Modeling Heterogeneity and Serial Correlation in Binary Time-Series Cross-sectional Data: A Bayesian Multilevel Model with AR(p) Errors

Xun Pang

Department of Politics, Princeton University, 035 Corwin Hall, Princeton, NJ 08544
e-mail: xpang@princeton.edu

This paper proposes a Bayesian generalized linear multilevel model with a *p*th-order autoregressive error process to analyze unbalanced binary time-series cross-sectional (TSCS) data. The model specification is motivated by the generic TSCS data structure and is intended to handle the associated inefficiency and endogeneity problems. It accommodates heterogeneity across units and between time periods in the form of random intercepts and random-effect coefficients. At the same time, its *p*th-order autoregressive error process, employed either by itself or in concert with other dynamic methods, adequately corrects serial correlation and improves statistical inference and forecasting. With a stationarity restriction on the error process, the model can also be used as a residual-based cointegration test on discrete TSCS data. This is especially valuable because cointegration testing on discrete TSCS data is methodologically challenging and rarely conducted in practice. To handle the estimation difficulties, I developed an efficient Markov chain Monte Carlo (MCMC) algorithm by orthogonalizing the error term with the Cholesky decomposition and adding an auxiliary variable. The parameter expansion method, that is, partial group move–multigrid Monte Carlo updating (PGM-MGMC), is employed to further improve MCMC mixing and speed up convergence. The paper also provides a computational scheme to approximate the Bayes's factor for the purposes of serial correlation diagnostics, lag order determination, and variable selection. Simulated and empirical examples are used to assess the model and techniques.

1 Introduction

The popularity and importance of time-series cross-sectional (TSCS) data in political science calls for statistical models that are able to handle heterogeneity and serial correlation and uncover the nonrandom pattern in the errors for statistical inference and forecasting. Conventional models and methods—such as the fixed- and random-effect estimators (FE and RE) and the panel-corrected standard errors (PCSE)—often treat heterogeneity and serial correlation only as "problems" to correct or control for. In this way, they waste information about political dynamics and differences among units and between

time periods. This is undesirable because heterogeneity and dynamics are of substantive importance to political scientists (Beck and Katz 1996, 2009; Beck, Katz, and Tucker 1998).

Path dependence or political inertia is a salient political phenomenon and can help explain the evolution of political institutions and events (Thelen 1999; Pierson and Skocpol 2002; Peters, Pierre, and King 2005). Adequately modeling serial correlation not only improves reliability of statistical inference but also reveals *sub rosa* political dynamic processes. Likewise, contemporary correlation caused by time-specific characteristics (common shocks from the system) is substantively interesting in studies of international and comparative political economy (IPE and CPE) because globalization has a major impact on almost all significant political economic phenomena (Garrett 1998; Woods 2001; Rudra 2002; Franzse and Hays 2007). Heterogeneity across units is also useful for cross-unit comparison. This is especially meaningful when the data contain a small or moderate number of theoretically interesting units, such as countries, states, or legislators. Thoroughly analyzing heterogeneity is also crucial for correctly understanding dynamics. Ignoring heterogeneity leads to ''spurious dynamics'' wherein temporal pseudodependence is simply caused by unmodeled differences instead of dynamics (Heckman 1981). Therefore, TSCS data analysis has two equally important tasks: overcoming the methodological challenges raised by the correlated TSCS structure and taking advantage of the unique *opportunities* presented by TSCS data for studying substantively interesting relationships. In the political methodology literature on TSCS analysis, most efforts are put into the former task, but the latter one is rather under addressed.

Methodologically, most existing methods are developed for linear models. The complexities caused by nonlinearity have rarely been addressed (for exceptions, see Beck, Katz, and Tucker 1998; Beck et al. 2002). In the broader literature on longitudinal data analysis, little effort has been made to thoroughly model intertemporal and contemporaneous dependencies in the nonlinear framework (Schafer and Yucel 2002; Renard, Molenberghs, and Geys 2004; Molenberghs and Verbeke 2005). However, TSCS data in political science often have discrete outcomes (most often, dichotomous outcomes): for example, democratization, exchange rate regime, international conflict, civil war, or state failure. Compared to continuous outcomes, discrete (especially dichotomous) data make it even more challenging to handle the already quite complex TSCS data structure.

This article proposes a Bayesian generalized linear multilevel model with a *p*th-order autoregressive error process (henceforth, GLMM-AR(p)) to model unbalanced binary TSCS data. In contrast to conventional methods that treat heterogeneity as an estimation nuisance, this Bayesian model analyzes heterogeneity with random-effect parameters, and their estimates (posterior distributions) can be easily used to learn the degree of heterogeneity in the time and spatial dimensions.[1] In this multilevel model, covariates are included at their appropriate levels to explain the individual- or group-level variation. The autoregressive error process is used to test and correct serial correlation. It also serves as a residual-based TSCS cointegration test to help avoid spurious regressions. To overcome the estimation difficulties caused by nonlinearity and high dimensionality, I use several methods to develop an efficient Markov chain Monte Carlo (MCMC) algorithm. I also propose a computational scheme for model comparison based on the Bayes's factor (BF).

---

[1]In this paper, I only consider contemporaneous correlation caused by time-specific shocks from the system. Contemporaneous correlation can also result from interactions among the units, which is often analyzed by applying spatial dynamic regressions (Franzse and Hays 2007, 2008a, 2008b). Integrating spatial dynamics is a future extension of the model presented in this paper.

The model and methods are applied to simulated and empirical binary TSCS data to improve the reliability of statistical inference and forecasting.

2   Motivating Examples and TSCS Data Structure

The GLMM-AR(p) model proposed in this paper is based on the generic characteristics of the TSCS data structure. Before introducing the model specification and going through the statistical details, it is necessary to first discuss the structural characteristics of TSCS data and the associated methodological issues. To do so, I use two typical TSCS data sets in political science as motivating examples.

The first application regards state failure in the sub-Saharan region. A large database on state failure was released by the Political Instability Task Force (PITF) in 2001 and has been analyzed in many studies (King and Zeng 2001b; Beck et al. 2002; Goldstone et al. 2000, for instance). The sub-Saharan data set contains 40 countries repeatedly measured from 1956 to 1995. The response variable, state failure, is dichotomous—a new occurrence of state failure in a country-year is coded 1; otherwise, the outcome is coded 0. Because of missing outcome data, the data structure is rather unbalanced. The minimum number of country observations is only 3 (Eritrea and Ethiopia), and the maximum is 40 (Liberia, South Africa, and Swaziland). In the cross-sectional dimension, only three countries are observed in each of the 4 years from 1956 to 1959 compared to 40 countries in 1993 and 1994. One of the interesting but debated substantive questions in the IPE and CPE literature is whether trade openness prevents state failure (King and Zeng 2001b; Beck et al. 2002; Goldstone et al. 2000).

The second example is also about state failure in the sub-Saharan countries but focused on civil war, one specific type of state failure. The response variable, civil war duration, is coded 1 if a country is in a new or ongoing civil war; otherwise, it is coded 0. The data are from Miguel, Satyanath, and Sergenti (2004) and analyze how economic growth affects the duration of civil war. The sample countries are observed between 1980 and 1999. The minimum number of country observations is 9 (Namibia) and year observations is 36 (1999), whereas the maximum number of country observations is 20 and year observations is 40.

The basic characteristics of the two TSCS data sets include multilevel and dynamic features and an unbalanced data structure. The data are multilevel in the sense that information is gathered both at the individual (country-year) and at the group (country and year) levels. Some variables of theoretical importance vary by time and unit, such as trade openness, economic growth, and democracy, but others are group-level characteristics, invariable across time or units, such as colonial heritage, geography, whether the state is an oil-exporting country, and the Cold War. Because of the multilevel feature, TSCS data are clustered in the time and spatial dimensions without nesting. Observations of the same year or of the same country can be regarded as members of the same group. Observations in the same group are correlated because they share the group-specific features (Gelman and Hill 2006; Beck and Katz 2007; Gill 2007; Shor et al. 2007). Furthermore, the effect of a covariate can vary by country or year, implying varying coefficients across groups (heterogeneous effects). To handle observed, unobserved, and coefficient heterogeneity, we can apply standard multilevel modeling in the classical or Bayesian frameworks (Schafer and Yucel 2002; Renard, Molenberghs, and Geys 2004; Molenberghs and Verbeke 2005, Chap. 14 and Chap. 22). Such models can be estimated with well-developed software, such as the lme4 R package, JAGS/BUGS, the SAS program, and so forth.

The most challenging part of discrete TSCS analysis lies in how to appropriately handle their time-series features in the multilevel and nonlinear situation. Both the state failure and

the civil war duration data can be regarded as a collection of multiple time series. There is always one time series for each unit in TSCS data, though the lengths of the time series may vary widely; therefore, all the concerns in time-series analysis are also relevant to TSCS data. Beck and Katz (1995), Beck, Katz, and Tucker (1998), and Beck et al. (2002) provide excellent guidelines for political scientists about how to handle serial correlation and how to use information in TSCS data to reveal political dynamics. However, several important questions have not yet been fully addressed. First, although we can use a variety of methods to capture dynamics in the data, most of them do not speak to whether the errors are still serially correlated. There are multiple sources of intertemporal correlation, including un-modeled dynamics, omitted variables, or measurement error. Dynamics in the errors may not be fully captured by lags or time splines. Serial correlation is the nonrandom temporal pattern that is not explained in a model. Uncovering this pattern is necessary to better un-derstand dynamics and reach more reliable inferences. Second, in the motivating examples as well as many other empirical studies, the response variable is discrete. Methods for serial correlation in linear models do not necessarily apply to nonlinear models. For example, a linear model with a lagged response has the same error structure as one with an AR(1) error process, but this is not the case in generalized linear models. Third, some of the most often-employed methods should be used with caution. Including lagged values in a struc-tural model should have a theoretical justification. If the errors are still correlated, lagged responses invite endogeneity (Achen 2001; Wilson and Butler 2007; Skrondal and Rabe-Hesketh 2008). Last but not least, cointegration testing is necessary to avoid spurious regressions in TSCS analysis (Hamilton 1994; Kao 1999; Choi 2001; Hubrich, Luetkepohl, and Saikkonen 2001; Im, Pesaran, and Shin 2003; Philips and Sul 2003; Bai and Ng 2004; Sul 2009).[2] In the state failure example, it is possible that either the mean level or the variance of the underlying failure propensity changes over time. Noncointegration is more likely in the civil war duration study. As Fearon and Laitin (2003) point out, civil wars tend to be cumulative and last for multiple periods. In the data set, a civil war often lasts for several years, sometimes more than a decade. This demonstrates strong path dependence. The underlying propensity for a country to stay in a war may drift widely, but the observed dichotomous response variable imposes formidable challenges to unit root and cointegra-tion testing. One possible solution is to directly model the error dynamic process to see whether it is stationary, which is actually a residual-based cointegration test on TSCS data (Kao 1999; Hubrich, Luetkepohl, and Saikkonen 2001; Im, Pesaran, and Shin 2003).

In general, binary TSCS data—such as the state failure and civil war duration examples—require a dynamic generalized linear multilevel model to handle heterogeneity and serial correlation. The model should analyze the direct and indirect effects of group-level characteristics and estimate the dynamic error process. The model also needs to ac-commodate unbalanced data structures, given that missing outcome data in the state failure and civil war duration example are a norm rather than an exception in political TSCS anal-ysis. However, developing such a model faces at least three major methodological chal-lenges. With heterogeneity in two dimensions and serial correlated errors, the overall error structure consists of three parts, namely, the unit-level errors, the time-level errors, and the serially correlated individual-level errors. The model should be able to analyze all three parts at the same time. Furthermore, because the covariance matrix of the individual-level

---

[2]In time-series analysis, cointegration refers to the situation that a set of variables are individually integrated of order 1, but some linear combination of these variables results is covariance stationary. In the regression sense, if the covariates are cointegrated with the response variable, the residual process is stationary.

errors is not diagonal, the likelihood function is intractable. This makes conventional data augmentation methods inefficient. In addition, the complication arising from correlated errors is further exacerbated by unbalanced data structures that increase the complexity of the covariance matrix. The following three sections will introduce such a model and related estimation methods to handle those computational problems.

3 Model Specification and Assumptions

In this section, I introduce the GLMM-AR(p) model and the necessary assumptions for identification and consistency. I give a general model specification that is designed to accommodate a variety of real data analysis situations. For a particular data set, some parts of the general specification may not be employed. For instance, if no group-level predictors are available, the group-level regressions can be omitted, and the random-effect coefficients are simply modeled as varying around their overall mean. Likewise, if the researcher has substantive theories suggesting that coefficients may not vary by time or unit, the model can be simplified to only have random intercepts. It is also possible that the final specification chosen by model comparison does not have a dynamic error process; nevertheless, it is still necessary to first try different lag orders to test for serial correlation. When developing the model, it is important to leave all important options open and let the model be general and flexible. In empirical studies, we should apply the model based on the particular data set and use the BF to choose a model *ex post* (refer to the next section for details). An R package, GLMMarp, is available online to help specify, estimate, and compare models in the general GLMM-AR(p) framework.

