above are shown dyads USA - North Korea, India - Pakistan, United Kingdom - Russia/USSR, Egypt - Israel.*

pair of countries are indexed with $i = 1, \dots m$. The variable t is the number of years after 1950, such that $t = 0$ in 1950 and $t = 60$ in 2010. The length of the conflict series is $n_i = 61$ if the two countries in the dyad existed over the whole period. If a country came into existence after 1950, the conflict series with the country will have its first observation in the year the newest country in the dyad was founded.

As in Chapter 2, we define three levels of conflict $j = 0, 1, 2$ for the response variable $y_{i,t}$. These three conflict levels are

- 0 - *No Conflict*: no observation in the MID data set between the two countries in the dyad in the t'th year. Observations at level 1 in the MID data set are also included here.[1]
- 1 - *Minor Conflict*: An observation at level 2-4 in the MID between the two countries in the t'th year.
- 2 - *Major Conflict*: An observation at level 5 (War) in the MID between the two countries in the t'th year.

With this categorization, the data consists of 50465 observations at the level zero, 1923 observations at the level one as well as 147 observations at the

[1] There are no observations at level 1 in the MID data set.

level two. Not surprisingly the huge majority of observations are at level zero. Countries are normally at peace with each other. We see that major conflicts are very rare events. In Section 6.1 we have plotted the conflict series in four rather conflict-riven dyads.

When it comes to the data for covariate vectors, these are taken from a variety of sources. The following variables are considered:

- *Democracy.* As a measure of democracy level in a given country in a given year, we use the polity score from the Polity IV data set. (Marshall & Jaggers, 2003) The polity score is a number on the scale from -10 to 10, where 10 means that that the country is fully democratized and -10 means that the country is not democratic at all. We standardize this variable by adding ten to each polity score and then dividing by twenty. As each dyad consists of two countries, the covariates used in the analysis are $\texttt{dem}_{\text{high}}$ and $\texttt{dem}_{\text{low}}$ which report highest and lowest democracy scores in the dyad.
- *Gross Domestic Product.* Economic development is a factor which influences a state's capacity to project power. To measure the difference in economic development in a dyad, we use data from the 2018 version of the Maddison project database. (Bolt, Inklaar, de Jong, & van Zanden, 2018) This data set consists of data on GDP pr. capita in each country. We define covariate `GDP.ratio` as the logarithm of the ratio between the highest and the lowest gdp pr. capita in the dyad at each year. We normalize the variable such that it is on the scale (0,1).
- *Military Capacity.* The difference in military capacity may affect the conflict level in a dyad. We use the National Material Capabilities,version 5.0, of the Correlates of War project (J. D. Singer, Bremer, & Stuckey, 1972; J. Singer, 1987) as a measure of the (potential) military strength of a country. The variable `cinc.ratio` is the logarithm of the ratio between the highest and the lowest cinc value in the dyad at each year.
- *Major Power.* The variable `major` is an indicator of whether China, France, Great Britain, Russia/USSR or the USA is a country in the dyad (Correlates of War Project, 2017). The variable has the value of 1 if this is the case, 0 otherwise.
- *Alliances.* The variable `alliance` is an indicator of whether the countries in the dyad are allied. The variable has the value of 1 if the countries have a defense pact, neutrality pact or entente according to the Correlates of War Alliances dataset. Data on alliances taken from version 4.1 of the Formal Alliances data set of the COW project. (Gibler, 2009)
- *Contiguity.* The variable `contiguity` is an indicator of whether the countries in the dyad are neighboring countries, according to the Correlates of War Direct Contiguity data set (Stinnett et al., 2002). Two countries

are neighboring countries if they have a shared land border, or if a straight line of maximum 400 miles can be drawn across open water between points on the border of the two states in the dyad.

6.2 The wide model and candidate models

We will use the *multinomial* Markov chain models with a *logit* link to analyze the conflict time series $\{y_{i,t}\}$ as a response to the time series of covariate vectors $\{x_{i,t}\}$.

A central assumption is thus that the dyadic conflict chains $\{y_i\}$ are Markov chains of order one. This is a reasonable assumption. The willingness of war in a dyad may evidently be different if the dyad saw conflict the previous year (Beck & Katz, 1998). Given the persistence of war in some dyads however, one could argue that the Markov dependency should extend further back than just to past conflict level at $t-1$. Some dyadic relationships resemble feuds, whose origins go even further back than 1950. In figure Section 6.1 we see that the relationship between India and Pakistan has been tense the whole period. Such an extension would have complicated the model setup considerably however, increasing the number of parameters greatly. We do not, therefore, consider this option.

Another central assumption is that the chains $\{y_{i,t}\}$ are independent. In the setting of the MID data, this means that the conflicts in every dyad of countries are independent. This is of course a simplification. In the age of globalization, violence in one corner of the world has an effect on the probability of violence in another corner. It should be reasonable, however, to assume that the conflict process between countries A and B has no influence on the conflict process between countries C and D. More problematic is the assumption that the conflict process between countries A and B is independent of the conflict process between countries A and C. The Markov chain model defined in Chapter 2 does not take interdependence between dyads into consideration. We will discuss this simplification further in Chapter 7.

To do model selection with the FIC, we define a *wide* model that is to play the role as the *true* data-generating mechanism in the analysis. As mentioned in Chapter 3, this model should include all covariates that *a priori* are thought to have some effects on the response variable. Consulting the international relations literature we find that the covariates defined in Section 6.1 are considered to be central explanatory variables to dyadic conflict levels. As mentioned above, it is widely agreed that democracy has an effect on the probability of conflict. Gartzke (2007) argues that also economic development has an effect on the probability of war, more advanced economies being, on the one hand, less dependent on conquerable resources, whereas, on the other hand, highly developed economies have a higher military capacity and therefore a greater potential of power projection. Russet and Oneal (2001) include a measure for alliances in their analysis, and Gowa (1999) claims that much of the ´long peace' after the second world war is explained by the alliance system of the cold war.

In accordance with the theoretical framework in this thesis, we let the *wide* model have only Markov-dependent effects and also Markov-dependent intercepts. In the notation of chapter two then, we write the *wide* model as

$$\begin{aligned} M_0: \qquad & u_{i,t} : \emptyset \\ & z_{i,t} : 1, \texttt{dem}_{\text{high}}, \texttt{dem}_{\text{low}}, \texttt{gdp.ratio}, \texttt{cinc.ratio}, \\ & \qquad\qquad \texttt{major}, \texttt{alliance}, \texttt{contiguity}. \end{aligned}$$

As there are three conflict levels and two parameters for each covariate at each level, the *wide* model has a total of $6 \cdot 8 = 48$ parameters.

The candidate models to be considered are the following:

- M_1: A model with only Markov-dependent effects, but no indicators.

$$\begin{aligned} & u_{i,t} : \emptyset \\ & z_{i,t} : 1, \texttt{dem}_{\text{high}}, \texttt{dem}_{\text{low}}, \texttt{gdp.ratio}, \texttt{cinc.ratio}, \end{aligned}$$

- M_2 : A model with only democracy variables considered. Markov-dependent effects and intercept.

$$\begin{aligned} & u_{i,t} : \emptyset \\ & z_{i,t} : 1, \texttt{dem}_{\text{high}}, \texttt{dem}_{\text{low}} \end{aligned}$$

- M_3: A model with only democracy variables considered. Markov independent effects, with Markov dependent intercept.

$$\begin{aligned} & u_{i,t} : \texttt{dem}_{\text{high}}, \texttt{dem}_{\text{low}} \\ & z_{i,t} : 1 \end{aligned}$$

- M_4: Multinomial model.

$$\begin{aligned} & u_{i,t} : 1, \texttt{dem}_{\text{high}}, \texttt{dem}_{\text{low}}, \texttt{gdp.ratio}, \texttt{cinc.ratio}, \\ & \qquad\qquad \texttt{major}, \texttt{alliance}, \texttt{contiguity} \\ & z_{i,t} : \emptyset \end{aligned}$$

- M_5: A model with all variables considered, but only Markov indpendent effects. Markov dependent intercept.

$$\begin{aligned} u_{i,t} &: \texttt{dem}_{\text{high}}, \texttt{dem}_{\text{low}}, \texttt{gdp.ratio}, \texttt{cinc.ratio}, \\ &\qquad \texttt{major}, \texttt{alliance}, \texttt{contiguity} \\ z_{i,t} &: 1 \end{aligned}$$

- M_6: A model with all variables considered. Indicators Markov-independent.

$$\begin{aligned} u_{i,t} &: \texttt{major}, \texttt{alliance}, \texttt{contiguity} \\ z_{i,t} &: 1, \texttt{dem}_{\text{high}}, \texttt{dem}_{\text{low}}, \texttt{gdp.ratio}, \texttt{cinc.ratio}. \end{aligned}$$

- M_7: A model with all covariates, except `gdp.ratio`. Markov dependent effects and intercept.

$$\begin{aligned} u_{i,t} &: \\ z_{i,t} &: 1, \texttt{dem}_{\text{high}}, \texttt{dem}_{\text{low}}, \texttt{cinc.ratio}, \texttt{major}, \texttt{alliance}, \texttt{contiguity} \end{aligned}$$

- M_8: A model with no indicators. Intercept and `dem.high` Markov-dependent.

$$\begin{aligned} u_{i,t} &: \texttt{dem}_{\text{low}}, \texttt{gdp.ratio}, \texttt{cinc.ratio} \\ z_{i,t} &: 1, \texttt{dem}_{\text{high}} \end{aligned}$$

- M_9: A model with no democracy effects.

$$\begin{aligned} u_{i,t} &: \emptyset \\ z_{i,t} &: 1, \texttt{dem}_{\text{high}}, \texttt{dem}_{\text{low}}, \texttt{gdp.ratio}, \texttt{cinc.ratio}, \\ &\qquad \texttt{major}, \texttt{alliance}, \texttt{contiguity} \end{aligned}$$

6.3 Model Selection with the AIC

Before launching the FIC machinery, I do model selection with the traditional *Aikake Information Criterion* (AIC). I do this for the sake of comparison with the FIC. The AIC-score of a model with parameter β is given by the formula

$$\text{AIC} = 2\,\ell_{M,n}(\hat{\beta}_{M,n}) - 2\,p.$$

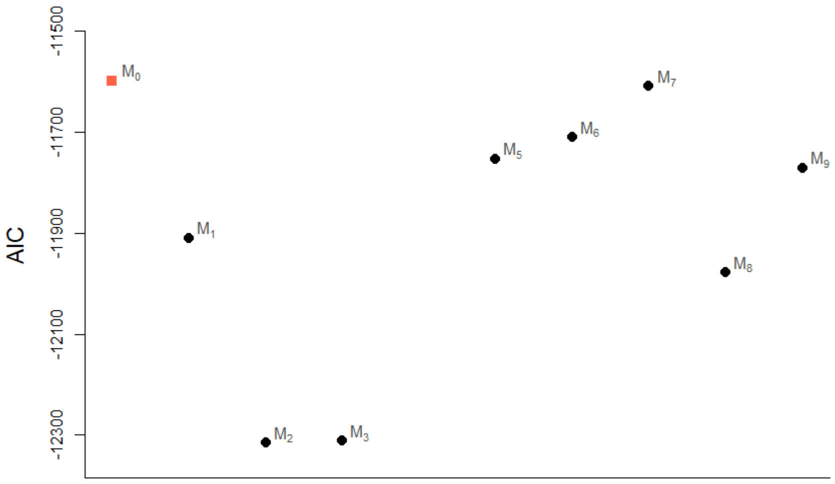

Figure 6.2: *AIC scores of wide model and candidate models. The y-axis shows the AIC-score. The x-axis has no values as it gives a listing of the models. The red square is the AIC score of the wide model. The black circles are the AIC scores of the candidate models. The multinomial model has an AIC score below the scale of the y-axis and is not in the plot.*

Models are ranged after the likelihood minus a penalty term for parameters. The model with the lowest AIC is deemed to be the model closest to the *true* data generating mechanism.

In figure Figure 6.2 I have plotted the AIC score for the *wide* model and the candidate models. The y-axis shows the AIC score. In the x-direction, the models are listed in order of the candidate number, so the x-axis has no values. We see from the plot that the *wide* model and M_7 are considered approximately equally good models. In fact, *the wide* model is considered to be closest to the *true* data-generating mechanism in that its AIC score is -11598, whereas the AIC score of M_7 is -11609. The model M_7 differs from the *wide* model only in that it has no effects of GDP per capita included.