3.1 *General Model Specification: GLMM-AR(p) Model*

Suppose that a data set consists of N unique units indexed i, where $i = 1, 2, \ldots, N$. Each unit i has T_i observations at times $1_i, \ldots, t_i, \ldots, T_i$, where $\{1_i, \ldots, t_i, \ldots, T_i\} \subseteq \{1, 2, \ldots, T\}$. With an unbalanced data structure, it is likely that $T_i \neq T$ and $T_i \neq T_j$ for $i \neq j$. Units i and j are observed in the same time period if and only if $t_i = t_j$. Therefore, t_i indicates the location of observation y_{i,t_i} in the sequence $\{1, 2, \ldots, T\}$, that is, the time-dimensional cluster it belongs to. In the model, there are two sources of clustering: each observation y_{i,t_i} belongs to clusters i and t' ($t_i = t'$) at the same time. By using the latent variable specification (Albert and Chib 1993), the GLMM-AR(p) with a probit link can be written as follows:

$$y_{i,t_i} = \mathbb{I}\left(z_{i,t_i} > 0\right), \tag{1}$$

$$z_{i,t_i} = \mathbf{x}_{1i,t_i}^{'}\boldsymbol{\beta}_1 + \mathbf{w}_{i,t_i}^{'}\boldsymbol{\beta}_{2,i} + \mathbf{s}_{i,t_i}^{'}\boldsymbol{\beta}_{3,t_i} + \xi_{i,t_i}, \tag{2}$$

$$\boldsymbol{\beta}_{2,i} = \mathbf{A}_i\boldsymbol{\beta}_2 + \mathbf{b}_i, \tag{3}$$

$$\boldsymbol{\beta}_{3,t_i} = \mathbf{F}_{t_i}\boldsymbol{\beta}_3 + \mathbf{c}_{t_i}, \tag{4}$$

$$\xi_{i,t_i} = \rho_1\xi_{i,t_i-1} + \cdots + \rho_p\xi_{i,t_i-p} + \boldsymbol{\epsilon}_{i,t_i}, \tag{5}$$

where $\mathbb{I}(\cdot)$ is the indicator function. This specification is an extension of commonly applied mixed-effect models in longitudinal data analysis. In this model, whether y_{i,t_i} takes value 0 or 1 is determined by a latent continuous variable z_{i,t_i} and a threshold 0. At the latent level, z_{i,t_i} is assumed to have a linear relationship with the covariates in equation (2). The error

term ξ_{i,t_i} follows an AR(p) process in equation (5). Although non-Toeplitz errors can be applied, the simple AR(p) error specification is adequate to approximate a white noise process ϵ with an appropriate order p.[3] Besides its computational convenience, the AR(p) error specification has several other advantages. It can serve as a serial correlation diagnostic following the classical Box–Jenkins procedure (Box, Jenskins, and Reinsel, 1994). It is also a residual-based cointegration test because by definition the autoregressive error process should be stationary under cointegration. Note that this error specification does not theoretically or mathematically exclude the use of other methods for dynamics modeling, such as lagged responses or lagged explanatory variables. If necessary, lags can be included as ordinary regressors in the design matrices in equation (2). As this does not lead to any additional methodological complexity, I do not specify them as separate terms in the model and will not discuss them as special specifications. This general specification allows three groups of covariates: those in $\mathbf{x}_{1i,t_i}$ with FEs on all the observations and those in $\mathbf{w}_{i,t_i}$ or $\mathbf{s}_{i,t_i}$ with unit- or time-specific effects. For identification, when the model has both time- and unit-specific intercepts, one of them has to be centered at 0; and if there is also an overall intercept, both random intercepts are mean 0. At the unit level in equation (3), heterogeneity across units is modeled, and the variation of random effects $\boldsymbol{\beta}_{2,i}$ is further explained by a covariate matrix $\mathbf{A}_i$, a parameter vector $\boldsymbol{\beta}_2$, and a group-level error term $\mathbf{b}_i$. The same interpretation applies to the time-level regression in equation (4).

To estimate the model, I construct its reduced form by denoting $\mathbf{x}_{i,t_i} = \left(\mathbf{x}_{1,it_i}, \mathbf{w}'_{i,t_i}\mathbf{A}_i, \mathbf{s}'_{i,t_i}\mathbf{F}_{t_i}\right)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3,)$, which can be written as follows:

$$z_{i,t_i} = \mathbf{x}'_{i,t_i}\boldsymbol{\beta} + \mathbf{w}'_{i,t_i}\mathbf{b}_i + \mathbf{s}'_{i,t_i}\mathbf{c}_{t_i} + \xi_{i,t_i}, \tag{6}$$

$$\xi_{i,t_i} = \rho_1\xi_{i,t_i-1} + \ldots + \rho_p\xi_{i,t_i-p} + \epsilon_{i,t_i}. \tag{7}$$

Here, the "interactions" $\mathbf{w}'_{i,t_i}\mathbf{A}_i$ and $\mathbf{s}'_{i,t_i}\mathbf{F}_{t_i}$ do not cause concern because model interpretation is solely based on the structural form in equations (1)–(5). The estimation form clearly demonstrates the complexity of the error structure. If we assume that $\{\mathbf{b}_i\}$ and $\{\mathbf{c}_t\}$ are uncorrelated, the error structure can be expressed as follows: $\boldsymbol{\Sigma}_{z_i} = \mathbf{W}'_i\boldsymbol{\Sigma}_{b_i}\mathbf{W}_i + T_i\mathbf{S}'_i\boldsymbol{\Sigma}_{c_{T_i}}\mathbf{S}_i + \boldsymbol{\Sigma}_{\xi_i}$, where $\boldsymbol{\Sigma}_{z_i}$, $\boldsymbol{\Sigma}_{b_i}$, $\boldsymbol{\Sigma}_{c_{T_i}}$, and $\boldsymbol{\Sigma}_{\xi_i}$ are the covariance matrices of $\mathbf{z}_i$, $\{\mathbf{b}_i\}$, $\{\mathbf{c}_{T_i}\}$, and ξ_i, respectively.[4] It is important to note that the covariance matrix $\boldsymbol{\Sigma}_{\mathbf{z}i}$ reflects both correlation and heteroskedasticity.

Several assumptions are required for identification and estimation.

1. The error term ϵ_{i,t_i} is white noise—that is, $\epsilon_{i,t_i} \sim \text{Normal}(0, 1)$.
2. There is no cross-sectional correlation in the individual-level error term ξ; mathematically, that means $\text{cor}(\xi_{i,t_i}, \xi_{k,t_k}) = 0$, $\forall i \neq k$.
3. The covariates in $\mathbf{x}_{i,t_i}$, $\mathbf{w}_{i,t_i}$, and $\mathbf{s}_{i,t_i}$ are sequentially exogenous. Formally, this assumption is $(\mathbf{x}_{i,t_i}, \mathbf{w}_{i,t_i}, \mathbf{s}_{i,t_i} \perp \xi_{i,t_i})|z_{i,t_{s_i}}, \quad \forall t_{s_i} \leqslant t_i$.
4. The error term ξ follows a pth-order autoregressive process.
5. The autoregressive process is stationary. The stationarity assumption is restrictive and requires that the dynamic process $\mathbf{z}_i$ is either stationary or cointegrated with

---

[3]A Toeplitz matrix is also known as a diagonal-constant matrix, which is a matrix with constant descending diagonals from left to right. The AR(p) and MA(q) covariance matrices are both Toeplitz.

[4]For model specification and estimation, this assumption is not necessary. Here, the assumption is used only for an easily written mathematical form as an illustration how complex the error structure is.

the explanatory variables (Chib and Greenberg 1994); otherwise, the model cannot be applied.[5] This assumption makes the GLMM-AR(p) a cointegration test following the line of residual-based testing (Engle and Granger 1987; Kao 1999; Pedroni 1999, 2004). Later in this paper, I use an empirical example to illustrate how the GLMM-AR(p) model serves as a tool to detect spurious regressions.

6. Finally, I specify prior distributions of the parameters as follows:

$$\boldsymbol{\beta} \sim N_{K_1}(\boldsymbol{\beta}_0, \mathbf{B}_0), \qquad \epsilon_{i,t_i} \sim N(0,1), \qquad \{\mathbf{b}_i\} \sim N_{K_2}(\mathbf{0}, \mathbf{D}), \qquad \mathbf{D}^{-1} \sim W_{K_2}(\boldsymbol{\upsilon}_0, \mathbf{D}_0),$$
$$\{\mathbf{c}_t\} \sim N_{K_3}(\mathbf{0}, \mathbf{E}), \qquad \mathbf{E}^{-1} \sim W_{K_3}(\boldsymbol{\eta}_0, \mathbf{E}_0), \qquad \boldsymbol{\rho} \sim U_p(\boldsymbol{\rho} : \boldsymbol{\rho} \in S_\rho),$$

where N is the normal density, W is the Wishart density, U is a uniform density, and S_ρ is the stationarity space of the p-th order autoregressive process.[6] Because ϵ_{i,t_i}, $\{\mathbf{b}_i\}$, and $\{\mathbf{c}_t\}$ are error terms, their prior means are $\mathbf{0}$. With these prior functional forms, all the parameters except $\boldsymbol{\rho}$ have conditional conjugacy. Other prior specifications are possible. For instance, the autoregressive coefficients $\boldsymbol{\rho}$ could have a multivariate normal prior distribution truncated in the stationarity space.

4 MCMC Algorithm

Because the error term in this model is serially correlated, the covariance matrix Σ_ξ has nonzero off-diagonal elements, making it challenging to estimate. One possible ''solution'' is to use so-called robust standard errors. However, this method is not as convenient as it seems to be because when constructing the weight function, we have to overcome three types of bias (Andrews 1991; Lumley and Heagerty 1999; Zeileis 2004).[7] More importantly, using robust standard errors discard valuable information likely contained in the error term. Such information could have been used for analyzing dynamics and heterogeneity and for improving forecasting. Other solutions in the literature include numerical methods, such as penalized quasilikelihood, marginal quasilikelihood, or expectation-maximization algorithms (Schafer and Yucel 2002; Renard, Molenberghs, and Geys 2004; Molenberghs and Verbeke 2005, Chap. 14 and Chap. 22). Those methods are burdened with many approximation steps such as linear or high-order Taylor expansion, Gauss–Hermite quadrature, pseudodata generation, and empirical Bayes's estimation. What is even worse is that these procedures often produce estimates and standard errors that are biased toward zero, especially with low-order expansions (Ng et al. 2006). The Bayesian approach has been widely used for analyzing multilevel models because it has the flexibility needed for the estimation of parameter-rich models (Hagenaars 1990; Singer and Willett 2003; Skrondal and Rabe-Hesketh, 2004; Yang, Fu, and Land 2004; Molenberghs

---

[5]With the stationarity restriction, autoregressive coefficients are not allowed to go out of the stationarity space in an MCMC simulation. If the error process is actually not stationary, an MCMC simulation will be abnormally halted because no legitimate sample of autoregressive coefficients can be drawn, or the MCMC draws of those parameters are highly concentrated around the edge of the stationarity space, causing the Markov chain impossible to pass convergence tests. In either case, the model cannot be applied because we should not draw any inference based on the simulation output.

[6]An autoregressive process is stationary if all the characteristic roots of the polynomial are outside the unit circle. For different autoregressive processes, the stationarity space is different.

[7]The three biases are the bias of the estimator of the variance, the bias due to omitted and down-weighted correlations (truncation bias), and the bias caused by evaluating the estimator at estimates rather than the true parameters (centering bias).

and Verbeke, 2005). In this section, I discuss these estimation problems and provide Bayesian solutions.

4.1 *Cholesky Decomposition and Auxiliary Variable Approach*

When developing an MCMC algorithm for model estimation, the major problem posed by the nondiagonal matrix Σ_{ξ_i} is the construction of the conditional distribution of $\{\mathbf{c}_t\}$. To sample $\{\mathbf{c}_t\}$, we have to first switch from the $\mathbf{z}_i$-dimension to the $\mathbf{z}_t$-dimension and then determine the joint error distribution in the $\mathbf{z}_t$-dimensional expression of the GLMM-AR(p) model:

$$\begin{cases} \mathbf{z}_1 = \mathbf{x}_1\boldsymbol{\beta}+\mathbf{w}_1\mathbf{b}_{N_1}+\mathbf{s}_1\mathbf{c}_1+\boldsymbol{\epsilon}_1, \\ \mathbf{z}_2 = \mathbf{x}_2\boldsymbol{\beta}+\mathbf{w}_2\mathbf{b}_{N_2}+\mathbf{s}_2\mathbf{c}_2+\boldsymbol{\epsilon}_2, \\ \vdots \\ \mathbf{z}_T = \mathbf{x}_T\boldsymbol{\beta}+\mathbf{w}_T\mathbf{b}_{N_T}+\mathbf{s}_T\mathbf{c}_T+\boldsymbol{\epsilon}_T. \end{cases} \quad (8)$$

This is a seemingly unrelated regression (SUR) system. The errors in the T equations are correlated because of serial dependence. The covariance matrix is $\mathrm{Var}(\boldsymbol{\epsilon}|\mathbf{X},\mathbf{w},\mathbf{s}) = \Sigma_{\epsilon_t}\otimes\mathbf{I}_T$ for a balanced data set and more complicated if $N_t \neq N_k$ for some $t \neq k$. Handling unbalanced structures is not easy in SUR analysis. In the Bayesian framework, it is feasible to calculate the covariance matrix $\mathrm{Var}(\epsilon|\mathrm{X},\mathbf{w},\mathbf{s})$ with unbalanced structures. However, it is computationally inefficient to calculate an $M \times M$ matrix $\left(M = \sum_{t=1}^{T}N_t\right)$ in each MCMC iteration by decomposing Toeplitz matrices and then mapping the elements into a large covariance matrix.

Another closely related problem is that the nondiagonal covariance matrix means that we have to sample from *truncated multivariate* normal distributions to augment $\{\mathbf{z}_i\}$. Most existing methods for doing this are inefficient and mix slowly (Geweke 1991, 1996; Sandor and Andras 2004; Rodriguez-Yam, Davis, and Scharf 2004). Correlated errors also make it difficult to compute the BF for model comparison. This is because to calculate the BF, we have to deal with the following integral in the likelihood computation:

$$L = f\left(y_{i,1_i},\ldots,y_{i,T_i}|\boldsymbol{\theta}\right) = \int_{a_{1_i}}^{b_{1_i}}\int_{a_{2_i}}^{b_{2_i}}\ldots\int_{a_{i,T_i}}^{b_{i,T_i}} p\left(z_{i,t_{1_i}},\ldots,z_{i,T_i}|\boldsymbol{\theta}\right)\mathrm{d}z_{i,t_{1_i}}\ldots\mathrm{d}z_{i,T_i}, \quad (9)$$

where $\left(a_{i,t_i},b_{i,t_i}\right)$ is the truncated region determined by the value of y_{i,t_i}, and $\boldsymbol{\theta}$ represents all the parameters in the model. This high-dimensional integration is difficult because $\mathbf{y}_i$ are not i.i.d. conditional on $\boldsymbol{\theta}$—that is, $f\left(y_{i,1_i},\ldots,y_{i,T_i}|\boldsymbol{\theta}\right) \neq \prod_{t=1}^{T_i}f\left(y_{i,t}|\boldsymbol{\theta}\right)$. To do numerical integration, we need random samples of $\mathbf{z}$ from its marginal distribution (unconditional on $\mathbf{y}$). This creates a paradox because the latent process $\mathbf{z}$ can only be inferred from the observed process $\mathbf{y}$. One solution is the method of Geweke, Hajivassiliou, and Keane (the GHK simulator) (Geweke 1991; Borsch-Supan and Hajivassiliou 1993; Keane 1994; Chib and Jeliazkov 2006), which is computationally expensive, especially with a large data set. Alternatively, if the latent response variable is updated by using a polynomial operator $P(L)z_{i,t_i}$, the likelihood can be approximated with the auxiliary particle filter (Mueller and Czado 2005; Pang 2008). However, this sampling importance resampling scheme is numerically unstable for a high-order Markov process and often requires numerous draws to obtain a valid approximation of the likelihood.

In this paper, I propose an algorithm which orthogonalizes the correlated errors in such a way that $\mathbf{z}$ can be sampled independently. With this method, $f\left(y_{i,1_i},\ldots,y_{i,T_i}|\boldsymbol{\theta}'\right) = \prod_{t=1_i}^{T_i}f\left(y_{i,t}|\boldsymbol{\theta}'\right)$, where $\boldsymbol{\theta}' = (\boldsymbol{\theta},\mathbf{u})$ and $\mathbf{u}$ is an auxiliary parameter vector. The basic idea

is as follows: first, I decompose the covariance matrix $\boldsymbol{\Sigma}_{\xi_i}$ into two parts $\Sigma_{\boldsymbol{\xi}} = \boldsymbol{\Omega}_i + \kappa_i \mathbf{I}_{T_i}$, where $\boldsymbol{\Omega}_i$ is a positive-definite matrix and $\boldsymbol{\kappa}_i$ is an arbitrary constant; and $\boldsymbol{\Omega}i$ is further decomposed into $\mathbf{V}_i^{'}\mathbf{V}_i$ in which $\mathbf{V}_i^{'}$ is a lower triangular matrix produced in the Cholesky decomposition.[8] Hence, $\Sigma_{\boldsymbol{\xi}_i} = \mathbf{V}_i^{'}\mathbf{V}_i + \kappa_i \mathbf{I}_{T_i}$. Now equation (6) can be reexpressed as follows:

$$\mathbf{z}_i = \mathbf{x}_i^{'}\boldsymbol{\beta} + \mathbf{w}_i^{'}\mathbf{b}_i + \mathbf{s}_i^{'}\mathbf{c}_{T_i} + \mathbf{V}_i^{'}\mathbf{u}_i + \boldsymbol{\varepsilon}_i, \tag{10}$$

where the new error term $\boldsymbol{\epsilon}_i \sim N_{T_i}(0, \kappa_i \mathbf{I}_{T_i})$, the auxiliary variable $\mathbf{u}_i \sim N_{T_i}(\mathbf{0}, \mathbf{I}_{T_i})$, and $\boldsymbol{\epsilon}_i$ and $\mathbf{u}_i$ are mutually independent. In equation (10), $\mathbf{z}_i$ has exactly the same covariance matrix as before $(\boldsymbol{\Sigma}_{\boldsymbol{\xi}_i})$ when integrating $\mathbf{u}_i$ out. Given $\mathbf{u}_i$, $\boldsymbol{\rho}$ and other parameters, the zs do not need to be updated conditional on one another because the new error term ϵi is i.i.d. There are general formulas to compute the Toeplitz covariance matrix $\Sigma_{\boldsymbol{\xi}_i}$ for AR(p) errors. In practice, I do not need to compute each $\Sigma_{\boldsymbol{\xi}_i}$; instead, in each iteration, I can simply compute $\Sigma_{T \times T}$ and construct the covariance matrix $\Sigma_{\boldsymbol{\xi}_i}$ for each i by taking the first T_i rows and columns of $\Sigma_{T \times T}$. With this method, all zs are updated in one block instead of $\sum_{i=1}^{N} T_i$ blocks, improving simulation efficiency. Importantly, because now we have $f(y_{i,1_i}, \ldots, y_{i,T_i}|\boldsymbol{\theta}^{'}) = \prod_{t=1_i}^{T_i} f(y_{i,t}|\boldsymbol{\theta}^{'})$, the likelihood can be computed as simply as in an ordinary probit model.

Define $(q_1, q_2, \ldots, q_{T_i}) \equiv \mathbf{q}_i \equiv \mathbf{V}_i^{'}\mathbf{u}_i$, and the full MCMC algorithm is as follows:

1. $\boldsymbol{\beta}, \{\mathbf{b}_i\}, \{\mathbf{u}_i\}| \cdot \sim \pi(\{\mathbf{u}_i\}|\boldsymbol{\beta}, \{\mathbf{b}_i\}, \cdot)\pi(\{\mathbf{b}_i\}|\boldsymbol{\beta}, \cdot)\pi(\boldsymbol{\beta}| \cdot).[9]

   • $\boldsymbol{\beta}| \cdot \sim N_{K_1}(\bar{\boldsymbol{\beta}}, \mathbf{B}_1)$, where $\mathbf{B}_1 = \left(\mathbf{B}_0^{-1} + \sum_{i=1}^{N} \mathbf{x}_i^{'}\mathbf{H}_i^{-1}\mathbf{x}_i\right)^{-1}$, $\bar{\boldsymbol{\beta}} = \mathbf{B}_1\left(\mathbf{B}_0^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^{N}\mathbf{x}_i^{'}\mathbf{H}_i^{-1}(\mathbf{z}_i - \mathbf{s}_i^{'}\mathbf{c}_{T_i})\right)$, and $\mathbf{H}_i = (\boldsymbol{\Omega}_i + \mathbf{w}_i^{'}\mathbf{D}\mathbf{w}_i)$;

   • $\mathbf{b}_i|\boldsymbol{\beta}, \cdot \sim N_{K_2}(\bar{\mathbf{b}}_i, \mathbf{D}_i)$, where $\mathbf{D}_i = \left(\mathbf{D}^{-1} + \mathbf{w}_i^{'}(\boldsymbol{\Omega}_i)^{-1}\mathbf{w}_i\right)^{-1}$ and $\bar{\mathbf{b}}_i = \mathbf{D}_i\mathbf{w}_i^{'}(\boldsymbol{\Omega}_i)^{-1}(\mathbf{z}_i - \mathbf{x}_i^{'}\boldsymbol{\beta} - \mathbf{s}_i^{'}\mathbf{c}_{T_i})$;

   • $\mathbf{u}_i| \cdot \sim N(\bar{\mathbf{u}}_i, \mathbf{U}_i)$, where $\mathbf{U}_i = (\mathbf{I}_{T_i} + \mathbf{V}_i\mathbf{V}_i^{'}/\kappa_i)^{-1}$, and $\bar{\mathbf{u}}_i = \mathbf{U}_i\mathbf{V}_i(\mathbf{z}_i - \mathbf{x}_i^{'}\boldsymbol{\beta} - \mathbf{w}_i^{'}\mathbf{b}_i - \mathbf{s}_i^{'}\mathbf{c}_{T_i})/\kappa_i$.

2. $\{\mathbf{c}_t\}| \cdot \sim N_{K_3}(\bar{\mathbf{c}}_t, \mathbf{E}_i)$, where $\mathbf{E}_i = \left(\mathbf{E}^{-1} + \mathbf{s}_i^{'}(\boldsymbol{\kappa}_{N_t}\mathbf{I}_N)^{-1}\mathbf{s}_i\right)^{-1}$ and $\bar{\mathbf{c}}_t = \mathbf{E}_i\mathbf{s}_i^{'}(\boldsymbol{\kappa}_{N_t}\mathbf{I}_N)^{-1}$ $(\mathbf{z}_i - \mathbf{x}_i^{'}\boldsymbol{\beta} - \mathbf{s}_i^{'}\mathbf{b}_{N_t} - \mathbf{q}_t)$, where $\boldsymbol{\kappa}_{N_t} = \{\kappa_1, \kappa_j, \kappa_N\}_t$.

3. $z_{i,t_i}| \cdot \sim TN(\mathbf{x}_{i,t_i}^{'}\boldsymbol{\beta} + \mathbf{w}_{i,t_i}^{'}\mathbf{b}_i + \mathbf{s}_{i,t_i}^{'}\mathbf{c}_{t_i} + q_{i,t_i}, \kappa_i)$.

4. $\mathbf{D}^{-1}|\{\mathbf{b}_i\} \sim W_{K_2}(v_1, \mathbf{D}_1)$ and $\mathbf{E}^{-1}|\{\mathbf{c}_t\} \sim W_{K_3}(\eta_1, \mathbf{E}_1)$, where $v_1 = v_0 + N$, $\mathbf{D}_1 = \left(\mathbf{D}_0^{-1} + \sum_{i=1}^{N}\mathbf{b}_i\mathbf{b}_i^{'}\right)^{-1}$, $\eta_1 = \eta_0 + T$, and $\mathbf{E}_1 = \left(\mathbf{E}_0^{-1} + \sum_{t=1}^{T}\mathbf{c}_t\mathbf{c}_t^{'}\right)^{-1}$.

5. $\boldsymbol{\rho}| \cdot \sim \Psi(\boldsymbol{\rho}) \times N(\hat{\boldsymbol{\rho}}, \mathbf{P})$, I use a Metropolis–Hasting algorithm to update $\boldsymbol{\rho}$ with a tailored kernel $N(\hat{\boldsymbol{\rho}}, \mathbf{P})$ as in Chib (1993) .

4.2 *Improving Mixing: PGM-MGMC*

There are two important issues in applying MCMC methods: convergence and mixing. To draw reliable inferences based on MCMC samples, we first have to ensure that the

---

[8]I choose $\varrho_i/2$ for κ_i, where ϱ_i is the smallest eigenvalue of Ω_i. This choice follows Chib and Jeliazkov (2006) and is to make the algorithm numerically stable.

[9]Sampling $\boldsymbol{\beta}$, $\{\mathbf{b}_i\}$, and $\{\mathbf{u}_i\}$ in one block improves the efficiency of this algorithm because they are correlated by construction; however, it is not feasible to include $\{\mathbf{c}_t\}$ in this block because of the complex covariance structure caused by the unbalanced data structure.

simulation output is actually from the target distribution. If the algorithm is correctly developed, the Markov chain is guaranteed to converge eventually to the target distribution. The second major concern is mixing—how fast the chain explores the sample space and produces a representative sample of the posterior distribution. When mixing is slow, the chain stays in a local region for a long time. If we stop the simulation too early, we will reach conclusions based on local draws, which is as misleading as nonconvergence. For models with only a few parameters, mixing is often not a big concern, but for multilevel models with a large number of random-effect and low-level parameters, the Markov chain can mix so slowly that Bayesian models may be impractical (Carlin 1996; Olsen and Schafer 2001). The GLMM-AR(p) model has serially correlated erors and random effects in two dimensions. The algorithm presented above has the slow-mixing problem, as shown in preliminary simulations. This problem is quite serious for the random-effect parameters because their within-chain correlation decreases so slowly that autocorrelation is still above 0.5 even after 200 lags. In those simulation studies, with 500,000 post-burn-in iterations, the posterior samples still fail formal convergence tests using the coda R package. This highlights the necessity of improving the efficiency of the algorithm. The solution used in this paper is to add a PGM-MGMC updating stage into the algorithm. The PGM-MGMC sampler moves the Markov chain in a coordinate-free way in each MCMC iteration. It noticeably reduces within-chain correlation and speeds up MCMC mixing. This is illustrated in the simulated and empirical examples later in the paper.