The other candidate models are somewhat behind these two preferred models. Notice especially that the models M_2 and M_3 are judged to be rather far away from the true data-generating mechanism by the AIC. These two models have only effects of democracy variables included, in addition to an intercept. Worst ranked of all the candidate models is however the multinomial model. This model has an AIC score of -16168, way less than all the other models. As this is the only model that treats the data as independent, the exceedingly bad performance of the *multinomial* model is a sure sign that there is some Markov-dependency among the data.

6.4 Model selection with the AFIC, first focus parameter

We then select models with regard to the research question. First, we choose the transition probability $\pi_{12}(x)$ as the focus parameter. We use the developed FIC with this focus parameter to select the model deemed best at estimating $\pi_{12}(x)$.

Strictly speaking, the selection mechanism is now the Average-FIC, the AFIC. We are not interested in finding the best estimate of $\pi_{12}(x)$ for a specific value of x, but in finding the best estimate of $\pi_{12}(x)$ over the whole space Γ of covariates. That is, the aggregated focus parameter is actually

$$\pi_{12} = \int \pi_{12}(x)\mathrm{d}C(x)$$

As the cumulative weight function $W(t)$ discussed in Section 4.3 we use the empirical distribution of covariates.

Methodologically there is of course nothing wrong in selecting models with regards to a single covariate value. If we were mainly interested in giving an estimate of the probability of conflict for a specific dyad in a specific year, this would have been the right thing to do. If we were interested in giving the best estimate of $\pi_{12}(x)$ only for a subset of countries, for example only for democratic countries, we could have used the AFIC with respect to only the subset of democratic countries. As the research question is universal in its scope, we are asking about the effect of democracy on any interstate conflict at any time. We are thus interested in selecting the model that best estimates $\pi_{12}(x)$ over the whole range of covariates. The π_{12} parameter is, therefore, the correct choice of *focus parameter*.

This being said, we need to choose an individual value of x for the purpose of visualization. We choose the covariate values in the dyad USA- North Korea in 2010 as our reference. We denote these particular covariate values x^*. The reason for choosing these values as the reference for visualization is that in 2010 the USA and North Korea were at the opposite extremes of democracy values. Being fully democratic, USA had a Polity score of 10. Being fully autocratic, North Korea, had a Polity score of -10. Thus $\texttt{dem}_{\mathrm{high}} = 1$ and $\texttt{dem}_{\mathrm{low}} = 0$ in x^*. When reading the plots one needs in keep in mind that $\pi_{12}(x*)$ is only the reference value for plotting, *not* the only focus parameter of the AFIC procedure.

Now to the model selection. The FIC chooses the model which strikes the best balance between estimated variance and bias of the focus parameter. A model with many relevant parameters will typically have a small bias.On the other hand, the more parameters a model has, the more variability does it have in its estimates. We saw above that the vast majority of observations in the

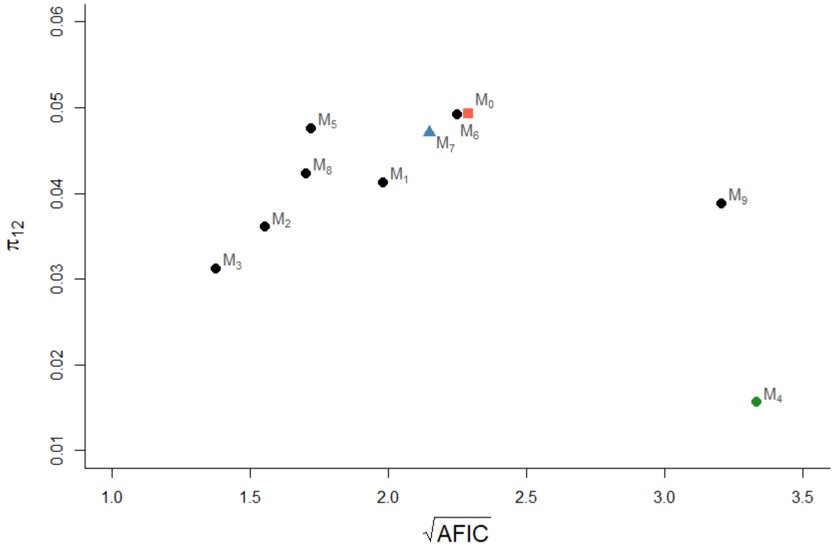

Figure 6.3: *Root-AFIC score and $\pi_{12}(x^*)$ estimates of wide model and candidate models. The red square is the wide model. The blue triangle is the model M_7. The green circle is the multinomial model M_4. The black circles are the remaining candidate models.*

dyadic time series are at the peace level. There were some observations at the minor conflict level, and only a few observations at the major conflict level. As the observations of conflict are so few it would therefore be reasonable to expect that a model with many parameters include too much variance to give precise estimates of $\pi_{12}(x)$. Although the *wide* model is selected by the AIC to be the model closest to the true data-generating mechanism, we would not be surprised if this model, with 48 parameters, was deemed by the FIC to be too wide for estimation of π_{12}.

This is indeed the case. In Section 6.4 is an AFIC-plot of the *wide* model and candidate models. The x-axis in the plot is the $\sqrt{AFIC}$ value. The y-axis is the value of the reference parameter $\pi_{12}(x^*)$ of the dyad USA-North Korea in 2010. The red square and the blue triangle are the *wide* model and the model M_7 which was deemed almost as good by the AIC. We see from the plot that neither of these are considered by the AFIC to give very good estimates of the focus parameter. Both are unbiased, but they have huge aggregated variances, 423.9 for the *wide* model and 409.4 for the model M_7. The model with the smallest variance is the standard multinomial model M_4. This model is the green circle in the AFIC plot. The aggregated variance of this model is only 47.1. But its aggregated squared bias is much too high: 11.0. The multinomial

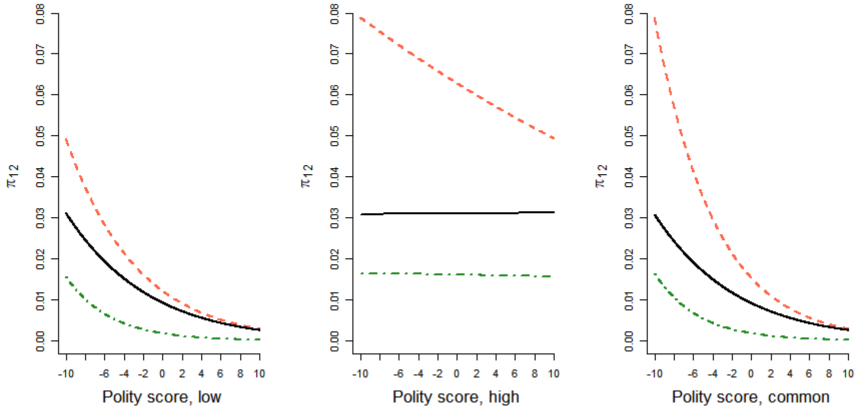

Figure 6.4: *Probability curves of $\hat{\pi}_{12}(x^*)$ with varying democracy levels. The black whole line is the estimate of FIC winner M_3. The red dotted line is the estimate of the wide model. The green dotted line is the estimate of the multinomial model. In the left panel, covariate values are held at x^* but with $\texttt{dem}_{low}$ varying. The middle panel has the same setup, but with $\texttt{dem}_{high}$ varying. In the right panel covariate values are held at x^* but with both $\texttt{dem}_{low} = \texttt{dem}_{high}$ varying.*

model is therefore considered also by the FIC to be the worst model.

We see from the plot that the model chosen by the FIC is actually the very simple model M_3. This model takes only democracy effects $\texttt{dem}_{\text{low}}$ and $\texttt{dem}_{\text{high}}$ into consideration, and it considers these effects to be Markov independent. In this model, the Markov-dependency is taken care of by the intercept. This model is considered to have a very low aggregated bias (0.09) for π_{12}, but due to its simpleness, it has a much lower aggregated variance (202.2) than the *wide* model. It is therefore considered to be much more precise for the estimation of π_{12}. The model M_2 is also deemed good. Interestingly, both M_2 and M_3 were considered by the AIC to be models at some distance from the true data-generating mechanism.

In Figure 6.4 we have plotted the transition probability of $\pi_{12}(x)$ for the value of x^*, but where the democracy-variables are varying. In the left panel, the curve of $\pi_{12}(x^*)$ is plotted when the value of $\texttt{dem}_{\text{low}}$ is varying. This corresponds to the counter-factual situation where North Korea had a higher democracy level in 2010, but all other covariate values in x^* remained unchanged. In the middle panel the curve of $\pi_{12}(x^*)$ is plotted for changing values of $\texttt{dem}_{\text{high}}$. This correspond to the counter-factual situation where the USA had a lower democracy level in 2010, but all other covariate values in x^* remained unchanged. In the right panel the probability curve of $\pi_{12}(x^*)$ is plotted when $\texttt{dem}_{\text{low}} = \texttt{dem}_{\text{high}}$. This corresponds to the counter-factual situation where

North Korea and the USA had the same democracy level in 2010, but all other covariate values in x^* remained unchanged. The black line is the estimates $\hat{\pi}_{12}(x^*)$ of the FIC-winner M_3, the red dotted line is the corresponding estimates of the *wide* model. The green dotted line is the estimates of the multinomial model.

In the left panel of Figure 6.4 , we see that all models estimate a drop in probability when the country with the lowest democracy score becomes more democratic. The drop is somewhat greater in the *wide* model than in the FIC winner. The two models agree on the estimates for high values of $\texttt{dem}_{\text{low}}$. For low values of $\texttt{dem}_{\text{low}}$, the *wide* model seems to overestimate the probability of escalation π_{12}. The multinomial model underestimates the probability of escalation for all values, which is not surprising as it does distinguish between escalation to war and escalation to *minor conflict*.

The middle panel of Figure 6.4 is interesting. The *wide* model estimates a considerable increase in escalation probability π_{12} when the democracy level of the most democratic country in the dyad drops. In other words, according to the *wide* model, probabilities of escalation rise also if the most democratic country becomes less democratic. The FIC-winner gives another conclusion, however. According to this model, the democracy level $\texttt{dem}_{\text{high}}$ has no considerable effect on escalation probability π_{12}.

When it comes to the right panel of Figure 6.4, both the *wide* model and the FIC winner M_3 predicts a drop in escalation probability π_{12} when $\texttt{dem}_{\text{low}} = \texttt{dem}_{\text{high}}$ increases. This drop is less marked in the M_3 model since $\texttt{dem}_{\text{low}}$ is the only value with a considerable effect on π_{12}.

In figure Figure 6.5 we have plotted the corresponding curves of $\hat{\pi}_{12}(x^*)$ of M_3 with 95% pointwise probability bands. See Section 3.6. In the left plot, we see that there seems to be a significant effect of $\texttt{dem}_{\text{low}}$ on $\pi_{12}(x)$. However, we can draw no proper conclusion, as the confidence bands are not *simultaneous*. On the other hand, it is perfectly warranted to conclude that $\texttt{dem}_{\text{high}}$ has no effect on $\pi_{12}(x)$ according to the model M_3.

6.5 Model selection with the AFIC, second focus parameter

Let us now do model selection with the more involved focus parameter $f_{12}(x)$. Recall from Section 5.4 that this parameter is the probability that the Markov chain will enter state 2 before it enters state 0, when the chain starts in state 1. In the context of conflict modeling, this translates into being the probability that a minor conflict between two countries will develop into a war before peace is obtained. In many ways, this focus parameter f_{12} reflects better the focus question. When we ask about the probability of escalation, we are not merely interested in the probability of a minor conflict developing into war next year, but the probability of the minor conflict being the beginning of a chain of conflicts that in the end will lead to war.