The basic idea of multigrid methods is to use a sequence of auxiliary "coarse-grid" problems in addition to the original "fine-grid" problem so that the information is more efficiently stored and convergence is accelerated (Briggs 1987, Chap. 3; Goodman and Sokal 1989). This method was first applied in statistical physics and Euclidean quantum physics. Goodman and Sokal (1989) extended the deterministic multigrid method into a MGMC algorithm by applying partial resampling and fiber construction. Liu and Sabatti (2000) generalized the Gibbs sampler by using the MGMC method to decompose the sample space into disjoint orbits to facilitate information transmission.[10] After a transformation group is chosen, the Markov chain is moved by from one orbit to another without leaving the target sample space (Liu and Wu 1999). This enables a faster exploration of the sample space and achieves the effects of reparameterization, blocking, and grouping, but it decomposes the sample space more freely than other methods. Liu and Sabatti (2000) applied this method to state space models and showed that the move could be dramatic and autocorrelation was reduced. Mueller and Czado (2005) also applied the PGM-MGMC method to reduce autocorrelation in their autoregressive ordinal probit model and demonstrated efficiency gains.

The posterior distribution of the GLMM-AR(p) setup facilitates the development of a mover distribution. From this mover distribution, we can randomly draw a mover to transform a subset of m-dimensional parameter vector $\omega \equiv (\{\mathbf{z}_i\}, \boldsymbol{\beta}, \{\mathbf{b}\}, \{\mathbf{c}\})$. To apply the PGM-MGMC method, I choose the scale group $\Gamma = \{\chi > 0 : \chi(x) = \chi \times x\}$ and calculate the unimodular Haar measure as $L(d\chi) = \chi^{-1} d\chi$. Now, it is straightforward to see that the mover, χ, can be sampled from a standard gamma distribution:

---

[10]In group theory, define a group G and a set X, and an *orbit* of $x \in X$ is the set $S \subset X$ to which x can be moved by the elements of G. Disjoint orbits simply means that the orbits $S_1, S_2, \ldots$ are disjoint. Refer to Bogopolski (2008) or Aschbacher (2000) for more formally definitions and detailed theories about orbits.

$$\chi^{m-1}\pi(\chi\boldsymbol{\omega})\mathrm{d}\chi\propto\chi^{m-1}\exp\left(-\tfrac{1}{2}\sum_{i=1}^{N}\left(\chi\mathbf{z}_i-\mathbf{x}_i\chi\boldsymbol{\beta}-\mathbf{w}_i\chi\mathbf{b}_i-\chi\mathbf{c}\right)^{'}\boldsymbol{\Omega}_i^{-1}(\chi\mathbf{z}_i-\mathbf{x}_i\chi\boldsymbol{\beta}-\mathbf{w}_i\chi\mathbf{b}_i-\chi\mathbf{c})\right)$$

$$\times\exp\left(-\tfrac{1}{2}(\chi\boldsymbol{\beta})^{'}\mathbf{B}_0^{-1}(\chi\boldsymbol{\beta})\right)\exp\left(-\tfrac{1}{2}\sum_{i=1}^{N}\chi\mathbf{b}_i^{'}\mathbf{D}^{-1}\chi\mathbf{b}_i\right)\exp\left(-\tfrac{1}{2}\sum_{t-1}^{T}\chi\mathbf{c}_t^{'}\mathbf{E}^{-1}\chi\mathbf{c}_t\right)$$

$$\propto\chi^{m-1}\exp\left(-\tfrac{1}{2}\chi^2\left(\underbrace{\sum_{i=1}^{N}(\mathbf{z}_i-\mathbf{x}_i\boldsymbol{\beta}-\mathbf{w}_i\mathbf{b}_i-\mathbf{s}_i\mathbf{c}_{T_i})^{'}\boldsymbol{\Omega}_i^{-1}(\mathbf{z}_i-\mathbf{x}_i\boldsymbol{\beta}-\mathbf{w}_i\mathbf{b}_i-\mathbf{s}_i\mathbf{c}_{T_i})}_{Q_1}\right.\right.$$

$$\left.\left.+\underbrace{\boldsymbol{\beta}^{'}\mathbf{B}_0^{-1}\boldsymbol{\beta}+\sum_{i=1}^{N}\mathbf{b}_i^{'}\mathbf{D}^{-1}\mathbf{b}_i+\sum_{t=1}^{T}\mathbf{c}_t^{'}\mathbf{E}^{-1}\mathbf{c}_t}_{Q_2}\right)\right)$$

$$\propto\chi^{m-1}\exp\left(-\tfrac{1}{2}\chi^2\left(Q_1+Q_2\right)\right). \tag{11}$$

This kernel is proportional to a gamma distribution $\Gamma(a, b)$ of random variable χ^2 with parameters $a = (m + 1)/2$ and $b = (Q_1 + Q_2)/2$. I apply this PGM-MGMC sampler in each iteration in the MCMC algorithm to further update the parameters in the following way: $\{\chi\mathbf{z}_i^{(g)}\}\to\{\mathbf{z}_i^{(g)}\}$, $\chi\boldsymbol{\beta}^{(g)}\to\boldsymbol{\beta}^{(g)}$, $\{\chi\mathbf{b}_i^{(g)}\}\to\{\mathbf{b}_i^{(g)}\}$, and $\{\chi\mathbf{c}_t^{(g)}\}\to\{\mathbf{c}_t^{(g)}\}$.

5 Bayesian Model Comparison

Because we always have uncertainty in almost all respects of model specification, implementing information-based criteria in model decision making is important (Gill 2007). In the GLMM-AR(p) model, besides the uncertainty in other aspects of model specification, we normally have little prior information about the order of the autoregressive error process. In linear time-series regressions, the way to handle this problem is to first estimate models with different lag orders and then to determine the lag order by model comparison. However, the same procedure is not commonly performed for nonlinear models, and lag orders are often set for convenience sake (most often the first order). The BF is a comprehensive criterion for model comparison, but it is notoriously difficult to compute because of high-dimensional integration and numerical instability (Han and Carlin 2001). With correlated errors, completing this type of computation often requires additional samplers, such as importance samplers (Chib and Jeliazkov 2006) or recursive importance samplers (Mueller and Czado 2005). However, because of the way, I orthogonalize the errors, I can compute the BF in such a simple way that it only requires the full- and reduced-run MCMC outputs.[11] In this section, I present the computational scheme for the BF.

The BF is defined as the ratio of the marginal likelihood of two competing models. The marginal likelihood of a model is the quantity after marginalizing the parameters from the likelihood, that is, $f(y|M_j) = \int f(y|\theta, M_j)\pi(\theta)\mathrm{d}\theta$, where Mj refers to Model j. There are various approaches to approximate the BF (Han and Carlin 2001). Here, I use the marginal likelihood method developed by Chib (1995) and Chib and Jeliazkov (2001). To apply this

---

[11]Full MCMC simulations or full runs estimate all the parameters in the model, whereas reduced MCMC simulations or reduced runs refer to simulations using the same algorithms as the full runs but fixing a subset of parameters at certain values.

method, I first express the marginal likelihood as the normalizing constant in the Bayesian posterior setup: $m(y) = f(y|\theta)\pi(\theta)/\pi(\theta|y)$. Then, I fix θ at θ^*.[12] The marginal likelihood on a logarithmic scale can be computed by using the formula:

$$\ln \hat{m}(y) = \ln \hat{f}(y|\theta^*) + \ln \hat{\pi}(\theta^*) - \ln \hat{\pi}(\theta^*|y). \tag{12}$$

With the Cholesky decomposition and auxiliary parameter approach, the likelihood ordinate, $\hat{f}(y|\theta^*)$, is straightforward to compute. Denote by $\boldsymbol{\theta}$ all the parameters except the auxiliary variable $\mathbf{u}$, the likelihood ordinate can be approximated by integrating out $\mathbf{u}$:

$$\hat{f}(\mathbf{y}|\boldsymbol{\theta}^*) = \frac{1}{M} \sum_{m=1}^{M} \prod_{i=1}^{N} \prod_{t_i=1}^{T_i} (\Delta_{it_i})^{y_{it_i}} (1-\Delta_{it_i})^{1-y_{it_i}}, \tag{13}$$

$$\text{where} \quad \Delta_{it_i} = \Phi\left(\frac{\mathbf{x}_{it_i}'\boldsymbol{\beta}^* + \mathbf{w}_{it_i}'\mathbf{b}_i^* + \mathbf{s}_i'\mathbf{c}_{t_i}^* + q_{it_i}^{(m)}}{\sqrt{\kappa_i^{(m)}}}\right). \tag{14}$$

Because I integrate out $\mathbf{u}$ ($\mathbf{q}$ is a function of $\mathbf{u}$) with respect to the conditional distribution $\pi(\mathbf{u}|\boldsymbol{\theta}^*, \mathbf{z})$, a reduced run is required to recursively sample from $\pi(\mathbf{u}|\mathbf{z}, \boldsymbol{\theta}^*)$ and $\pi(\mathbf{z}|\mathbf{u}, \boldsymbol{\theta}^*)$.

To approximate the posterior ordinate $\pi(\boldsymbol{\theta}^*|y)$, I partition it in the following way:

$$\hat{\pi}(\boldsymbol{\beta}^*, \mathbf{b}^*, \mathbf{D}^*, \boldsymbol{\rho}^*, \mathbf{E}^*|\mathbf{y}) = \hat{\pi}(\boldsymbol{\rho}^*|\mathbf{y})\hat{\pi}(\mathbf{c}^*|, \boldsymbol{\rho}^*, \mathbf{y})\hat{\pi}(\mathbf{E}^*|\mathbf{c}^*, \boldsymbol{\rho}^*, \mathbf{y})\hat{\pi}(\mathbf{b}^*|\mathbf{E}^*, \mathbf{c}^*, \boldsymbol{\rho}^*, \mathbf{y})$$
$$\times \hat{\pi}(\mathbf{D}^*|\mathbf{b}^*, \mathbf{E}^*, \mathbf{c}^*, \boldsymbol{\rho}^*, \mathbf{y})\hat{\pi}(\boldsymbol{\beta}^*|\mathbf{D}^*, \mathbf{b}^*, \mathbf{E}^*, \mathbf{c}^*, \boldsymbol{\rho}^*, \mathbf{y}), \tag{15}$$

and compute each term on the right-hand side:

1. $\hat{\pi}(\boldsymbol{\rho}^*|\mathbf{y})$: denote by ψ all parameters except $\boldsymbol{\rho}$ and $\mathbf{u}$:

$$\hat{\pi}(\boldsymbol{\rho}^*|\mathbf{y}) = \frac{J^{-1}\sum_{i=1}^{N}(\alpha(\boldsymbol{\rho}^{(j)}, \boldsymbol{\rho}^*|\mathbf{y}, \boldsymbol{\psi}^{(j)}, \mathbf{u}^{(j)}, \mathbf{z}^{(j)})q(\boldsymbol{\rho}^{(j)}, \boldsymbol{\rho}^*|\mathbf{y}, \boldsymbol{\psi}^{(j)}, \mathbf{u}^{(j)}, \mathbf{z}^{(j)}))}{K^{-1}\sum_{k=1}^{K}(\alpha(\boldsymbol{\rho}^*, \boldsymbol{\rho}^{(k)}|\mathbf{y}, \boldsymbol{\psi}^{(k)}, \mathbf{u}^{(k)}, \mathbf{z}^{(k)}))}. \tag{16}$$

The numerator is the sample expectation with respect to $\pi(\psi, \mathbf{u}, \mathbf{z}|\mathbf{y})$. The MCMC output can be used to average out these parameters. The denominator is the sample expectation with respect to the conditional product measure $\pi(\psi, \mathbf{u}, \mathbf{z}|\mathbf{y})q(\boldsymbol{\rho}^*, \boldsymbol{\rho}|\mathbf{y}, \psi, \mathbf{u}, \mathbf{z})$. Here, one reduced run is needed by fixing $\boldsymbol{\rho}$ at $\boldsymbol{\rho}^*$ and updating all other parameters.

2. $\hat{\pi}(\mathbf{c}^*|\boldsymbol{\rho}^*, \mathbf{y})$: use the output of the reduced run conducted in Step 1 for this numerical integration.
3. $\hat{\pi}(\mathbf{E}^*|\mathbf{c}^*, \boldsymbol{\rho}^*, \mathbf{z}, \mathbf{y}) = \hat{\pi}(\mathbf{E}^*|\mathbf{c}^*, \mathbf{y})$: no reduced run is required.
4. $\hat{\pi}\left(\mathbf{b}^*|\mathbf{E}^*, \mathbf{c}^*, \boldsymbol{\rho}^*, \mathbf{y}\right) = \prod_{i=1}^{N} \hat{\pi}(\mathbf{b}_i^*|\mathbf{c}^*, \boldsymbol{\rho}^*, \mathbf{z}_i, \mathbf{y}_i)$: conduct a reduced run by fixing $\mathbf{E}$, $\mathbf{c}$, and $\boldsymbol{\rho}$.
5. $\hat{\pi}(\mathbf{D}^*|\mathbf{b}^*, \mathbf{E}^*, \mathbf{c}^*, \boldsymbol{\rho}^*, \mathbf{y}) = \hat{\pi}(\mathbf{D}^*|\mathbf{b}^*, \mathbf{y})$: no reduced run is needed.

---

[12]The parameters $\boldsymbol{\theta}$ can be fixed at any values. But for the sake of numerical stability, values in the high-density region of the posterior are preferred. In this paper, I choose θ^* to be the posterior mean.

6. $\hat{\pi}(\boldsymbol{\beta}^*|\mathbf{D}^*, \mathbf{b}^*, \mathbf{E}^*, \mathbf{c}^*, \boldsymbol{\rho}^*, \mathbf{y}) = \hat{\pi}(\boldsymbol{\beta}^*|\mathbf{b}^*, \mathbf{c}^*, \boldsymbol{\rho}^*, \mathbf{z}, \mathbf{y})$: conduct a reduced run by fixing $\mathbf{D}$, $\mathbf{b}$, $\mathbf{E}$, $\mathbf{c}$, and $\boldsymbol{\rho}$.

6 Simulation Study

I examine the performance of the model and estimation techniques with two Monte Carlo experiments. The only difference between their data-generating processes (DGPs) is the error dynamic process—the true autoregressive coefficients are $\rho_1 = 0.7$ and $\rho_2 = 0.2$ in the first data set and $\rho_1 = -0.5$ and $\rho_2 = -0.3$ in the second one. Chib and Jeliazkov (2006) argue that in hierarchical models, negative autocorrelation is easier to identify. Here, I use the two comparable data sets to check this point. In both data sets, the number of observations is 2187, and the data structures are unbalanced. The number of observations of a unit varies from 2 to 50 and that of a time period varies from 8 to 50. Both DGPs have a mixed-effect design: there are five covariates with FEs ($\mathbf{x}_i$), five with unit-specific effects ($\mathbf{w}_i$), and two with time-specific effects ($\mathbf{s}_i$). At the group levels, there are three unit-level predictors ($\mathbf{A}$) and two time-level predictors ($\mathbf{F}$).

I assign diffuse priors. For the auxiliary parameter vector, $\mathbf{u}_i$, its prior is $N_{T_i}(\mathbf{0}, \mathbf{I}_{T_i})$ by design. The prior choice is also straightforward for the group-level errors $\{\mathbf{b}_i\}$ and $\{\mathbf{c}_t\}$. As residuals, their distributions are centered at $\mathbf{0}$ and their covariance matrices are treated as hyperparameters. The coefficient parameter vector $\boldsymbol{\beta}$ is assigned with a multivariate normal prior centered at $\mathbf{0}$, having a diagonal covariance matrix, $400 \times \mathbf{I}$. This prior is vague and does not have notable effects on the posteriors, given that the sample size is 2187. In order to implement the PGM-MGMC algorithm, the priors of $\boldsymbol{\beta}$ are required to be centered at $\mathbf{0}$. The prior on the autoregressive parameter vector $\boldsymbol{\beta}$ is a multivariate uniform distribution truncated within the stationary space. For the hyperparameters $\mathbf{D}$ and $\mathbf{E}$, their priors can have a discernible influence on the posteriors because binary data often have very little information about such low-level parameters. However, excessively diffuse priors, such as setting $\mathbf{D}_0 = 100 \times \mathbf{I}$, are likely to cause trouble in inverting matrices in the Wishart updating. Cowles, Carlin, and Connett (1996) and Ibrahim and Klainman (1998) also observe that excessively vague priors of the covariance matrices invite numerical instability and slow convergence. My experience from many trials suggests that priors around $\mathbf{D}_0 = 20 \times \mathbf{I}$ (and the same for $\mathbf{E}_0$) are good choices for balancing prior influence and numerical stability for varying sample sizes. The sensitivity of posteriors to the prior choices was checked by using alternative priors with reasonable changes of locations and scales.

I conducted the simulations using two Macintosh machines, each equipped with eight 2.26 GHz Intel Xeon "Nehalem" processors and 6 GB of memory running Mac OS X. With the sample size 2187, the average time for one iteration is about 5.35 s. The MCMC outputs analyzed in this section are based on 100,000 iterations after discarding 10,000 burn-in iterations for each model. Competing models were running in parallel. The running time for each GLMM-AR(p) model is about 1 week. This computing time is acceptable considering that the code is written completely in R. In the Metropolis-Hastings (MH) step, because the proposal density is tailored, the acceptance rate is roughly between 75% and 90%. Multiple convergence diagnostics were conducted for all parameters except the augmented data and the auxiliary parameters $\mathbf{u}$. Because there are too many such parameters (2187×2), I randomly drew 100 of each and conducted diagnostics.

As observed by Carlin (1996) and Olsen and Schafer (2001), the slow MCMC mixing problem for nonlinear mixed-effect models is serious even after the improvement achieved by the Cholesky decomposition and auxiliary parameter approach. As shown in Fig. 1, for the FE parameters, the standard Gibbs chains are mixed well; but for the random-effect and
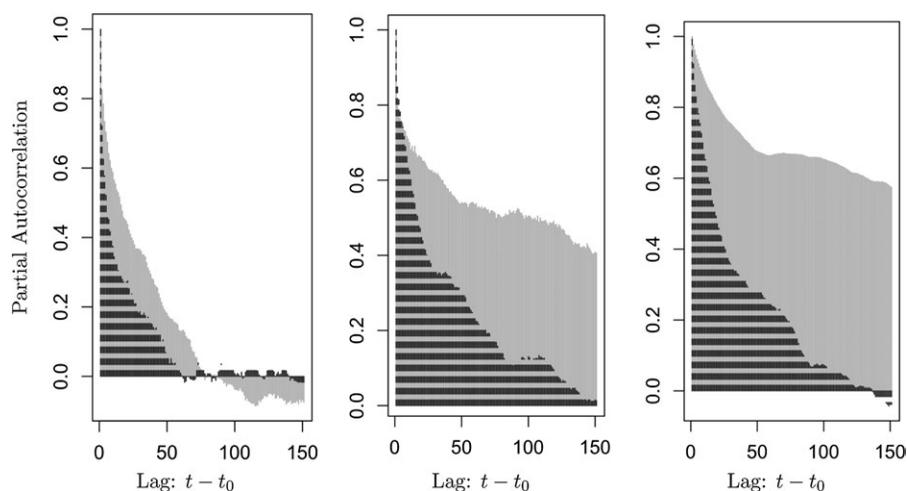
Fig. 1 Comparison of within-chain correlation: Gibbs versus Gibbs + PGM-MGMC.
Note. The gray shadow indicates the autocorrelation of the standard Gibbs chains, whereas the
black shadow shows the decrease of autocorrelation when applying the PGM-MGMC method. From
the left to the right, the graphs illustrate the autocorrelation of the Markov chains of a FE parameter,
a random-effect parameter, and one element in the covariance matrix **E**. These parameters are
randomly chosen.

the low-level parameters, autocorrelation decreases rather slowly—even after 150 lags, it is
still as high as 0.5. By using the PGM-MGMC mover, the mixing is considerably improved,
especially for those slowly mixed parameter chains, as shown by the blue areas in Fig. 1.

To select the lag order p for the autoregressive error process, I start by estimating the
GLMM-AR(0) model with the assumption of serial independence and compute its marginal
likelihood. Then I estimate the GLMM-AR(1) and compare its marginal likelihood with that
of the GLMM-AR(0). If the marginal likelihood of the AR(0) model is larger than that of the
AR(1) model, I choose 0 as the lag order; otherwise, I increase the lag order to estimate the
GLMM-AR(2) and compare it with the GLMM-AR(1). I continue this process until an in-
crease of the lag order leads to a decrease of the marginal likelihood and then I choose the lag
order associated with the largest marginal likelihood. This procedure illustrates how the
model can serve as a serial correlation diagnostic for binary TSCS data. It is much easier
to apply than the score test proposed in Gourieroux, Monfort, and Trognon (1985), espe-
cially for a high-order autocorrelation diagnostics. Notice that the lag order chosen by this
process is the one that fits the data best. It may not be the "true" lag order if the data do not
provide enough information to fully uncover the error process. The posteriors of the autor-
egressive coefficients can tell how strong and how lasting the autocorrelation is. In practice, I
estimated GLMM-AR(p) with the lag orders from 0 to 4 or 5 in parallel and then compared
them to decide whether a further increase is needed and what the final lag order is.

In Table 1, I report the marginal likelihood of the competing models.[13] For DGP I with
positive serial correlation, the GLMM-AR(3) is the best model among the five models with
the lag orders from 0 to 4. Meanwhile, for DGP II with negative serial correlation, the
GLMM-AR(2) fits the data best compared to other competing specifications. Figure 2

---

[13]Because the BF is just a pairwise comparison of marginal likelihoods of two competing models, we can directly
look at marginal likelihoods—the biggest value represents the best model.

Table 1 Log marginal likelihood

DGP	AR(0)	AR(1)	AR(2)	AR(3)	AR(4)
Simulation I ($\rho_1 = 0.7$, $\rho_2 = 0.2$)	-594.33	-582.44	-574.44	-560.85	-569.47
Simulation II ($\rho_1 = -0.5$, $\rho_2 = -0.3$)	-570.91	-551.46	-544.25	-558.36	—

summarizes the posteriors of all FE parameters and part of the mean parameters of random effects.[14] The GLMM-AR(0) assumes independent errors. It produces estimates biased toward 0 and artificially small error bands (high certainty levels) for both DPGs. This is for two reasons: first, with falsely assumed serial independence, estimators of standard errors are biased (Gourieroux, Monfort, and Trognon 1984; Poirier and Ruud 1988); in addition, with the probit link, the coefficient estimates are scaled by Σ_{ξ_i}. If the errors are correlated, the standard identification assumption $\Sigma_{\xi_i} = \mathbf{I}$ leads to smaller estimates. This is because the actual diagonal elements in Σ_{ξ_i} are greater than 1 with autocorrelated errors so that estimates are further scaled toward 0. This kind of bias also applies to the random coefficients at the two group levels, as illustrated in Fig. 3.

In these simulated studies, I did not find clear evidence that negative autocorrelation is easier to identify—their directions are all estimated correctly, although their scales are not. With a larger number of time periods or units, the autoregressive coefficients will be better estimated. In Fig. 2, most posteriors of the models with $p > 0$ cover true values, and they are similar across those models. Although all the true parameters except ρ are exactly the same in the two DGPs, the posteriors are slightly different for the two different data sets, implying that the two data sets are also different in terms of how informative the data are.

The simulation designs put much heterogeneity in both the time and the unit dimensions: the random effects β_{2i} and β_{3t} are generated from distributions $N(\mathbf{A}_i^{'}\beta_2, \mathbf{D})$ and $N(\mathbf{F}_t^{'}\beta_3, \mathbf{E})$ with large variance. Figure 3 shows that the estimated group-level residuals vary widely across units and between time periods. The graphs also demonstrate that an adequate correction for serial correlation is important for better identifying heterogeneity. The posteriors of the group-specific residuals in the GLMM-AR(0) model are more concentrated around zero. This demonstrates that neglecting serial correlation results in the underestimation of heterogeneity.

Other parameters such as the covariance matrices $\mathbf{D}$ and $\mathbf{E}$ are estimated less precisely than the higher level parameters. They are also relatively sensitive to prior specifications. However, the estimated correlation parameters are close to the true values. An increase of the number of units helps better estimate $\mathbf{D}$ and decreases its sensitivity to priors. Similarly, a larger number of time periods identifies $\mathbf{E}$ more precisely.