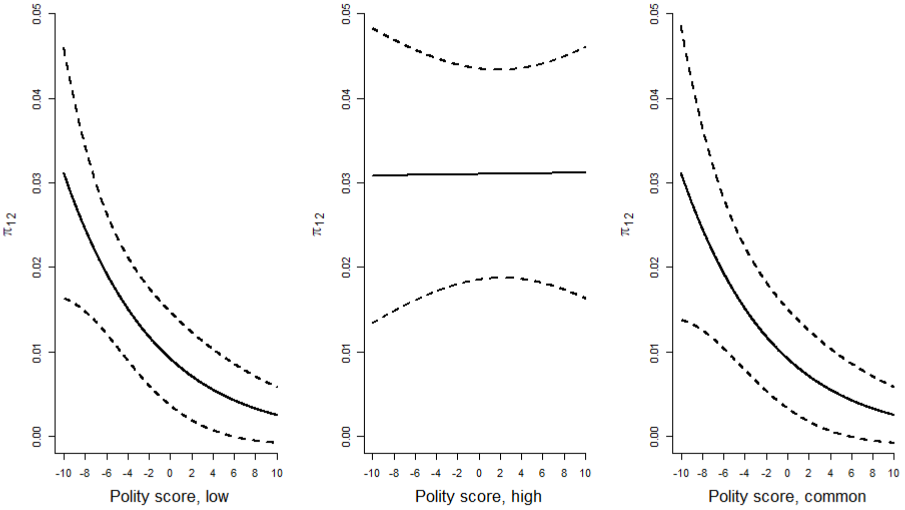

Figure 6.5: *Probability curves of $\hat{\pi}_{12}(x^*)$ with 95% pointwise confidence bands for the FIC-winner M_3. In the left panel, covariate values are held at x^* but with* `dem`$_{low}$ *varying. The middle panel has the same setup, but with* `dem`$_{high}$ *varying. In the right panel covariate values are held at x^* but with both* `dem`$_{low}$ $=$ `dem`$_{high}$ *varying*

In the same manner as in Section 6.4, we use the AFIC as we are not merely interested in f_{12} for a specific covariate value, but for the whole covariate space Γ. We use the empirical distribution of covariates as the cumulative weight function. Also for f_{12} we use the covariate value x^* of USA-North Korea in 2010 for visualization.

The AFIC plot of the selection process is given in Figure 6.6. In the plot, we see that neither the *wide* model nor the model M_7 are considered by the AFIC to give very precise estimates of f_{12} . As was the case for π_{12} these models are unbiased, but they are estimated to have much too high aggregated variance to be the preferred models. The model which is the preferred model is model M_5. This model has the same structure as the AFIC-winner M_3 for π_{12}, in that it has only Markov independent effects and Markov dependent intercepts. It is a more complex model however, as it takes all covariates into consideration. The very simple model M_3 is however considered to give only somewhat less precise estimates of f_{12}. As in the case above, the multinomial model is considered to have a huge bias, thus being a very bad model, also for the focus parameter f_{12}.

In Figure 6.7 we have plotted curves for the value $f_{12}(x^*)$ with democracy values changing in the same manner as in Figure 6.4. We see from the left panel where we let `dem`$_{\text{low}}$ change that both the FIC-winner M_5 and the *wide* model estimate a considerable drop in f_{12} when the least democratic state in

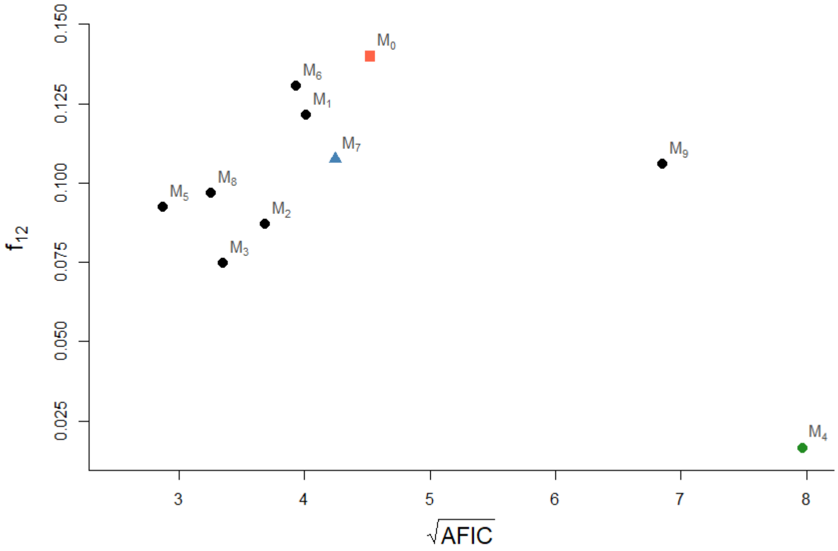

Figure 6.6: Root-AFIC score and $f_{12}(x^*)$ estimates of wide model and candidate models. The red square is the wide model. The blue triangle is the model M_7. The green circle is the multinomial model M_4. The black circles are the remaining candidate models.

the dyad becomes more democratic. In the range of high values of $\texttt{dem}_{\text{low}}$ they agree in their estimates. In the range of low values f $\texttt{dem}_{\text{low}}$ the *wide* model seems to be overestimating the probability of escalation. In the middle panel, we see that the probability curves are almost parallel for the *wide* model and the FIC-winner M_5, meaning that they agree in their estimates on the effect of $\texttt{dem}_{\text{high}}$. When the most democratic state gets less democratic, the probability f_{12} increases. The right panel shows curves for $\texttt{dem}_{\text{low}} = \texttt{dem}_{\text{high}}$, showing the same pattern of a clear lowering of escalation probabilities when democracy scores increase.

In figure Figure 6.8 the curves of f_{12} are plotted with 90% pointwise confidence bands for the AFIC winner M_5. A more conservative story is told in these plots. Although there is a clear tendency of lowering escalation probabilities f_{12} with increasing democracy scores, we see from these plots that the effect is not significant. Concentrating on the rightmost plot, we see that there is a small set of values (0.42,0.45) that are inside the confidence bands for all democracy values. The bands are pointwise confidence bands, but as the simultaneous confidence will be wider than the pointwise bands, we may draw the conclusion that although there is a clear tendency to lowering probabilities with increasing democracy scores, the effect of democracy on f_{12} is *not* significant, not even at

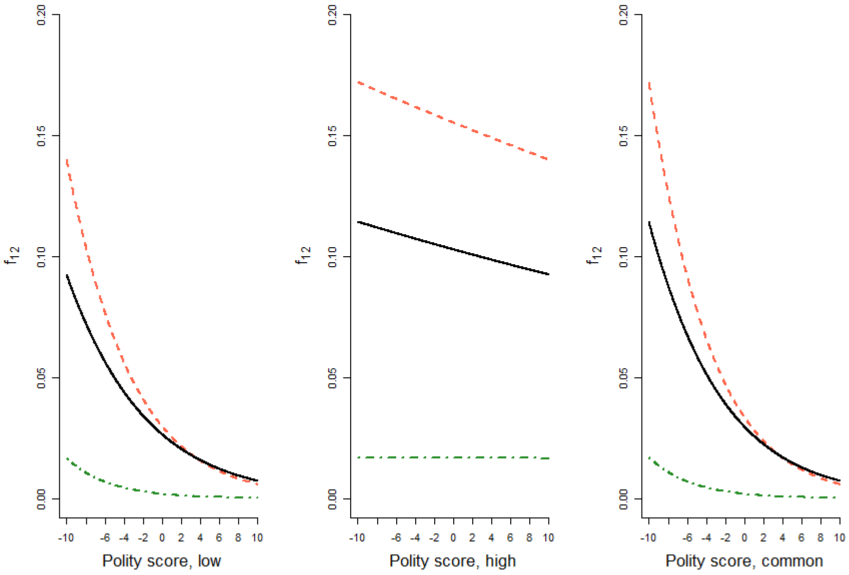

Figure 6.7: *Probability curves of $f_{12}(x^*)$ with varying democracy democracy levels. The black whole line is the estimate of FIC winner M_5. The red dotted line is the estimate of the wide model. The green dotted line is the estimate of the multinomial model. In the left panel, covariate values are held at x^* but with $\texttt{dem}_{low}$ varying. The middle panel has the same setup, but with $\texttt{dem}_{high}$ varying. In the right panel covariate values are held at x^* but with both $\texttt{dem}_{low} = \texttt{dem}_{high}$ varying.*

the $\alpha = 0.1$ level.

6.6 Summary of the MID Analysis

To summarize what we have learned from the analysis of the MID data set.

First, we have seen that the models deemed by the FIC to give the best estimates of escalation probabilities π_{12} and f_{12} are much simpler models than the *wide* model prefered by the AIC. Whereas the *wide* model has Markov-dependency in all effects, the models M_3 and M_5 chosen by the FIC have Markov dependency only in the intercepts, the other effects are Markov independent. This is not least due to the scarceness of conflict observations compared to the abundance of peace observations. The *wide* model is good at estimating transitions at the 0-1 level, it is too *wide* to analyze transitions at the 1-2 level.

When it comes to the results of the analysis, there seems to be a clear tendency of lowering probabilities with increasing democracy scores. This is

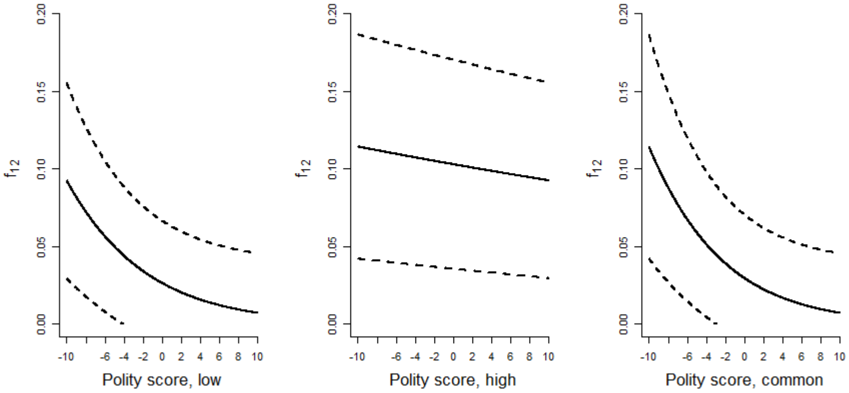

Figure 6.8: *Probability curves of $\hat{f}_{12}(x^*)$ with 90% pointwise confidence bands for the FIC-winner M_5. In the left panel, covariate values are held at x^* but with $\texttt{dem}_{low}$ varying. The middle panel has the same setup, but with $\texttt{dem}_{high}$ varying. In the right panel, covariate values are held at x^* but with both $\texttt{dem}_{low} = \texttt{dem}_{high}$ varying.*

the case both for the probability π_{12} of direct escalation and the more involved but also more interesting probability f_{12} of long term escalation. There may be sufficient evidence that there is a significant effect of $\texttt{dem}_{\text{low}}$ on the probability of war next year, but we need simulataneous confidence bands to conclude with confidence. When it comes to the more interesting question of f_{12}, the lowering effect of democracy scores is not significant.

CHAPTER 7

Concluding Remarks

The achievements of this thesis are fourfold. First, I have defined a dynamic multinomial logit model for *inhomogenous* Markov chains. This model is closely related to the model developed by Kaufmann (1987), Fahrmeir and Kaufmann (1987), Fokianos and Kedem (1998, 2003) and Kedem and Fokianos (2002). It differs however in the way past values of the Markov chains are treated. An advantage of the dynamic multinomial model as it is defined in this thesis is that it allows for asymptotic theory of maximum likelihood estimators under misspecification.

The second achievement of the thesis has been the development of this asymptotic theory of maximum likelihood estimators under misspecification. I have shown that maximum likelihood estimates $\hat{\beta}$ and $\hat{\beta}_M$ of the *wide* model and the candidate model respectively, have approximate joint normal distribution

$$\begin{pmatrix} \sqrt{n}(\hat{\beta} - \beta_{\text{true}}) \\ \sqrt{n}(\hat{\beta}_M - \beta_{M,0,n}) \end{pmatrix} \approx_d \mathrm{N}\left(0, \begin{pmatrix} J_n^{-1} & J_n^{-1} C_{M,n} J_{M,n}^{-1} \\ J_{M,n}^{-1} C_{M,n} J_n^{-1} & J_{M,n}^{-1} K_{M,n} J_{M,n}^{-1} \end{pmatrix}\right)$$

for large samples. I have found expressions for the variance matrices J_n, $K_{M,n}$ and $C_{M,n}$ and I have proposed a strategy for estimation of these matrices.

The third achievement of this thesis has been the development of a *Focused Information Criterion* for the defined dynamic multinomial logit models. I have done this in the scheme of the FIC with a fixed *wide* model. On the basis of the joint distribution of maximum likelihood estimators $\hat{\beta}$ and $\hat{\beta}_M$ I have demonstrated that also maximum likelihood estimates $\hat{\mu}$ and $\hat{\mu}_M$ of focus parameters in the *wide* model and the candidate model have approximate joint normal distribution

$$\begin{pmatrix} \sqrt{n}(\hat{\mu} - \mu_{\text{true}}) \\ \sqrt{n}(\hat{\mu}_M - \mu_{M,0,n}) \end{pmatrix} \approx_d \mathrm{N}\left(0, \begin{pmatrix} \nu_{\text{wide}} & \nu_{M,c} \\ \nu_{\text{M,c}} & \nu_M \end{pmatrix}\right)$$

for large samples. This approximate normal distribution of focus parameter estimates makes it possible to estimate mean squared errors of $\hat{\mu}$ and $\hat{\mu}_M$. It thus allows ranking of the *wide* model and candidate models by their precision in estimators of the focus parameter μ, which is the essence of the FIC.