7 Empirical Examples

In this section, I use the two motivating examples discussed at the beginning of the paper to demonstrate the implementation of the model. The state failure example shows that, by controlling for multiple sources of confounding and making good use of information

---

[14]There are $5 \times 3 + 2 \times 2$ such parameters, and it would take too much space to report all of them in the paper.
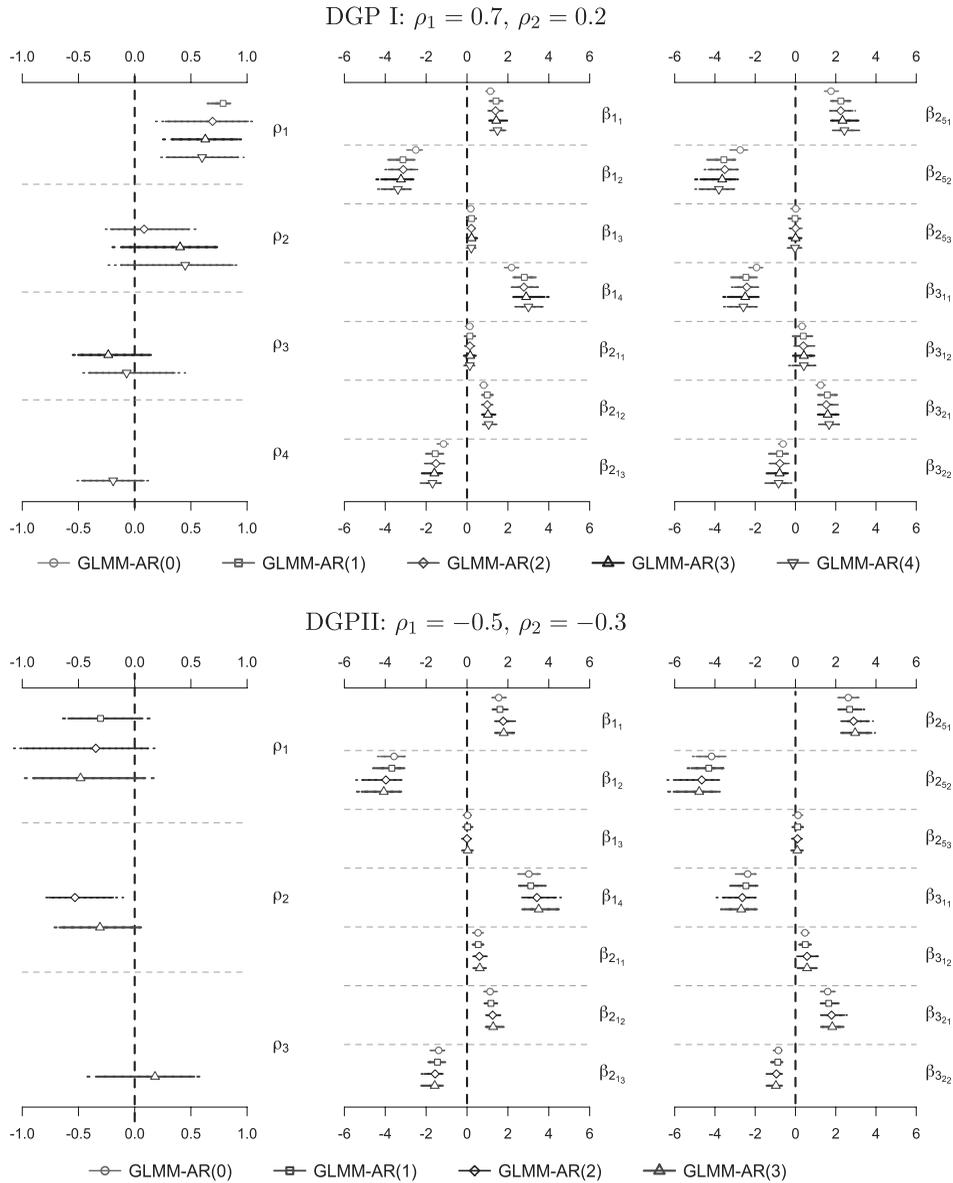
Fig. 2 Posteriors of GLMM-AR(p) models with different lag orders.
Note. The solid lines are 90% credible intervals and the dashed lines are 95% credible intervals. Posteriors of the competing models are grouped by parameters. The true values of the parameters presented in the graphs are as follows: for the FE parameters, $\beta_{1_1} = 2$, $\beta_{1_2} = -4$, $\beta_{1_3} = 0$, and $\beta_{1_4} = 4$; for the mean parameters of the unit-specific effects, $\beta_{2_{11}} = 0.5$, $\beta_{2_{12}} = 1$, $\beta_{2_{13}} = -2$, $\beta_{2_{13}} = -2$, $\beta_{2_{51}} = 3$, $\beta_{2_{52}} = -4$, and $\beta_{2_{53}} = 0$; and for the mean parameters of the time-specific effects, $\beta_{3_{11}} = -3$, $\beta_{3_{12}} = 0.2$, $\beta_{3_{21}} = 2$, and $\beta_{3_{22}} = -1$.

in the error term, the GLMM-AR(p) model fits the data better than conventional models and considerably improves within-sample forecasting. The civil war example highlights the necessity of serial correlation diagnostics in TSCS data analysis: falsely assuming independent errors ignores the possibility of nonstationarity and may result in spurious

b_5: Residual variation of random-effect parameter $\beta_{2_5,i}$
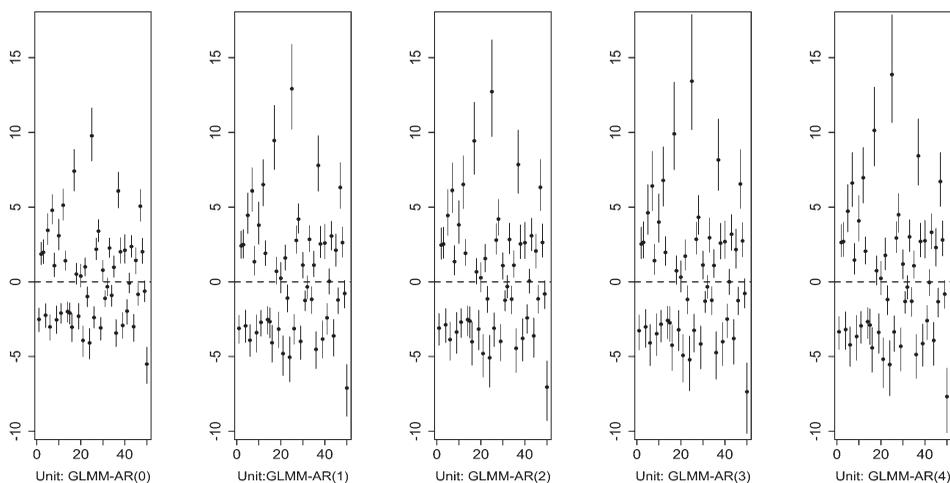
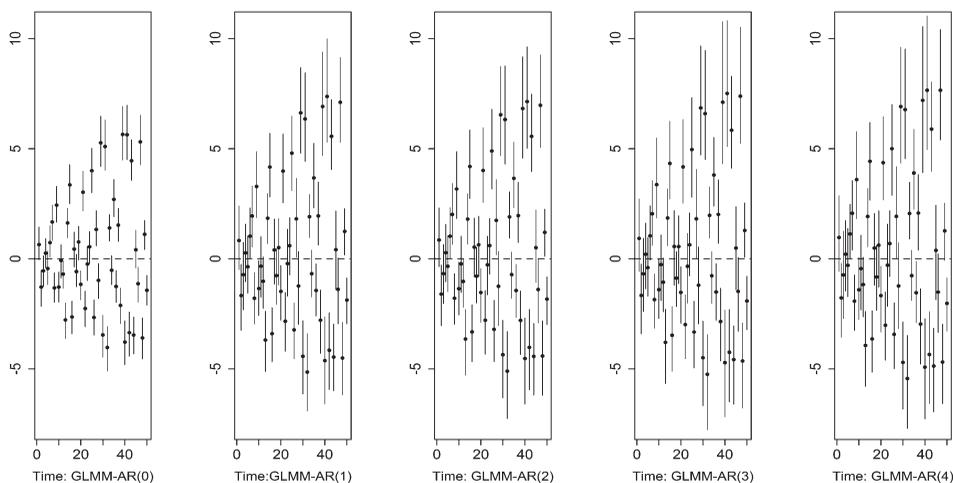c_1: Residual variation of random-effect parameter $\beta_{3_1,t}$

Fig. 3 Heterogeneity in unit and time dimensions (two examples).
Note. The dots in the graphs indicate posterior mean. The vertical lines are the 95% credible intervals of the posteriors of the group-level errors, b_5 and c_1, measuring the unexplained variation of the random-effect parameters, $\beta_{2_5,i}$ and $\beta_{3_1,t}$. Each vertical line is a group-specific residual distribution. Salient heterogeneity is reflected by the varying locations and scales of the posteriors.

regressions. This empirical example demonstrates how the GLMM-AR(p) model helps detect suspicious cases of noncointegration.

7.1 *State Failure in the Sub-Saharan Africa*

Since 1994, the PITF has been building statistical models to explain and forecast state failure (see *Phase I Findings* to *Phase IV Findings* available at http://globalpolicy.gmu .edu/pitf). King and Zeng (2001b) gave a comprehensive critique of the methodology applied by PITF for drawing causal inferences and conducting prediction and provided suggestions on how to improve state failure data analysis. From a different perspective,

Table 2 Within-group data variation: state failure example

Variable	Symbol	Within-country variation			Within-year variation		
		Minimum	Mean	Maximum	Minimum	Mean	Maximum
State failure	Failure	0.25	0.42	0.58	0.39	0.49	0.58
Democracy	Demo	0.00	1.51	3.72	2.76	3.38	5.13
Party fractionalization	Partyfrac	0.00	0.08	2.26	0.00	0.12	0.86
Party legitimacy	Partyleg	0.00	0.77	1.32	0.82	1.09	1.53
Regime durability	Regundur	0.00	0.08	0.28	0.06	0.19	0.22
Calories per capita (consumed)	Calory	0.03	0.17	0.35	0.29	0.35	0.44
GDP per capita	Gdppc	0.12	0.57	6.90	0.54	1.68	3.61
Neighbors in conflict	neighI	0.00	0.70	1.78	0.00	0.98	1.58
Neighbors in civil/ethnic war	neighII	0.00	0.62	1.36	0.00	0.85	1.53
Infant mortality	Infmort	0.07	0.29	0.99	0.04	0.53	0.73
Political terror scale	Terror	0.49	0.76	1.15	0.00	0.84	1.23
Political discrimination	Discrim	0.00	0.66	0.89	0.00	0.68	0.84
Secondary school enrollment	Enroll	0.00	0.01	0.03	0.01	0.02	0.02
Change of democracy	Cdemo	0.00	1.14	2.53	0.00	1.08	4.95
Trade openness	Trade	0.10	0.34	0.57	0.02	0.48	0.62

Note. The table presents the data variation in the time and spatial dimensions. In the within-county (within-year) variation column is the summary of how the variables vary within each country (year). If a variable is invariant or slowly moving (with any 0 or close to 0 value in the row of a variable), it will cause identification and estimation problems for the FE estimator. GDP, gross domestic product.

I use the state failure data of the sub-Saharan countries to illustrate how the GLMM-AR(p) model improves statistical inference and forecasting.

In the imputed data set, there are 1214 country-years, and there are 446 state failures.[15] The proportion of state failure is 36.74%, which means that state failure is not a rare event in the sub-Saharan region. Hence, case–control resampling (Breslow 1996; King and Zeng 2001a, 2001c) is not necessary or preferred given the fact that it unnecessarily discards observations and breaks the data structure. Following PITF and King and Zeng (2001b), I lagged all the explanatory variables for 2 years. This is not only to avoid simultaneity but also to use the statistical model to do 2-year-ahead prediction. Table 2 summarizes the within-country and within-year variation of the data.[16] There are three variables, party fractionalization, regime durability, and male secondary school enrollment, which have little cross-year variation for most countries, causing serious problems with the FE model. The cross-sectional variation of almost all variables changes considerably from year to year, implying heterogeneity between the sample time periods. In the literature, there is a debate about the effect of trade openness on state failure: PITF and Beck et al. (2002) find that it significantly reduces the risk of state failure, but King and Zeng (2001b) argue that the effect is unclear and uncertain. If the overall effect of trade

---

[15]There is a large proportion of missing data in the original data set. Bayesian data augmentation is a preferred method to handle missingness than multiple imputation (Gelman et al. 1995; Gill 2007). However, due to the more than 2000 parameters in the GLMM-AR(p) model, Bayesian data augmentation for such a large proportion of missing data will greatly increase the computing time. For this reason, I use the multiple imputation method to handle the missingness before implementing the Bayesian MCMC simulation.

[16]The variables are chosen from the "candidate" covariates in *PITF Phase III Findings* (p. 24) by using stochastic search variable selection method (Pang and Gill 2010) with 0.5 as the threshold value.

openness is ambiguous, it may be because its effect is so heterogeneous across countries that a cross-sectional generalization is misleading. Therefore, I assign the covariate trade openness with a country-specific coefficient. Country- and year-specific intercepts are also included in the GLMM-AR(p) model to control for unobserved heterogeneity.

Because I do not have strong prior beliefs about the parameters in the model, the priors used in this example are uninformative. For the purpose of model comparison, I assign the same priors to the common parameters in competing models. I followed two basic rules when specifying priors: the first rule is that the prior should not affect the posterior in an important way and the second is that the sample space defined by the prior should be mathematically and substantively proper. I set $\boldsymbol{\beta} \sim N(\mathbf{0}, 400 \times \mathbf{I})$. Centering the prior mean at 0 is conventionally interpreted as hypothesis testing on whether the coefficients are different from 0 (Alston et al. 2005; Gill 2007). The variance is chosen based on my experience: such a variance specification normally ensures that the prior has negligible influence on the posterior. About the priors of $\mathbf{D}$ and $\mathbf{E}$, the smaller $\mathbf{D}_0^{-1}$, v_0, $\mathbf{E}_0^{-1}$, and η_0, the smaller effect the priors have on the posteriors (refer to Step 4 of the MCMC algorithm). However, because the sample countries are all in the sub-Saharan region, excessively large prior variances on $\mathbf{D}$ and $\mathbf{E}$ are not substantively defensible. I set $\mathbf{D} \times \mathbf{E} \sim W(40, 20 \times \mathbf{I}_{40}) \times W(40, 20 \times \mathbf{I}_{40})$. This prior choice also helps avoid numerical instability. Other parameters are easy to choose, except the lag order p. The lag order is determined *ex post* by model selection instead of being parameterized in the model. I checked robustness by using different sets of priors with reasonable changes. With the sample size of 1214, the priors do not have visible influence on $\boldsymbol{\beta}$ and $\boldsymbol{\rho}$, but with the number of groups in both time and spatial dimensions equal to 40, the priors of $\mathbf{D}$ and $\mathbf{E}$ have observable effects on their posteriors. Nevertheless, they do not have visible influences on the substantively interesting parameters.

I estimated the competing models using the same computing platform as in the simulation studies. The average running time for a single iteration is about 3.72 s. Compared to the simulated examples in the previous section, the state failure models require more iterations (I did 200,000 iterations after 10,000 burn-in iterations). This is because colinearity in the empirical data causes higher cross-chain correlation, which slows down mixing and increases the difficulty of assessing (non)convergence. Each GLMM-AR(p) model takes about 9 days to complete the simulation process. I estimated six competing models, including the GLMM-AR(p) models with $p = 0, 1, 2, 3$, the completely pooled probit model (PROBIT), and a GLMM model without considering time heterogeneity (GLMM-CL1). The PROBIT, GLMM-CL1, and GLMM-AR(0) all assume that the errors are serially uncorrelated. The PROBIT model does not analyze any unobserved heterogeneity either. The GLMM-CL1 model only accommodates heterogeneity across units. The GLMM-AR(p) models took a longer time to converge than the PROBIT and GLMM-CL1 models, but with the PGM-MGMC method, their mixing time was considerably shortened. Figure 4 shows how much the method reduces within-chain correlations and improves MCMC mixing.

The posteriors and marginal likelihood of each model are reported in Fig. 5. Positive serial correlation is detected: $\hat{E}(\rho_1) = 0.37$ in the GLMM-AR(1) and $\hat{E}(\rho_1) = 0.27$ and $\hat{E}(\rho_2) = 0.22$ in the GLMM-AR(2); ρ_1 and ρ_2 are greater than 0 at a 95% credibility level. The PROBIT ignores serial correlation and heterogeneity, and its estimates and standard errors are noticeably different from the other five models. The model exaggerates the effects of almost all the variables and leads to over-confidence on the estimates. The pooled model suggests that nine explanatory variables are important with high certainty. It also strangely supports the argument that higher male secondary school enrollment leads to higher risk of state failure, which is contradictory to current theory (Sambanis 2002;
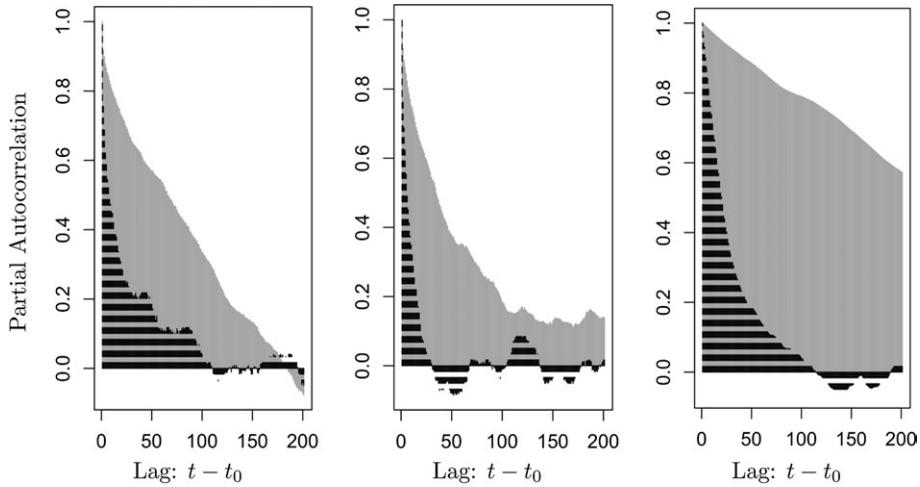
Fig. 4 Mixing improvement by PGM-MGMC updating: state failure example.
Note. The gray shadow indicates the autocorrelation of the standard Gibbs chains, whereas the black shadow shows the decrease of autocorrelation when applying the PGM-MGMC. From the left to the right, the graphs illustrate the autocorrelation of the Markov chains of a FE parameter, a random-effect parameter, and one element in the covariance matrix **D**. These parameters are randomly chosen.

Fearon and Laitin 2003; Collier and Hoeffler 2004). The other five models all find that six of the nine ''important'' variables in the PROBIT lose both importance and certainty after controlling for heterogeneity. As shown in Fig. 5, the posteriors of the GLMM-AR(p) models with $p > 0$ are similar across the models, and the BFs suggest that the GLMM-AR(2) has
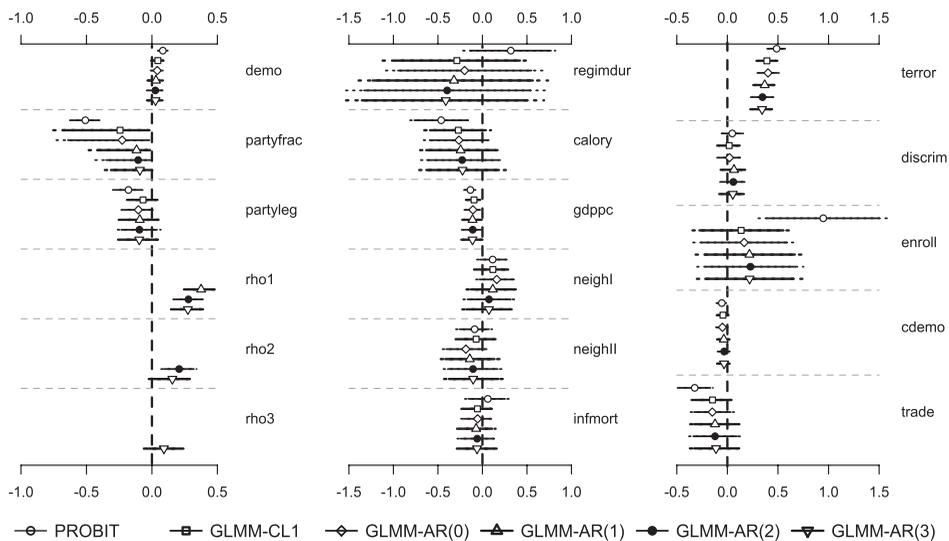

Fig. 5 Posterior summary of six competing models: state failure example.
Note. Log marginal likelihood: PROBIT $= -584.794$, GLMM-CL1 $= -511.185$, GLMM-AR(0) $= -503.931$, GLMM-AR(1) $= -470.357$, GLMM-AR(2) $= -468.638$, and GLMM-AR(3) $= -473.560$. The solid lines are 90% credible intervals and dashed lines are 95% credible intervals. Posteriors of the competing models are grouped by parameters.
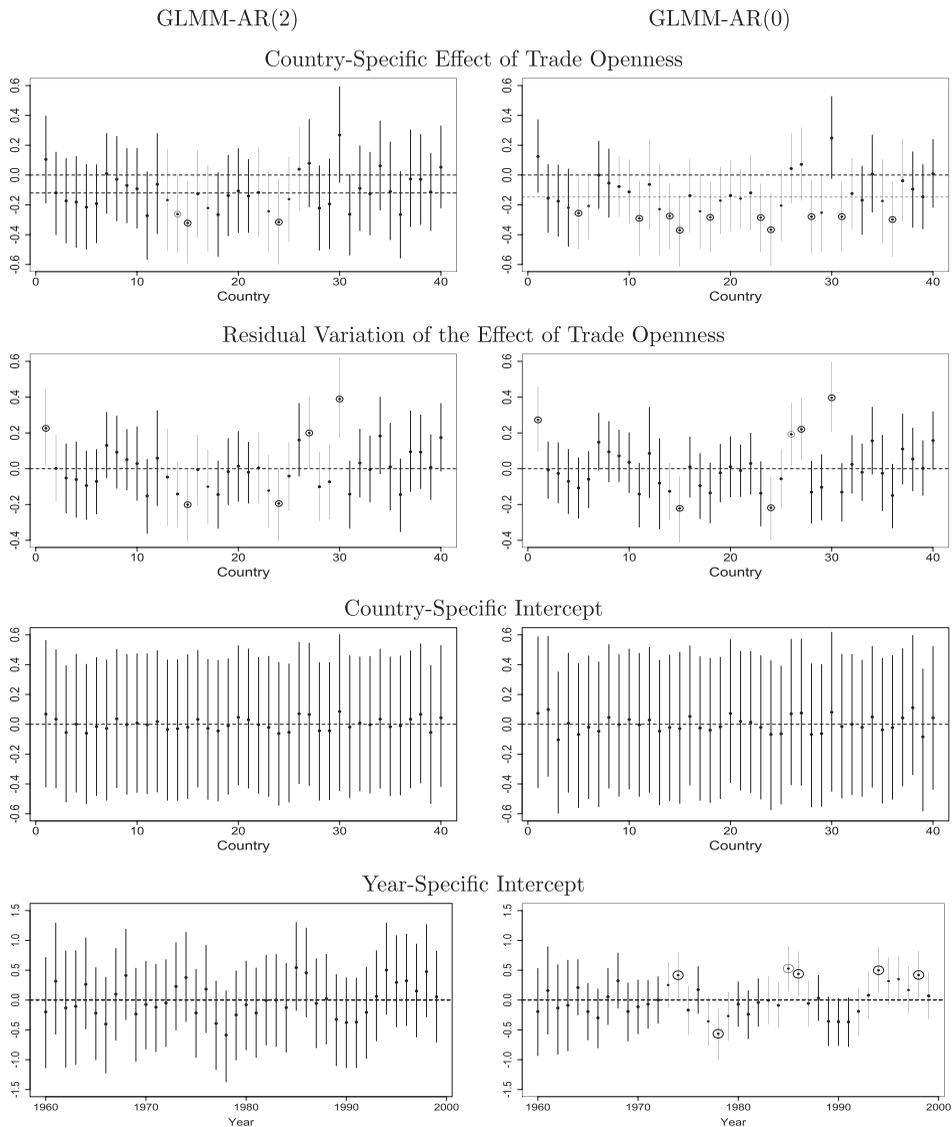
Fig. 6 Random effects, country-level residuals, and random intercepts.
Note. The vertical lines in the graphs are 90% credible intervals of the posteriors. The zero horizontal line is a reference line, and the nonzero horizontal lines in the first-row graphs is the overall mean of the country-specific coefficient associated to trade openness. The other three random parameters are residuals that are mean 0 by definition, a fact that is confirmed by the estimates.

the best goodness of fit, slightly better than the GLMM-AR(1) ($\log_{10} = 0.74$), but decisively better than the GLMM-AR(3) ($\log_{10} = 2.14$). It fits the data much better than the multilevel models without modeling serial correlation (the $\log_{10}$ BF of the GLMM-AR(2) versus the GLMM-AR(0) is 15.33 and that of the GLMM-AR(2) versus the GLMM-CL1 is 18.48). The PROBIT model has poor model quality, and the $\log_{10}$ BF of it versus the best model is as small as -50.45.

    Figure 6 compares the posteriors of the random-effect parameters of the GLMM-AR(2) and the GLMM-AR(0). The graphs show that correcting serial correlation makes

a difference in estimating the random-effect coefficients and group-level residuals. Heterogeneity is detected in both the time and the spatial dimensions. The effect of trade openness varies from country to country, a fact that can substantively be attributed to the different structure of international trade in different countries. An open economy implies that a country is engaged in the international community, which has been suggested to be the mechanism through which trade openness reduces state failure risk (*PITF Phase III Report* [Goldstone et al. 2000]). However, if the major products are primary goods, this trade structure may be highly correlated with high domestic instability. This is because primary commodity exports may imply high dependency on natural resources, a weak government, and profitable rebellions for lootable resources, all of which lead to high risk of state failure (Doyle and Sambanix 2000; Sambanis 2001; Collier and Hoeffler 2004). In the GLMM-AR(2) model, there are only three countries (Gabon, Gambia, and Malawi) in which trade openness decreases state failure risk at a 90% credibility level. For other countries, an open market has no clear effect on state failure. The country-specific random intercept shows relative homogeneity—there is not much unobserved heterogeneity after the country-level errors are modeled (as shown in graphs in the second row of Fig. 6). However, the year-specific intercept demonstrates noticeable heterogeneity between time periods. Compared to the random effects in the GLMM-AR(2), the posteriors of the GLMM-AR(0) have smaller error bands and consequently put more confidence on the effect of trade openness—at a 90% credibility level, trade openness has a negative effect on state failure in 10 instead of the 3 countries in the GLMM-AR(2).