Finally, I have used the developed FIC for dynamic multinomial logit models to analyze the *Militarized Interstate Dispute* data set. The focus parameter of the analysis was the probability of conflict escalation. The analysis showed a clear tendency towards reduction of escalation probability when countries are more democratic. Results did not turn out to be significant, however.

Through simulations and the analysis of the MID data set we have learned that the FIC is indeed a fruitful selection strategy when it comes to dynamic modeling. We have seen that the FIC may select different models depending on which focus parameter we are interested in estimating. With the FIC we may discover modeling aspects of the dynamic process that are lost of sight when global criteria like the AIC is used. We have seen that surprisingly simple models are preferred by the FIC for estimation of conflict escalation although a much more complex model is selected by the AIC.

We have however also encountered considerable estimation challenges. An indisputable weak point of the developed FIC with a fixed *wide* model is the complexity of the covariance matrix $K_{M,n}$ of the random score vector under misspecification. This complexity makes the calculation of estimates very hard and time consuming. The $K_{M,n}$ consists of an astronomical number of terms and taking them all in account is computationally infeasible. We have argued that we may reduce the number of terms drastically under the mild assumption that the Markov chain moderately fast finds its equilibrium distribution over the whole covariate space Γ. The estimation of the matrix will still be a complicated affair however.

I finish this thesis by proposing some additional themes that deserve to be further explored.

7.1 Alternative Data

I have used the developed methodology to analyze armed conflict data from the MID data set. Conflict research is however only one of many fields where the developed methodology model may find its application. The defined multinomial logit model and its associated FIC may be used to analyze any kind of dynamic systems where the Markov assumption is reasonable. As long as data from different chains are assumed to be independent, the developed methodology is a promising alternative to traditional methods.

The examples of dynamic systems that fulfills such criteria are numerous. In the introduction we mentioned that Markov chain models are used in a healthy diversity of fields. Medicine, genetics, engineering, economics and meteorology were mentioned. Concrete examples of data sets are given in Kedem and Fokianos (2002). These authors use dynamic multinomial models to analyze DNA sequence data, to make soccer predictions, to analyse sleep phases of humans.

The developed FIC should definitely be used to analyze data also from these diverse fields.

7.2 Alternative Modeling

The methodology in this thesis may be extended.

First, I have assumed throughout the thesis that the number of categories of the Markov chains is $K = 3$. It should not be difficult to generalize the theory to other choices of K. The case of $K = 2$ would even simplify matters. The choice of two categories would reduce the multinomial distribution to a binomal distribution and the binomial distribution is even simpler mathematically. It should also be possible to increase the number K of categories. An increase of categories would however increase the number of parameters extensively and the dreadful $K_{M,n}$ matrix would pose even greater estimation challenges. Formally however, the theory would have remained the same.

Secondly, I have developed the dynamic multinomial model within the framework of generalized linear models. As link function I have used the *logit* function. This has ensured the crucial property of a *concave* likelihood function, which guarantees unique least false parameter values. Not least beacuse of this mathematicall finesse the logit is a popular choiche.

It is possible to use other link functions for the dynamic multinomial model. A reasonable choice could for example be to consider the *probit* link. This would have changed the concrete mathematical expressions of matrices, score functions, likelihood and the like, but the general framework of the development would have remained the same. With the same proceedings we could hopefully have developed large sample distributions of ML estimates under misspecification and a FIC for dynamic multinomial *probit* models too. Other link functions should also be tried.

7.3 FIC in a Local Misspecification Context

The *Focused Information Criterion* developed in thesis assumes a *fixed* true model. Originally however the FIC was developed by N. L. Hjort and Claeskens (2003) and Claeskens and Hjort (2008b) in a local misspecification context. In this original version of the FIC, the true model is considered to be changing with sample size. The true multinomial logit model would in this version have been on the form

$$f(y|x, \theta_0, \gamma_0 + \frac{\delta}{\sqrt{n}}).$$

This model would only be $O(n^{-\frac{1}{2}})$ away from a narrow model on the form

$$f(y|x, \theta_0, \gamma_0).$$

In the orignial FIC, candidate models are considered to lie between these two models, which means that they need to be submodels of the true *wide* model.

The potential advantage of this orignal FIC approach is that we may get rid of the dreadful $K_{M,n}$ matrix. The variance matrices of random score vectors should in this local misspecification setting be attainable simply through manipulations of the comparatively simple matrix J_{wide}, which in the original FIC denotes the information matrix of the *wide* model, evaluated at the narrow model.

The development of such an original FIC for the dynamic mulitnomial logit model should be tried. It would surely be no easy undertaking. Plausibly it would be just as elaborate as the present development. But the rewards could be considerable in that the resulting expressions in this original FIC scheme could be simpler.

On the other hand, the original FIC approach involves reduced flexibility. As candidate models need to be submodels of the *wide* model, we may loose many of the interesting candidate models. In the MID analysis, we saw for example that the FIC winners had no Markov-dependent effects. These models were not submodels of the *wide* model, which had only Markov dependent effects. The multinomial model would also be excluded in this scheme. Thus in the original FIC scheme we would have lost the attractive of the dynamic multinomial logit model defined in this thesis, which consits in the accomodation of models with different Markov structure in the effects. Only by working out this original FIC for dynamic multinomal logit models could we assess the advantages and drawbacks however.

7.4 Interaction between Chains

Another issue for discussion is potential interaction effects between Markov chains. A central assumption of the methodology developed in this thesis is the assumption that different Markov chains are independent. This is not always a reasonable assumption for real dynamic systems. Often there is a considerable interaction effect between chains. Could the developed methodology be extended to allow for such interaction effects between Markov chains?

In the context of conflict modelling this is a pressing matter. We discussed the problem briefly in Section 6.2. In the analysis of the MID data set we assumed that each chain of observations of conflict between two countries are independent. This implied that the conflict observations between countries A and B were treated as independent from conflict observations between countries A and C, clearly a crude assumption.

With this interaction challenge in mind Cranmer, Desarmais, and Menninga (2016) argue that conflict researchers should abandon dyadic designs completely. As a response to this Poast (2016) argues that dyadic designs are reasonable as long as relevant variables which account for interaction between dyads are included in the model setup. Such a relevant variable could for example be an indicator of whether there was conflict time $t-1$ in any of the chains in which country A is one of the parts.

Could such an inclusion of interaction effects be achieved in the framework of this thesis? Probably not. That is, under correct model specification interaction would plausibly pose no problem. We have seen in Section 3.2 that the matrix

J_n involves no terms including $t-1$, and this would probably not change if we included past interaction in the model.

Under model misspecification the matter would have been different. Including past interaction effects would here break all the gates of Mordor wide open. The resulting correlation structure would result in a candidate random score variance matrix that would make the dreadful $K_{M,n}$ matrix look like an annoying little *gamin* in comparison.

However, the problem of interaction effects could maybe be alleviated in the original FIC approach. As we in this FIC scheme only consider the information matrix of the *wide* model, it would plausibly be the case that past interaction effects would case no extra complexity. This constitutes definitely another argument for investigating the original FIC version with regard to the dynamic multinomial logit model defined in this thesis. The loss in flexibility in the choice of candidate models could maybe be outweighed by inclusion of models which take interaction into consideration. The answer to this question only future research can reveal.

Appendices

APPENDIX A

Appendix

In this appendix I give proofs and expressions that are too comprehensive to be included in the text. For readability, I write

$$\Phi_{kj}^{(s)}(t) = \left| P_{kj}^{(s)}(t) - P_{2j}^{(s)}(t) \right|.$$

I also write in this appendix

$$\Sigma_{rj|k}(t) = \text{Cov}_{\text{wide}} \left\{ y_{t,r}, y_{t,j} \middle| y_{t-1,k} = 1 \right\}$$

In Appendix A.1 I give a proof of Lemma 3.1.1 in Section 3.1. In Appendix A.2 I find an expression for the unconditional variance of y_t. In Appendix A.3 I prove that for functions on the general form ψ_{tkj} the sum of covariances is only $o_p(1)$ away from an N-dimensional function of covariate values. In Appendix A.4 I give expressions for matrices $V_{M,n}$, $Q_{M,n}$ and $W_{M,n}$ which together with $K_{M,n}$ constitute the complex covariate matrix $K_{M,n}$. In Appendix A.5 I give a short note on R scripts used in this thesis.

A.1 Proof of the Correlation Structure.

We first prove Lemma 3.1.1 given in Section 3.1

Lemma A.1.1. *Let covaritate vectors $x_0, x_1, \ldots, x_t$ be generated by some unknown covariate distribution C(x) in accordance with the assumptions in Section 2.3. Given these covariate values, let $\{y_{i,t}\}$ be a Markov chain generated by the wide model. Define functions $\psi_{t,k,j}$ and $\psi^*_{t,k,j}$ as defined in Equation* (3.1). *It is then the case for all $t > 0$ that*

$$\lim_{t\to\infty} \sum_{s=0}^{t-1} \text{Cov}_{wide} \left\{ \psi_{t,k,j}, \psi_{t-s,k',j'} \right\} < \infty.$$

Proof. Define first

$$\lambda_{kj}^{(t)} = \mathrm{E}_{\mathrm{wide}} \left\{ f_{kj}^{(t)} y_{tj} + g_{kj}^{(t)} \middle| y_{t-1,k} = 1 \right\} = f_{kj}^{(t)} \pi_{kj}(x_t) + g_{kj}^{(t)},$$

and

$$\lambda_{kj}^{*(t)} = \mathrm{E}_{\mathrm{wide}} \left\{ f_{kj}^{*(t)} y_{tj} + g_{kj}^{*(t)} \middle| y_{t-1,k} = 1 \right\} = f_{kj}^{*(t)} \pi_{kj}(x_t) + g_{kj}^{*(t)}.$$

Consider for $s = 0$

$$\begin{aligned} &\mathrm{Cov}_{\mathrm{wide}} \left\{ \psi_{t,k,j}, \psi_{t,k',j'}^{*} \right\} \\ &\qquad = \mathrm{Cov}_{\mathrm{wide}} \left\{ \left(f_{kj}^{(t)} y_{tj} + g_{kj}^{(t)} \right) y_{t-1,k}, \left(f_{k'j'}^{*(t)} y_{tj'} + g_{k'j'}^{*(t)} \right) y_{t-1,k'} \right\}. \end{aligned}$$

By the law of total covariance, we may write this as

$$\begin{aligned} &= \mathrm{E}_{\mathrm{wide}}\ \mathrm{Cov}_{\mathrm{wide}} \left\{ \left(f_{kj}^{(t)} y_{tj} + g_{kj}^{(t)} \right) y_{t-1,k}, \left(f_{k'j'}^{*(t)} y_{tj'} + g_{k'j'}^{*(t)} \right) y_{t-1,k'} \middle| y_{t-1} \right\} \\ &\qquad + \mathrm{Cov}_{\mathrm{wide}} \left\{ \mathrm{E}_{\mathrm{wide}} \left\{ \left(f_{kj}^{(t)} y_{tj} + g_{kj}^{(t)} \right) y_{t-1,k} \middle| y_{t-1} \right\}, \right. \\ &\qquad\qquad \left. \mathrm{E}_{\mathrm{wide}} \left\{ \left(f_{k'j'}^{*(t)} y_{tj'} + g_{k'j'}^{*(t)} \right) y_{t-1,k'} \middle| y_{t-1} \right\} \right\} \\ &= \mathrm{E}_{\mathrm{wide}}\ \mathrm{Cov}_{\mathrm{wide}} \left\{ f_{kj}^{(t)} y_{tj} y_{t-1,k}, f_{k'j'}^{*(t)} y_{tj'} y_{t-1,k'} \middle| y_{t-1} \right\} \\ &\qquad\qquad + \mathrm{Cov}_{\mathrm{wide}} \left\{ \lambda_{kj}^{(t)} y_{t-1,k}, \lambda_{k'j'}^{*(t)} y_{t-1,k'} \right\} \\ &= f_{kj}^{(t)} f_{k'j'}^{*(t)}\ \mathrm{Cov}_{\mathrm{wide}} \left\{ y_{t,j} y_{t,j'} \middle| y_{t-1,k} = 1 \right\} \delta_{kk'}\ \mathrm{E}_{\mathrm{wide}}\ y_{t-1,k'} \\ &\qquad\qquad + \lambda_{kj}^{(t)} \lambda_{k'j'}^{*(t)}\ \mathrm{Cov}_{\mathrm{wide}} \left\{ y_{t-1,k}, y_{t-1,k'} \right\}. \qquad \text{(A.1)} \end{aligned}$$

We know that

$$\left| \mathrm{Cov}_{\mathrm{wide}} \left\{ y_{t,k}, y_{t,k'} \right\} \right| = \left| \mathrm{E}_{\mathrm{wide}}\ y_{t,k} y_{t,k'} - \mathrm{E}_{\mathrm{wide}}\ y_{t,k}\, \mathrm{E}_{\mathrm{wide}}\ y_{t,k'} \right| < 1.$$

As $f_{kj}^{(t)}, f_{kj}^{*(t)}, \lambda_{kj}^{(t)}$ and $\lambda_{kj}^{*(t)}$ are uniformly bounded functions for all k, j, t, there exists an upper bound G, such that

$$\mathrm{Cov}_{\mathrm{wide}}\left\{\psi_{t,k,j},\psi^*_{t,k',j'}\right\} < \left| f^{(t)}_{kj} f^{*(t)}_{k'j'} + \lambda^{(t)}_{kj}\lambda^{*(t)}_{k'j'} \right| < G.$$

Thus the lemma holds for $s = 0$.

Consider then $s > 0$ and $t > s$. Let $y_{0:t-1}$ denote all observations in the chain y until $t-1$. We may then use the law of total covariance to write

$$\begin{aligned}
&\mathrm{Cov}_{\mathrm{wide}}\left\{\psi_{t,k,j},\psi^*_{t-s,k',j'}\right\}\\
&\quad= \mathrm{Cov}_{\mathrm{wide}}\left\{\left(f^{(t)}_{kj}y_{tj} + g^{(t)}_{kj}\right)y_{t-1,k},\psi^*_{t-s,k',j'}\right\}\\
&\quad= \mathrm{E}_{\mathrm{wide}}\ \mathrm{Cov}_{\mathrm{wide}}\left\{\left(f^{(t)}_{kj}y_{tj} + g^{(t)}_{kj}\right)y_{t-1,k},\psi^*_{t-s,k',j'}\middle| y_{0:t-1}\right\}\\
&\qquad\qquad + \mathrm{Cov}_{\mathrm{wide}}\left\{\mathrm{E}_{\mathrm{wide}}\left(\left(f^{(t)}_{kj}y_{tj} + g^{(t)}_{kj}\right)y_{t-1,k}\middle| y_{0:t-1}\right),\right.\\
&\qquad\qquad\qquad\qquad\qquad\qquad \left.\mathrm{E}_{\mathrm{wide}}\left(\psi^*_{t-s,k',j'}\middle| y_{0:t-1}\right)\right\}\\
&\quad= \mathrm{Cov}_{\mathrm{wide}}\left\{\lambda^{(t)}_{kj}y_{t-1,k},\psi^*_{t-s,k',j'}\right\}.
\end{aligned}$$

The last equation follows since $\psi^*_{t-s,k',j'}$ is a constant when we condition $y_{0:t-1}$.

Now, if $s > 1$ we may use the law of total covariance again. This time conditioning on $y_{0:t-2}$. We have that

$$\mathrm{E}_{\mathrm{wide}}\left(y_{t-1,k} \mid y_{t-2,r} = 1\right) = \sum_{r=0}^{2}\pi_{rk}(x_{t-1})y_{t-2,r}.$$

With the same procedure of conditioning as in the last case, we get when we condition on $y_{0:t-2}$

$$\begin{aligned}
\mathrm{Cov}_{\mathrm{wide}}\left\{\psi_{t,k,j},\psi^*_{t-s,k',j'}\right\} &= \mathrm{Cov}_{\mathrm{wide}}\left\{\lambda^{(t)}_{kj}\sum_{r=0}^{2}\pi_{rk}(x_{t-1})y_{t-2,r},\psi^*_{t-s,k',j'}\right\}\\
&= \mathrm{Cov}_{\mathrm{wide}}\left\{\lambda^{(t)}_{kj}\sum_{r=0}^{2}P^{(1)}_{rk}(t-2)y_{t-2,r},\psi^*_{t-s,k',j'}\right\}.
\end{aligned}$$

Using this strategy of conditioning a total of s times, we will get

$$\begin{aligned}
&\mathrm{Cov}_{\mathrm{wide}}\left\{\psi_{t,k,j}, \psi^*_{t-s,k',j'}\right\} \\
&\quad= \mathrm{Cov}_{\mathrm{wide}}\left\{\lambda^{(t)}_{kj} \sum_{r=0}^{2} P^{(s-1)}_{rk}(t-s) y_{t-s,r}, \psi^*_{t-s,k',j'}\right\} \\
&\quad= \mathrm{Cov}_{\mathrm{wide}}\left\{\lambda^{(t)}_{kj} \sum_{r=0}^{2} P^{(s-1)}_{rk}(t-s) y_{t-s,r},\right. \\
&\qquad\qquad\qquad\qquad\left.\big(f^{*(t-s)}_{k'j'} y_{t-s,j'} + g^{*(t-s)}_{k'j'}\big) y_{t-s-1,k'}\right\}.
\end{aligned}$$

Conditioning again on y_{t-s-1}, we reach the expression

$$\begin{aligned}
&\mathrm{Cov}_{\mathrm{wide}}\left\{\psi_{t,k,j}, \psi^*_{t-s,k',j'}\right\} \\
&\quad= \lambda^{(t)}_{kj} f^{*(t-s)}_{k'j'} \sum_{r=0}^{2} P^{(s-1)}_{rk}(t-s) \Sigma_{rj'|k'}(t-s)\, \mathrm{E}_{\mathrm{wide}}\; y_{t-s-1,k'} \\
&\qquad+ \lambda^{(t)}_{kj} \lambda^{(t-s)}_{k'j'} \sum_{r=0}^{2} P^{(s)}_{rk}(t-s-1)\, \mathrm{Cov}_{\mathrm{wide}}\left\{y_{t-s-1,r}, y_{t-s-1,k'}\right\}.
\end{aligned} \tag{A.2}$$

Now, we have that

$$\begin{aligned}
&\sum_{r=0}^{2} P^{(s-1)}_{rk}(t-s-1)\, \mathrm{Cov}_{\mathrm{wide}}\left\{y_{t-s-1,r}, y_{t-s-1,k'}\right\} \\
&\quad= \sum_{r=0}^{1} P^{(s-1)}_{rk}(t-s-1)\, \mathrm{Cov}_{\mathrm{wide}}\left\{y_{t-s-1,r}, y_{t-s-1,k'}\right\} \\
&\qquad\qquad+ P^{(s-1)}_{2k}(t-s-1)\, \mathrm{Cov}_{\mathrm{wide}}\left\{1 - y_{t-s-1,0} - y_{t-s-1,1}, y_{t-s-1,k'}\right\} \\
&\quad= \sum_{r=0}^{1} \left(P^{(s-1)}_{rk}(t-s-1) - P^{(s-s-1)}_{2k}(t-1)\right) \mathrm{Cov}_{\mathrm{wide}}\left\{y_{t-s-1,r}, y_{t-s-1,k'}\right\} \\
&\quad= \sum_{r=0}^{1} \Phi^{(s-1)}_{rk}(t-s-1)\, \mathrm{Cov}_{\mathrm{wide}}\left\{y_{t-s-1,r}, y_{t-s-1,k'}\right\}.
\end{aligned} \tag{A.3}$$

Inserting this in (A.2) , we get

$$\begin{aligned}
&\mathrm{Cov}_{\mathrm{wide}}\left\{\psi_{t,k,j}, \psi^*_{t-s,k',j'}\right\} \\
&\quad = \lambda^{(t)}_{kj} f^{*(t-s)}_{k'j'} \sum_{r=0}^{1} \Phi^{(s-1)}_{rk}(t-s)\Sigma_{rj'|k'}(t-s)\,\mathrm{E}_{\mathrm{wide}}\; y_{t-s-1,k'} \\
&\qquad + \lambda^{(t)}_{kj}\lambda^{(t-s)}_{k'j'} \sum_{r=0}^{1} \Phi^{(s)}_{rk}(t-s-1)\,\mathrm{Cov}_{\mathrm{wide}}\left\{y_{t-s-1,r}, y_{t-s-1,k'}\right\}. \qquad (\mathrm{A.4})
\end{aligned}$$

We know from Section 2.4 that $|\Phi^{(s)}_{rk}(t)| < \kappa^s$ for all t, s, where κ is the *ergodic coefficient* of the inhomogenous Markov chain. Taking absolute values of (A.4), we have that there for all given covariates $x_0, x_1, \dots x_t$ and for each $s < t$, there is a constant G such that

$$\left|\mathrm{Cov}_{\mathrm{wide}}\left\{\psi_{t,k,j}, \psi^*_{t-s,k',j'}\right\}\right| \leq 2\left|\lambda^{(t)}_{kj} f^{*(t-s)}_{k'j'}\kappa^{s-1}\right| + 2\left|\lambda^{(t)}_{kj}\lambda^{(t-s)}_{k'j'}\kappa^{s}\right| < G\kappa^{s-1}.$$

Thus

$$\lim_{t\to\infty}\left|\sum_{s=0}^{t-1}\mathrm{Cov}_{\mathrm{wide}}\left\{\psi_{t,k,j}, \psi^*_{t-s,k',j'}\right\}\right| < G + G\sum_{s=1}^{\infty}\kappa^{s-1} < \infty$$

as $\kappa < 1$. This proves the theorem.

■

A.2 Unconditional Variance of Response Variables

In this section we find exact expressions for the covariance between $\psi_{t,k,j}$ and $\psi_{t,k',j'}$.

Lemma A.2.1. *Let covaritate vectors $x_0, x_1, \dots, x_t$ be generated by some unknown covariate distribution C(x) in accordance with the assumptions in Section 2.3. Given these covariate values, let $\{y_{i,t}\}$ be a Markov chain generated by the wide model. For all given covariates $x_0, x_1, \dots x_t$, it is for each s such that $0 \leq s < t$ the case that*

$$\begin{aligned}
&\mathrm{Cov}_{wide}\left\{y_{t-s,j}, y_{t-s,j'}\right\} \\
&\quad = \sum_{w=0}^{t-s-1}\sum_{l=0}^{2}\sum_{k=0}^{1}\sum_{k'=0}^{1}\Phi^{(w)}_{kj}(t-s-w)\Phi^{(w)}_{k'j'}(t-s-w) \\
&\qquad\qquad\qquad \cdot\Sigma_{kk'|l}(t-s-w)\,\mathrm{E}_{wide}\; y_{t-s-w-1,l} \\
&\qquad + \sum_{k=0}^{1}\sum_{k'=0}^{1}\Phi^{(t-s)}_{kj}(0)\Phi^{(t-s)}_{k'j'}(0)\,\mathrm{Cov}_{wide}\left\{y_{0,k}, y_{0,k'}\right\}. \qquad (\mathrm{A.5})
\end{aligned}$$

Proof. We use the law of total covariance to write

$$\begin{aligned}
&\mathrm{Cov}_{\text{wide}}\left\{ y_{t-s,j}, y_{t-s,j'} \right\} \\
&\quad = \mathrm{E}_{\text{wide}}\ \mathrm{Cov}_{\text{wide}}\left\{ y_{t-s,j}, y_{t-s,j'} \middle| y_{t-s-1} \right\} \\
&\quad + \mathrm{Cov}_{\text{wide}}\left\{ \mathrm{E}_{\text{wide}}\left\{ y_{t-s,j} \middle| y_{t-s-1} \right\}, \mathrm{E}_{\text{wide}}\left\{ y_{t-s,j'} \middle| y_{t-s-1} \right\} \right\}.
\end{aligned}$$

For the first term write

$$\mathrm{E}_{\text{wide}}\ \mathrm{Cov}_{\text{wide}}\left\{ y_{t-s,j}, y_{t-s,j'} \middle| y_{t-s-1} \right\} = \sum_{k=0}^{2} \Sigma_{jj'|k}(t-s)\, \mathrm{E}_{\text{wide}}\ y_{t-s-1,k}.$$

For the second term write

$$\begin{aligned}
&\mathrm{Cov}_{\text{wide}}\left\{ \mathrm{E}_{\text{wide}}\left\{ y_{t-s,j} \middle| y_{t-s-1,k} \right\}, \mathrm{E}_{\text{wide}}\left\{ y_{t-s,j'} \middle| y_{t-s-1,k'} \right\} \right\} \\
&\quad = \mathrm{Cov}_{\text{wide}}\left\{ \sum_{k=0}^{2} \pi_{kj}(x_{t-s}) y_{t-s-1,k}, \sum_{k'=0}^{2} \pi_{k'j'}(x_{t-s}) y_{t-s-1,k'} \right\} \\
&\quad = \sum_{k=0}^{2} \sum_{k'=0}^{2} P_{kj}^{(1)}(t-s-1) P_{k'j'}^{(1)}(t-s-1) \\
&\qquad\qquad\qquad \cdot \mathrm{Cov}_{\text{wide}}\left\{ y_{t-s-1,k}, y_{t-s-1,k'} \right\}.
\end{aligned}$$

Conditioning further, we get

$$\begin{aligned}
&\mathrm{Cov}_{\text{wide}}\left\{ y_{t-s,j}, y_{t-s,j'} \right\} \\
&\quad = \sum_{w=0}^{t-s-1} \sum_{l=0}^{2} \sum_{k=0}^{2} \sum_{k'=0}^{2} P_{kj}^{(w)}(t-s-w) P_{k'j'}^{(w)}(t-s-w) \\
&\qquad\qquad\qquad \cdot \Sigma_{kk'|l}(t-s-w)\, \mathrm{E}_{\text{wide}}\ y_{t-s-w-1,l} \\
&\qquad + \sum_{k'=0}^{2} \sum_{k=0}^{2} P_{kj}^{(t-s)}(0) P_{k'j'}^{(t-s)}(0)\, \mathrm{Cov}_{\text{wide}}\left\{ y_{0,k}, y_{0,k'} \right\}.
\end{aligned}$$

From (A.3) we know that this is equivalent to the expression stated in the lemma. ■

Inserting (A.5) into (A.1), we we get the following general expression for s=0:

$$
\begin{aligned}
&\mathrm{Cov}_{\mathrm{wide}}\left\{\psi_{t,k,j}, \psi^*_{t,k',j'}\right\} \\
&\quad = f^{(t)}_{kj} f^{*(t)}_{kj} \Sigma_{jj'|k}(t)\delta_{kk'}\, \mathrm{E}_{\mathrm{wide}}\; y_{t-1,k'} \\
&\qquad + \lambda^{(t)}_{kj}\lambda^{*(t)}_{k'j'} \sum_{w=0}^{t-s-1} \sum_{l=0}^{2}\sum_{r=0}^{1}\sum_{r'=0}^{1} \Phi^{(w)}_{rk}(t-s-w)\Phi^{(w)}_{r'k'}(t-s-w) \\
&\qquad\qquad \cdot \Sigma_{rr'|l}(t-s-w)\, \mathrm{E}_{\mathrm{wide}}\; y_{t-s-w-1,l} \\
&\qquad + \lambda^{(t)}_{kj}\lambda^{*(t)}_{k'j'} \sum_{r=0}^{1}\sum_{r'=0}^{1} \Phi^{(t)}_{rk}(0)\Phi^{(t)}_{r'k'}(0)\, \mathrm{Cov}_{\mathrm{wide}}\left\{y_{0,r}, y_{0,r'}\right\},
\end{aligned}
$$

Inserting (A.5) into (A.4), we we get the following general expression for $s > 0$:

$$
\begin{aligned}
&\mathrm{Cov}_{\mathrm{wide}}\left\{\psi_{t,k,j}, \psi^*_{t,k',j'}\right\} \\
&\quad = \lambda^{(t)}_{kj} f^{*(t-s)}_{k'j'} \sum_{r=0}^{2} P^{(s-1)}_{rk}(t-s)\Sigma_{rj'|k'}\, \mathrm{E}_{\mathrm{wide}}\; y_{t-s-1,k'} \\
&\qquad + \lambda^{(t)}_{kj}\lambda^{*(t-s)}_{k'j'} \sum_{w=0}^{t-s-1} \sum_{l=0}^{2}\sum_{r=0}^{1}\sum_{r'=0}^{1} \Phi^{(s+w)}_{rk}(t-s-w)\Phi^{(w)}_{r'k'}(t-s-w) \\
&\qquad\qquad \cdot \Sigma_{rr'|l}(t-s-w)\, \mathrm{E}_{\mathrm{wide}}\; y_{t-s-w-1,l} \\
&\qquad + \lambda^{(t)}_{kj}\lambda^{*(t-s)}_{k'j'} \sum_{r=0}^{1}\sum_{r'=0}^{1} \Phi^{(t)}_{rk}(0)\Phi^{(t-s)}_{r'k'}(0)\, \mathrm{Cov}_{\mathrm{wide}}\left\{y_{0,r}, y_{0,r'}\right\}.
\end{aligned}
$$

Both for $s = 0$ and $s > 0$ the last term is negligible when t is of moderate size or higher.

A.3 Closeness to Finite Functions

The following lemma is used in the proof of Theorem 3.2.1.

Lemma A.3.1. *Let covaritate vectors $x_0, x_1, \ldots, x_t$ be generated by some unknown covariate distribution C(x) in accordance with the assumptions in Section 2.3. Given these covariate values, let $\{y_{i,t}\}$ be a Markov chain generated by the wide model. Let $\psi_{t,k,j}$ be a function in accordance with the definition above. For each $\epsilon > 0$ there exists an N and a function $f_N(x_t, \ldots, x_{t-N})$ such that for all covariates $x_0, x_1, \ldots x_t$ it is the case that*

$$
\lim_{t\to\infty} P\left(\left|\sum_{s=1}^{\infty} \mathrm{Cov}_{wide}\left\{\psi_{t,k,j}, \psi^*_{t-s,k',j'}\right\} - f_N(x_t, \ldots x_{t-N})\right| \geq \epsilon\right) = 0.
$$

Proof. Let N_1, N_2, N_3, N_4 be integers. Define $N = \max(N_1 + N_2, N_1 + N_3 + N_4)$. Define for readability the variables $t_{\text{past},1} = t - s - N_2$, $t_{\text{past},2} = t - s - w - 2 - N_4$ and $t_{sw} = t - s - w$. Define for $k, k', j, j' = 0, 1, 2$ the function

$$\begin{aligned}
&f(x_t, \dots x_{t-N}) \\
&\quad = \sum_{s=1}^{N_1} \Big\{ \lambda_{kj}^{(t)} f_{k'j'}^{*(t-s)} \sum_{r=0}^{1} \Phi_{rk}^{(s-1)}(t-s) \Sigma_{rj'|k'}(t-s) P_{k'k'}^{(N_2)}(t_{\text{past},1}) \\
&\quad + \lambda_{kj}^{(t)} \lambda_{k'j'}^{*(t-s)} \sum_{w=0}^{N_3} \sum_{l=0}^{2} \sum_{q=0}^{2} \sum_{r=0}^{1} \Phi_{rk}^{(s+w)}(t_{sw}-1) \Phi_{qk'}^{(w)}(t_{sw}-1) \\
&\qquad \cdot \Sigma_{rq|l}(t_{sw}-1) P_{ll}^{(N_4)}(t_{\text{past},2}) \Big\}.
\end{aligned}$$

Consider for each $t \geq N$

$$\begin{aligned}
&\left| \sum_{s=1}^{\infty} \mathrm{Cov}_{\text{wide}} \left\{ \psi_{t,k,j}, \psi_{t-s,k',j'}^{*} \right\} - f_N(x_t, \dots x_{t-N}) \right| \\
&= \Bigg| \sum_{s=N_1+1}^{\infty} \mathrm{Cov}_{\text{wide}} \left\{ \psi_{t,k,j}, \psi_{t-s,k',j'}^{*} \right\} \\
&\quad + \sum_{s=1}^{N_1} \lambda_{kj}^{(t)} f_{k'j'}^{*(t-s)} \sum_{r=0}^{1} \Phi_{rk}^{(s-1)}(t-s) \Sigma_{rj'|k'}(t-s) \\
&\qquad \cdot \left\{ \mathrm{E}_{\text{wide}} \; y_{t-s-1,k'} - P_{k'k'}^{(N_2)}(t_{\text{past},1}) \right\} \\
&\quad + \sum_{s=1}^{N_1} \lambda_{kj}^{(t)} \lambda_{k'j'}^{*(t-s)} \sum_{w=N_3+1}^{\infty} \sum_{l=0}^{2} \sum_{q=0}^{2} \sum_{r=0}^{1} \Phi_{rk}^{(s+w)}(t_{sw}-1) \Phi_{qk'}^{(w)}(t_{sw}-1) \\
&\qquad \cdot \Sigma_{rq|l}(t_{sw}-1) \, \mathrm{E}_{\text{wide}} \; y_{t_{sw}-2,l} \\
&\quad + \sum_{s=1}^{N_1} \lambda_{kj}^{(t)} \lambda_{k'j'}^{*(t-s)} \sum_{w=0}^{N_3} \sum_{l=0}^{2} \sum_{q=0}^{2} \sum_{r=0}^{1} \Phi_{rk}^{(s+w)}(t_{sw}-1) \Phi_{qk'}^{(w)}(t_{sw}-1) \\
&\qquad \cdot \Sigma_{rq|l}(t_{sw}-1) \mathrm{E}_{\text{wide}} \; y_{t_{sw}-2,l} - P_{ll}^{(N_4)}(t_{\text{past},2}) \Bigg|.
\end{aligned}$$

By repetead us of the triangle equality, we get

$$
\begin{aligned}
&\left|\sum_{s=1}^{\infty} \mathrm{Cov}_{\mathrm{wide}}\left\{\psi_{t,k,j}, \psi^{*}_{t-s,k',j'}\right\} - f(x_t, \dots x_{t-N})\right| \\
&\leq \left|\sum_{s=N_1+1}^{\infty} \mathrm{Cov}_{\mathrm{wide}}\left\{\psi_{t,k,j}, \psi^{*}_{t-s,k',j'}\right\}\right| \\
&+ \sum_{s=1}^{N_1}\left|\lambda^{(t)}_{kj} f^{*(t-s)}_{k'j'} \sum_{r=0}^{1} \Phi^{(s-1)}_{rk}(t-s)\Sigma_{rj'|k'}\right|\left|\mathrm{E}_{\mathrm{wide}}\ y_{t-s-1,k'} - P^{(N_2)}_{k'k'}(t_{\mathrm{past},1})\right| \\
&+ \sum_{s=1}^{N_1}\Bigg|\lambda^{(t)}_{kj} \lambda^{*(t-s)}_{k'j'} \sum_{w=N_3+1}^{\infty} \sum_{l=0}^{2}\sum_{q=0}^{2}\sum_{r=0}^{1} \Phi^{(s+w)}_{rk}(t_{sw}-1)\Phi^{(w)}_{qk'}(t_{sw}-1) \\
&\qquad\qquad\qquad\qquad\qquad \cdot \Sigma_{rq|l}(t_{sw}-1)\, \mathrm{E}_{\mathrm{wide}}\ y_{t_{sw}-2,l}\Bigg| \\
&+ \sum_{s=1}^{N_1}\left|\lambda^{(t)}_{kj} \lambda^{*(t-s)}_{k'j'} \sum_{w=0}^{N_3} \sum_{l=0}^{2}\sum_{q=0}^{2}\sum_{r=0}^{1} \Phi^{(s+w)}_{rk}(t_{sw}-1)\Phi^{(w)}_{qk'}(t_{sw}-1)\Sigma_{rq|l}(t_{sw}-1)\right| \\
&\qquad\qquad\qquad\qquad\qquad \cdot \left|\mathrm{E}_{\mathrm{wide}}\ y_{t_{sw}-2,l} - P^{(N_4)}_{ll}(t_{\mathrm{past},2})\right|.
\end{aligned}
$$

We know that $f^*_{kj}(t)$,$\lambda_{kj}(t)$ and $\lambda^*_{kj}(t)$ are uniformly bounded for all k, j, t. In accordance with the assumptions on $C(x)$ in Section 2.3 we may then for each $\epsilon > 0$ find an $N_1 \in \mathbb{N}$ such that

$$
\lim_{t\to\infty} P\left(\left|\sum_{s=N_1+1}^{t-1} \mathrm{Cov}_{\mathrm{wide}}\left\{\psi_{t,k,j}, \psi^{*}_{t-s,k',j'}\right\}\right| \geq \frac{\epsilon}{4}\right) = 0.
$$

For such a pair ϵ and N_1, we may also find $N_3 \in \mathbb{N}$ such that for each $s \leq N_1$ it is the case that

$$
\begin{aligned}
\lim_{t\to\infty} P\Bigg(\Bigg|\lambda^{(t)}_{kj} \lambda^{*(t-s)}_{k'j'} &\sum_{w=N_3+1}^{t-s-2} \sum_{l=0}^{2}\sum_{q=0}^{2}\sum_{r=0}^{1} \Phi^{(s+w)}_{rk}(t_{sw}-1)\Phi^{(w)}_{qk'}(t_{sw}-1) \\
&\Sigma_{rq|l}(t_{sw}-1)\, \mathrm{E}_{\mathrm{wide}}\ y_{t_{sw}-2,l}\Bigg| < \frac{\epsilon}{4N_1}\Bigg) = 0.
\end{aligned}
$$

From Section 2.4 we know that $\lim_{t\to\infty} \mathrm{E}_{\mathrm{wide}}\ y_{t,k}$ is the limiting probability π_k which exists as the chain is *strongly ergodic*. We have that $\pi_k = \lim_{s\to\infty} \mathbb{P}^{(s)}_{kk}(t)$ for all t.

From the assumptions on the covariate distribution in REF in Section 2.3, we see that this implies that for all $x_0, x_1, \dots x_t$ it is the case for every $k = 0, 1, 2$ there exists for each $\epsilon^* > 0$ an $N^* \in \mathbb{N}$ such that

$$
\lim_{t\to\infty} P\left(\left|\mathrm{E}_{\mathrm{wide}}\ y_{t,k} - P^{(N^*)}_{kk}(t - N^*)\right| \geq \epsilon^*\right) = 0.
$$

By using the triangle inequality, it follows that there for each $\epsilon > 0$ is an integer $N_2 \in \mathbb{N}$ such that

$$\lim_{t\to\infty} P\Bigg(\bigg|\lambda_{kj}^{(t)} f_{k'j'}^{*(t-s)} \sum_{r=0}^{1} \Phi_{rk}^{(s-1)}(t-s)\Sigma_{rj'|k'}(t-s)\bigg| \\ \bigg|\mathrm{E}_{\text{wide}}\ y_{t-s-1,k'} - P_{k'k'}^{(N_2)(t_{\text{past},1}}\bigg| \geq \frac{\epsilon}{4N_1}\Bigg) = 0.$$

Equivalently, there exists an integer $N_4 \in \mathbb{N}$ such that

$$\lim_{t\to\infty} P\Bigg(\bigg|\lambda_{kj}^{(t)} \lambda_{k'j'}^{*(t-s)} \sum_{w=0}^{N_3}\sum_{l=0}^{2}\sum_{q=0}^{2}\sum_{r=0}^{1} \Phi_{rk}^{(s+w)}(t_{sw}-1)\Phi_{qk'}^{(w)}(t_{sw}-1)\Sigma_{rq|l}(t_{sw}-1)\bigg| \\ \cdot \bigg|\mathrm{E}_{\text{wide}}\ y_{t_{sw}-2,l} - P_{ll}^{(N_4)}(t_{\text{past}_2})\bigg| \geq \frac{\epsilon}{4N_1}\Bigg) = 0.$$

Using the the triangle equality repeatedly, it then follows that there for each $\epsilon > 0$ there exist and $N \in \mathbb{N}$ such that

$$\lim_{t\to\infty} P\Bigg(\bigg|\sum_{s=1}^{\infty} \mathrm{Cov}_{\text{wide}}\left\{\psi_{t,k,j}, \psi_{t-s,k',j'}^{*}\right\} - f_N(x_t, \ldots x_{t-N})\bigg| \geq \epsilon\Bigg) = 0.$$

as is what we set out to prove.

■

A.4 Expressions for covariance matrices

In this section we give expressions of the $K_{M,n}$ matrix defined in Section 3.2. Recall that this matrix is on the form

$$K_{M,n} = J_{M,n}^{*} + V_{M,n} + (W_{M,n} + W_{M,n}^{t}) + (Q_{M,n} + Q_{M,n}^{t}).$$

Using the expressions in *Appendix A*.2 we may now, *nicht ganz ohne mühe*, find expressions for each of these matrices.
The first matrix $J_{M,n}^{*}$ is the matrix

$$J_{M,n}^{*} = \frac{1}{n}\sum_{t=1}^{t} \begin{pmatrix} J_{M,\gamma,t}^{*} & J_{M,\gamma b_0,t}^{*} & J_{M,\gamma b_1,t}^{*} & J_{M,\gamma b_2,t}^{*} \\ J_{M,b_0\gamma,t}^{*} & J_{M,b_0,t}^{*} & 0 & 0 \\ J_{M,b_1\gamma,t}^{*} & 0 & J_{M,b_1,t}^{*} & 0 \\ J_{M,b_2\gamma,t}^{*} & 0 & 0 & J_{M,b_2,t}^{*} \end{pmatrix},$$

with blocks

$$J^*_{M,\gamma,t} = \sum_{k=0}^{2} \begin{pmatrix} u_{M,t}u^{\mathrm{t}}_{M,t}\pi_{k,0}(1-\pi_{k,0}) & -u_{M,t}u^{\mathrm{t}}_{M,t}\pi_{k,0}\pi_{k,1} \\ -u_{M,t}u^{\mathrm{t}}_{M,t}\pi_{k,0}\pi_{k,1} & u_{M,t}u^{\mathrm{t}}_{M,t}\pi_{k,0}(1-\pi_{k,0}) \end{pmatrix} \mathrm{E}_{\text{wide}}\ y_{t-1,k},$$

and for $k = 0, 1, 2$

$$J^*_{M,b_k,t} = \begin{pmatrix} z_M z^{\mathrm{t}}_M \pi_{k,0}(1-\pi_{k,0}) & -z_M z^{\mathrm{t}}_M \pi_{k,0}\pi_{k,1} \\ -z_M z^{\mathrm{t}}_M \pi_{k,0}\pi_{k,1} & z_M z^{\mathrm{t}}_M \pi_{k,0}(1-\pi_{k,0}) \end{pmatrix} \mathrm{E}_{\text{wide}}\ y_{t-1,k},$$

$$J^*_{M,\gamma b_k,t} = \begin{pmatrix} u_{M,t} z^{\mathrm{t}}_M \pi_{k,0}(1-\pi_{k,0}) & -u_M z^{\mathrm{t}}_M \pi_{k,0}\pi_{k,1} \\ -u_{M,t} z^{\mathrm{t}}_M \pi_{k,0}\pi_{k,1} & u_M z^{\mathrm{t}}_M \pi_{k,0}(1-\pi_{k,0}) \end{pmatrix} \mathrm{E}_{\text{wide}}\ y_{t-1,k},$$

and $J^*_{M,b_k\gamma}(x_M) = J^*_{M,\gamma b_k}(x_t)^{\mathrm{t}}$. The probability π_{kj} is an abbreviation for $\pi_{kj}(x_t)$. The matrix $J^*_{M,n}$ resembles the Fisher information matrix $J_{M,n}$ of the candidate model. It is however not the same. It is important that these matrices are not mixed. The matrix $J^*_{M,n}$ uses the *true* probabilities $\pi_{kj}(x_t)$, whereas the candidate Fisher information matrix $J_{M,n}$ uses the probabilities $\pi_{M,kj}(x_{M,t})$ of the candidate model.

The second matrix $V_{M,n}$ is

$$V_{M,n} = \frac{1}{n}\sum_{t=1}^{n} \begin{pmatrix} V_{\gamma,t}(x_t) & V_{\gamma b_0,t} & V_{\gamma b_1,t} & V_{\gamma b_2,t} \\ V_{b_0\gamma,t} & V_{00,t} & V_{01,t} & V_{02,t} \\ V_{b_1\gamma,t} & V_{10,t} & V_{11,t} & V_{12,t} \\ V_{b_1\gamma,t} & V_{20,t} & V_{21,t} & V_{22,t} \end{pmatrix},$$

where

$$V_{\gamma,t} = \sum_{k=0}^{2}\sum_{k'=0}^{2} \nu_{k',k'}(u_{M,t}, u_{M,t}) \qquad V_{\gamma b_{k'},t} = \sum_{k=0}^{2} \nu_{k,k'}(u_{M,t}, z_{M,t})$$

$$V_{b_k\gamma,t} = \sum_{k'=0}^{2} \nu_{k',k'}(z_{M,t}, u_{M,t}) \qquad V_{k,k',t} = \nu_{k',k'}(z_{M,t}, z_{M,t}),$$

and

$$\nu_{k,k'}(d_{M,t}, h_{M,t}) = \begin{pmatrix} \phi_{k,0}\phi_{k',0}(x_{M,t})d_{M,t}h^{\mathrm{t}}_{M,t}\xi^{(t)}_{k,k'} & \phi_{k,0}\phi_{k',1}(x_{M,t})d_{M,t}h^{\mathrm{t}}_{M,t}\xi^{(t)}_{k,k'} \\ \phi_{k,1}\phi_{k',0}(x_{M,t})d_{M,t}h^{\mathrm{t}}_{M,t}\xi^{(t)}_{k,k'} & \phi_{k,1}\phi_{j,1}(x_{M,t})d_{M,t}h^{\mathrm{t}}_{M,t}\xi^{(t)}_{k,k'} \end{pmatrix}.$$

The vectors $d_{M,t}$,$h_{M,t}$ are here either $u_{M,t}$ or $z_{M,t}$ and $\xi_{k,k'}(t)$ is

$$\xi^{(t)}_{k,k'} = \sum_{w=0}^{t-2} \sum_{l=0}^{2} \sum_{q,r=0}^{1} \Phi^{(w)}_{rk}(t-w-1)\Phi^{(w)}_{qk'}(t-w-1) \\ \cdot \Sigma_{rq|l}(t-w-1)\, \mathrm{E}_{\text{wide}}\; y_{t-w-2,l}.$$

The third matrix $W_{M,n}$ is

$$W_{M,n} = \frac{1}{n} \sum_{t=2}^{n} \sum_{s=1}^{t-1} \begin{pmatrix} W^{(s)}_{\gamma,t} & W^{(s)}_{b_0\gamma,t} & W^{(s)}_{b_1\gamma,t} & W^{(s)}_{b_2\gamma,t} \\ W^{(s)}_{\gamma b_0,t} & W^{(s)}_{00,t} & W^{(s)}_{01,t} & W^{(s)}_{02,t} \\ W^{(s)}_{\gamma b_1,t} & W^{(s)}_{10,t} & W^{(s)}_{11,t} & W^{(s)}_{12,t} \\ W^{(s)}_{\gamma b_2,t} & W^{(s)}_{20,t} & W^{(s)}_{21,t} & W^{(s)}_{22,t} \end{pmatrix},$$

with block matrices

$$\begin{aligned} W^{(s)}_{\gamma,t} &= \sum_{k=0}^{2} \sum_{k'=0}^{2} \Omega_{k,k'}(u_{M,t}, u_{M,t-s}), \\ W^{(s)}_{b_{k'}\gamma,t} &= \sum_{k=0}^{2} \Omega_{k,k'}(u_{M,t}, z_{M,t-s}), \\ W^{(s)}_{\gamma b_k,t} &= \sum_{k'=0}^{2} \Omega_{k,k'}(z_{M,t}, u_{M,t-s}), \\ W^{(s)}_{kk',t} &= \Omega_{k,k'}(z_{M,t}, z_{M,t-s}), \end{aligned}$$

where

$$\Omega_{k,k'}(d_{M,t}, h_{M,t-s}) = \\ \begin{pmatrix} \phi_{k0}(x_{M,t}) d_{M,t} h^{\mathrm{t}}_{M,t-s} \omega_{0,k,k'}(t-s) & \phi_{k0}(x_{M,t}) d_{M,t} h^{\mathrm{t}}_{M,t-s} \omega_{1,k,k'}(t-s) \\ \phi_{k1}(x_{M,t}) d_{M,t} h^{\mathrm{t}}_{M,t-s} \omega_{0,k,k'}(t-s) & \phi_{k1}(x_{M,t}) d_{M,t} h^{\mathrm{t}}_{M,t-s} \omega_{1,k,k'}(t-s) \end{pmatrix},$$

and

$$\omega_{j,k,k'}(t-s) = \sum_{r=0}^{1} \Phi^{s-1}_{rk}(t-s) \Sigma_{rj|k'}(t-s) \mathrm{E}_{\text{wide}} y_{t-s-1,k'}.$$

The last matrix $Q_{M,n}$ is given by

$$Q_{M,n} = \frac{1}{n} \sum_{t=2}^{n} \sum_{s=1}^{t-2} \begin{pmatrix} Q^{(s)}_{\gamma,t} & Q^{(s)}_{b_0\gamma,t} & Q^{(s)}_{b_1\gamma} & Q^{(s)}_{b_2\gamma,t} \\ Q^{(s)}_{\gamma b_0,t} & Q^{(s)}_{00,t} & Q^{(s)}_{10,t} & Q^{(s)}_{20,t} \\ Q^{(s)}_{\gamma b_1,t} & Q^{(s)}_{10,t} & Q^{(s)}_{11,t} & Q^{(s)}_{21,t} \\ Q^{(s)}_{\gamma b_2,t} & Q^{(s)}_{20,t} & Q^{(s)}_{21,t} & Q^{(s)}_{22,t} \end{pmatrix},$$

with blocks

$$\begin{aligned}
&Q^{(s)}_{\gamma,t} = \sum_{k=0}^{2}\sum_{k'=0}^{2}\Theta_{k,k'}(u_{M,t},u_{M,t-s}),\\
&Q^{(s)}_{k,k',t}\Theta_{k,k'}(z_{M,t},z_{M,t-s}),\\
&Q^{(s)}_{\gamma b_{k'},t} = \sum_{k=0}^{2}\Theta_{k,k'}(u_{M,t},z_{M,t-s}),\\
&Q^{(s)}_{b_k\gamma,t}\sum_{k'=0}^{2}\Theta_{k,k'}(z_{M,t},u_{M,t-s}),
\end{aligned}$$

and

$$\begin{aligned}
&\Theta_{k,k'}(d_{M,t},h_{M,t-s})\\
&\quad=\begin{pmatrix}\phi_{k0}(x_{M,t})\phi_{k'0}(x_{M,t-s})\xi^{(t-s)}_{k,k'} & \phi_{k0}(x_{M,t})\phi_{k'1}(x_{M,t-s})\xi^{(t-s)}_{k,k'}\\ \phi_{k1}(x_{M,t})\phi_{k'0}(x_{M,t-s})\xi^{(t-s)}_{k,k'} & \phi_{k1}(x_{M,t})\phi_{k'1}(x_{M,t-s})\xi^{(t-s)}_{k,k'}\end{pmatrix},
\end{aligned}$$

where

$$\begin{aligned}
\xi^{(t-s)}_{k,k'} = d_{M,t}h^{\mathrm{t}}_{M,t-s}sum_{w=0}^{t-s-2}\sum_{l=0}^{2}\sum_{q=0}^{1}\sum_{r=0}^{1}\Phi^{(s+w)}_{rk}(t-s-w-1)\\
\Phi^{(w)}_{qk'}(t-s-w-1)\Sigma_{rq|l}(t-s-w-1)\,\mathrm{E}_{\mathrm{wide}}\;y_{t-s-w-2,l}.
\end{aligned}$$

A.5 A Note on R Scripts

In this thesis I have used the statistical programming language R (R Development Core Team, 2008) for simulations and analysis of the MID data. The scripts I have used are too comprehensive to be included in the thesis. The R-scripts are available upon request to `j.k.haug@gmail.com`.

Bibliography

Aalen, O. O., & Johansen, S. (1978). An empirical transition matrix for non-homogeneous Markov chains based on censored observations. *Scand. J. Statist. 5*(3), 141–150.

Agresti, A. (2013). *Categorical Data Analysis* (3rd edition). Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken.

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models.* Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken.

Basawa, I. V., & Prakasa Rao, B. L. S. (1980). *Statistical Inference for Stochastic Processes.* Probability and Mathematical Statistics. Academic Press, London.

Beck, R., & Katz, J. (1998). Taking time seriously: Time-series-cross-section analysis with a binary dependent variable. *American Journal of Political Science*, *42*(4), 1260–88.

Berchtold, A., & Raftery, A. E. (2002). The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. *Statist. Sci. 17*(3), 328–356.

Billingsley, P. (1961a). *Statistical Inference for Markov Processes.* Statistical Research Monographs, Vol. II. The University of Chicago Press, Chicago.

Billingsley, P. (1961b). Statistical methods in Markov chains. *Ann. Math. Statist. 32*, 12–40.

Bolt, J., Inklaar, R., de Jong, H., & van Zanden, J. (2018). Rebasing 'Maddison': New income comparisons and the shape of long-run economic development. Maddison Project Working paper 10.

Brillinger, D. R., Morettin, P. A., Irizarry, R. A., & Chiann, C. (2000). Some wavelet-based analyses of Markov chain data. *Signal Processing*, *80*, 1607–1627.

Brillinger, D. R. (1996). An analysis of an ordinal-valued time series. In *Athens Conference on Applied Probability and Time Series Analysis, Vol. II (1995)* (Vol. 115, pp. 73–87). Lect. Notes Stat.

Casella, G., & Berger, R. L. (2002). *Statistical Inference* (2nd edition). Duxbury, Pacific Grove.

Claeskens, G., Cunen, C., & Hjort, N. (2019). Model selection via focused information criteria for complex data in ecology and evolution. *Frontiers*. Submitted for publication.

Claeskens, G., & Hjort, N. L. (2008a). Minimizing average risk in regression models. *Econometric Theory*, *24*(2), 493–527.

Claeskens, G., & Hjort, N. L. (2008b). *Model Selection and Model Averaging*. Cambridge Series in Statistical and Probabilistic Mathematics.

Clauset, A. (2017). The enduring threat of a large interstate war. Tech. rep.

Clauset, A. (2018). Trends and fluctuations in the severity of interstate wars. *Science Advances*, *4*.

Correlates of War Project. (2017). State system membership list.

Cranmer, S., Desarmais, B., & Menninga, E. (2016). A critique of dyadic design. *International Studies Quarterly*, *60*(2), 355–362.

Cunen, C., Hjort, N., & Nygård, H. (2019). Statistical sightings of better angels: Analysing the distribution of battle deathsin interstate conflict over time. Submitted for publication.

Cunen, C., Walløe, L., & Hjort, N. (2019). Focused model selection for linear mixed models, with an application to whale ecology. Submitted for publication.

Dobrushin, R. (1956). Central limit theorem for non-stationary Markov chains. I,II. *Teor. Veroyatnost. i Primenen. 1*, 72–89, 365–425.

Fahrmeir, L., & Kaufmann, H. (1987). Regression models for nonstationary categorical time series. *J. Time Ser. Anal. 8*(2), 147–160.

Fokianos, K., & Kedem, B. (1998). Prediction and classification of non-stationary categorical time series. *J. Multivariate Anal. 67*(2), 277–296.

Fokianos, K., & Kedem, B. (2003). Regression theory for categorical time series. *Statist. Sci. 18*(3), 357–376.

Gaddis, J. (1989). *The Long Peace: Inquiries into the History of the Cold War*. Oxford University Press, Oxford.

Gartzke, E. (2007). The capitalist peace. *51*(1), 166–191.

Gat, A. (2006). *War in Human Civilization*. Oxford University Press, Oxford.

Gibler, D. (2009). *International military alliances 1648-2008*. CQ Press.

Goldstein, J. S. (2011). *Winning the War on War: The Decline of Armed Conflict Worldwide*. Penguin Press, New York.

Gowa, J. (1999). *Ballots and Bullets: The Elusive Democratic Peace*. Princeton University Press, New Jersey.

Hajnal, J. (1956). The ergodic properties of non-homogeneous finite Markov chains. *Proc. Cambridge Philos. Soc. 52*, 67–77.

Hajnal, J. (1958). Weak ergodicity in non-homogeneous Markov chains. *Proc. Cambridge Philos. Soc. 54*, 233–246.

Hall, P., & Heyde, C. C. (1980). *Martingale Limit Theory and its Application*. Probability and Mathematical Statistics. Academic Press, London.

Hegre, H. (2014). Democracy and armed conflict. *Journal of Peace Research*, *51*(2), 159–172.

Hegre, H., Karlsen, J., Nygård, H. M., Strand, H., & Urdal, H. (2013). Predicting armed conflict, 2010–2050. *International Studies Quarterly*, *57*(2), 250–270.

Hjort, N. L., & Varin, C. (2008). ML, PL, QL in Markov chain models. *Scand. J. Statist. 35*(1), 64–82.

Hjort, N., & Claeskens, G. (2006). Focused information criteria and model averaging for the Cox hazard regression model. *J. Amer. Statist. Assoc. 101*(476), 1449–1464.

Hjort, N. L., & Claeskens, G. (2003). Frequentist model average estimators. *J. Amer. Statist. Assoc. 98*(464), 879–899.

Hjort, N. L., & Pollard, D. (1993). Asymptotics for minimisers of convex processes. Statistical research report, Department of Mathematics, University of Oslo.

Hubbard, R. A., Inoue, L. Y. T., & Fann, J. R. (2008). Modeling nonhomogeneous Markov processes via time transformation. *Biometrics*, *64*(3), 843–850.

Jullum, M., & Hjort, N. L. (2017). Parametric or nonparametric: The FIC approach. *Statist. Sinica*, *27*(3), 951–981.

Karlin, S., & Taylor, H. M. (1975). *A First Course in Stochastic Processes* (2nd edition). Academic Press, London.

Kaufmann, H. (1987). Regression models for nonstationary categorical time series: Asymptotic estimation theory. *Ann. Statist. 15*(1), 79–98.

Kedem, B., & Fokianos, K. (2002). *Regression Models for Time Series Analysis*. Wiley Series in Probability and Statistics.

Ko, V., Hjort, N., & Hobæk Haff, I. (2019). Focused information criteria for copulae. *Scandinavian Journal of Statistics*. Submitted for publication.

Lemke, D., & Reed, W. (2001). The relevance of politically relevant dyads. *Journal of Conflict Resolution*, *45*(1), 126–144.

Levy, J. (1989). The caues of war. In *Behaviour, Society and Nuclear war* (pp. 209–313). Oxford University Press, New York.

Maoz, Z., Johnson, P., Kaplan, J., Ogunkoya, F., & Shreve, A. (2018). The dyadic militarized interstate disputes (mids) dataset version 3.0: Logic, characteristics, and comparisons to alternative datasets. *Journal of Conflict Resolution*.

Marshall, M. G., & Jaggers, K. (2003). *Polity IV project: Political regime characteristics and transitions, 1800–2003.*

Martinussen, T., & Scheike, T. H. (2006). *Dynamic Regression Models for Survival Data*. Statistics for Biology and Health. Springer, New York.

Meyn, S. P., & Tweedie, R. L. (1993). *Markov Chains and Stochastic Stability*. Communications and Control Engineering Series.

Pinker, S. (2011). *The Better Angels of Our Nature: Why violence has declined*. Viking books, Toronto.

Poast, P. (2016). Dyads are dead, long live dyads! the limits of dyadic designs in international relations research. *International Studies Quarterly*, *60*(2), 369–374.

Przeworski, A., Alvarez, M. E., Cheibub, J. A., & Limongi, F. (2000). *Democracy and Development, Political Institutions and Well-Being in the World, 1950-1990.* Cambridge University Press, Cambridge.

R Development Core Team. (2008). R: A language and environment for statistical computing.

Ross, S. M. (2014). *Introduction to Probability Models* (Eleventh). Elsevier/Academic Press, Amsterdam.

Russet, B., & Oneal, J. (2001). *Triangulating Peace: Democracy, Interdependence, and International Organizations.* Norton, New York.

Sarymsakov, T. A. (1953). On the ergodic principle for nonstationary Markov chains. *Doklady Akad. Nauk SSSR (N.S.) 90*, 25–28.

Seneta, E. (2014). Inhomogeneous Markov chains and ergodicity coefficients: John Hajnal (1924–2008). *Comm. Statist. Theory Methods*, *43*(7), 1296–1308.

Sethuraman, S., & Varadhan, S. R. S. (2005). A martingale proof of Dobrushin's theorem for non-homogeneous Markov chains. *Electron. J. Probab. 10*, no. 36, 1221–1235.

Singer, J. D., Bremer, S., & Stuckey, J. (1972). Capability distribution, uncertainty, and major power war, 1820-1965. In B. Russet (Ed.), *Peace, war and numbers.* Sage, Beverly Hills.

Singer, J. (1987). Reconstructing the correlates of war dataset on material capabilities of states, 1816-1985. *International Interactions*, *14*, 115–32.

Stinnett, D. M., Tir, J., Schafer, P., Diehl, P., & Gochman, C. (2002). The correlates of war project direct contiguity data, version 3. *Conflict Management and Peace Science*, *19*(2).

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*(1), 1–25.

Zhang, X., & Liang, H. (2011). Focused information criterion and model averaging for generalized additive partial linear models. *Ann. Statist. 39*(1), 174–200.