One of the main goals of the PITF studies is to build a forecasting system based on statistical models. Prediction is another means of assessing model quality. Here, I also compare these models by assessing their within-sample prediction performance. To avoid setting arbitrary or *post hoc* thresholds for classification, I simply report the numeric predictive probabilities of state failure of all the sample country-years. In this way, we can compare the predictive failure probabilities of the observed failure group and the observed nonfailure group. As shown in the upper graphs in Fig. 7, the PROBIT performs poorly in terms of distinguishing failure cases from nonfailure ones, and the density kernels of the two groups largely overlap. The GLMM-CL1 separates the two densities better than the pooled model, but the two kernels are still poorly separated. The GLMM-AR(0) considers heterogeneity both across units and between time periods, and it further reduces the overlapping area. By considering the error dynamics and using the information ignored by the former models, the GLMM-AR(2) distinguishes the two groups much more accurately. In the graph, the two density kernels are well separated—only their tails are slightly connected.

An alternative way to evaluate and compare the predicting performance of competing models is the *receiver operating characteristic (ROC) curve*, used by King and Zeng (2001b). The idea of the ROC curve is simple: at a given rate of correct classification of one group (say, the failure group), a model performs better than others if it has the highest rate of correct classification of the other group (say, the nonfailure group). Graphically, the dominating curve in an ROC graph represents the best model. This approach has the advantage of not using arbitrary cutoff values. In Fig. 7, the diagonal line is a reference line, indicating the extreme situation that the densities of the two groups completely overlap. In the figure, the ROC curve of the pooled probit model has the lowest location, indicating the worst prediction. The GLMM-CL1 model improves forecasting by modeling heterogeneity across countries, and its curve globally dominates that of the pooled probit model. The GLMM-AR(0) model has an ROC curve globally—but only slightly—above that of the GLMM-CL1, but it is dominated everywhere by the curve of the GLMM-AR(2) model.
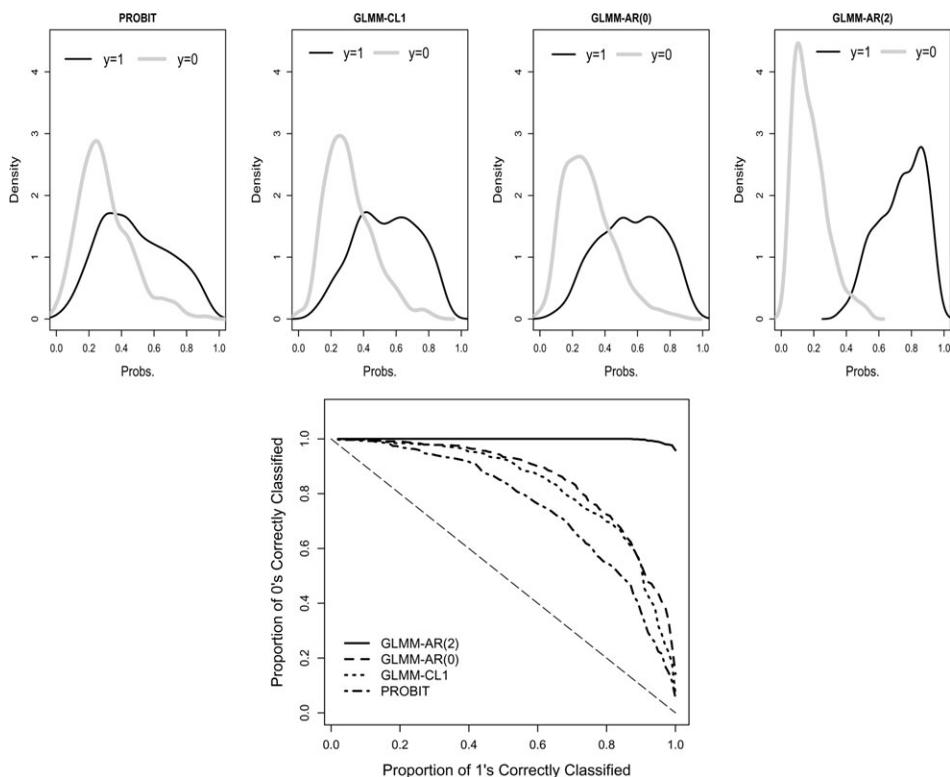
Fig. 7 Comparison of within-sample predictive probabilities.
Note. In the graphs of predictive probabilities, the gray lines are distributions of the predictive failure probability of the observed nonfailure group and the black lines are those of the observed failure group. In the ROC graph, the 45° line is a reference line, and the ROC curves plot the rate of correct classification of one group with a given rate of correct classification the other.

The ROC curve of the GLMM-AR(2) model is almost a horizontal line, showing that there is barely a trade-off between the two types of classification.

7.2 *Civil War Duration in Sub-Saharan Africa*

Political scientists are fully aware of the nonstationarity problem in time-series analysis (Beck 1993; Durr 1993; Smith 1993; Williams 1993; Box-Steffensmeier and Tomlinson 2000; DeBoef 2001), but this concern has not been commonly extended to TSCS studies. In this example, I use the civil war duration example to demonstrate that the GLMM-AR(p) model can serve as a cointegration test to help detect spurious regressions. The model analyzes the residual (error) process with the restriction that the process be stationary. Whenever this restriction is violated, the MCMC simulation will show abnormal symptoms: the Markov chain is halted because of the difficulty of sampling legitimate draws for the autoregressive coefficients, or the MCMC samples of the autoregressive coefficients are highly concentrated around the edge of the unit circle. In either situation, we should change the model specification to achieve cointegration by altering variable selection, transforming the data (for instance, by differencing), or changing other aspects of model specification.

The civil war literature includes studies on war *onset*, war *duration*, and war *termination* (Sambanis 2002). In the literature, some argue that civil war onset and duration are

Table 3 Within-group data variation: civil war duration example

Variable	Symbol	Within-country variation			Within-year variation		
		Minimum	Mean	Maximum	Minimum	Mean	Maximum
Ongoing war	war	0.00	0.17	0.51	0.36	0.43	0.47
Income per capita $(\log)_{t-1}$	gdpl	0.02	0.11	0.33	0.55	0.60	0.66
Population $(\log)_{t-1}$	popl	0.05	0.15	0.21	1.17	1.21	1.23
Mountainous	mountainous	0.00	0.00	0.00	1.43	1.45	1.50
Ethnolinguistic fractionalization	ethfrac	0.00	0.02	0.51	0.23	0.24	0.25
Religious fractionalization	relfrac	0.00	0.02	0.51	0.30	0.33	0.36
Oil-exporting country	oil	0.00	0.00	0.00	0.18	0.19	0.20
Political instability	instab	0.00	0.35	0.51	0.22	0.38	0.51
Anocracy$_{t-1}$	anocracy	0.00	0.29	0.51	0.22	0.39	0.51
Democracy$_{t-1}$	democracy	0.00	3.48	7.21	3.99	5.05	5.96
GDP growth rate$_{t-1}$	growth	0.02	0.05	0.20	0.04	0.07	0.13

Note. The table presents the data variation in the time and spatial dimensions. In the within-county (within-year) variation column is the summary of how the variables vary within each country (year). If a variable is invariant or slowly moving (with any 0 or close to 0 value in the row of a variable), it will cause identification and estimation problems for the FE estimator. GDP, gross domestic product.

different phenomena that require different theories to explain (Collier, Hoeffler, and Soderbom 2004; Fearon 2004), others, such as Miguel, Satyanath, and Sergenti (2004), suggest that the theories applied to civil war onset are also relevant to civil war duration. Most empirical civil war data are TSCS. Popular models applied in the extant empirical civil war studies include the pooled model and the FE and RE estimators, none of which consider the cointegration issue. This paper is not intended to solve the theoretical debate. Instead, the GLMM-AR(p) model is applied here to show that TSCS analysis of civil war duration without cointegration testing can produce unreliable inferences, which might shed light on the debate. The response process, civil war duration, may be unstationary, but many of the explanatory variables have little temporal variation (as shown in Table 3). This highlights the necessity of cointegration testing and justifies the concern with spurious regressions when applying the FE, RE, or pooled probit model.

Miguel, Satyanath, and Sergenti (2004) use *rainfall* as an instrumental variable for economic growth for the purpose of controlling for endogeneity. However, this is a weak instrumental variable, and it is also difficult to argue that rainfall affects civil war duration only through economic growth (in other words, it is uncorrelated with the error term). Furthermore, if economic growth time series cannot help achieve cointegration, it is unlikely that the rainfall data can. For these reasons, I use the lagged economic growth rate directly. Miguel, Satyanath, and Sergenti (2004) also tested whether the effect of economic growth varies across countries. I specify the GLMM-AR(p) model with both country- and year-specific random intercepts and a random-effect coefficient for the variable of economic growth. The definitions and within-group variations of all the variables are summarized in Table 3. I do not include two often-selected variables in civil war onset studies—a dummy variable for being a noncontiguous state and an indicator for being new state—because they have little overall variation across the sample country-years. How the effect of economic growth varies across countries is further explained with
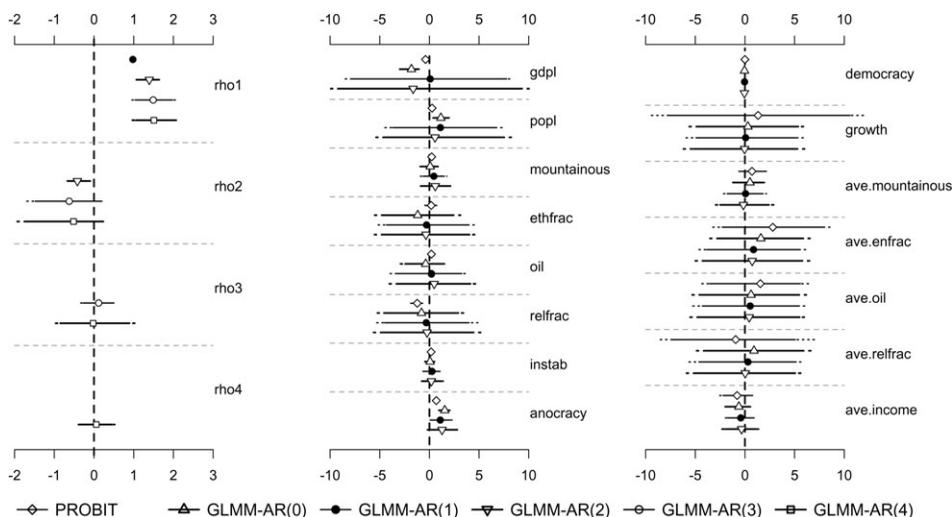
Fig. 8 Summary of MCMC outputs of competing models.
Note. Log marginal likelihood: PROBIT $= -399.514$ and GLMM-AR(0) $= -327.632$. For the GLMM-AR(p) models with $p = 1, 2, 3, 4$, because the MCMC processes were abnormally halted, the Markov chains did not converge and no marginal likelihood was calculated.

the same five variables as in Miguel, Satyanath, and Sergenti (2004), including how mountainous a country is, whether the state is an oil-exporting country, economic development level, male secondary school enrollment, and democracy. There are 743 observations (country-years) in total among which are 182 ongoing wars.

With a similar prior assignment as in the state failure study, I estimated six models with or without considering serial correlation. Figure 8 summarizes the MCMC outputs of the competing models. The GLMM-AR(0) model controls for heterogeneities in both time and spatial dimensions and produces different estimates from those of the pooled probit model. Compared to the probit model, the GLMM-AR(0) has much larger error bands for most parameters and detects fewer statistically important variables. Nonetheless, both models suggest that theories applied to civil war onset are relevant to civil war duration—the reliability and directions of the effects of most covariates are similar to those in civil war onset models. [17] Then I estimated GLMM-AR(p) models with $p = 1, 2, 3, 4$. All the simulations were abnormally halted because of the difficulty of drawing legitimate samples of the autoregressive coefficients. The left graph in Fig. 8 summarizes the samples of the autoregressive coefficients before the MCMC simulations were halted. The draws in all the GLMM-AR models are close to the edge of the unit circle, showing a significant tendency toward going out of the stationarity space. In the middle and right graphs of Fig. 8 are the summaries of the MCMC draws of the GLMM-AR(1) and -AR(2) models. Note that the outputs cannot tell us anything about the parameters because they are not sampled from the target distributions. I use them to demonstrate how different the empirical results could be when we seriously consider the residual dynamics and the noncointegration problem.

The evidence found in the MCMC simulation process suggests that the explanatory variables are unlikely to be cointegrated with civil war duration, and statistical inferences

---

[17]For the detailed empirical findings in civil war onset studies, refer to the quantitative studies on civil war onset, such as Fearon and Laitin (2003), Cederman and Girardin (2007), and Fearon, Kasara, and Laitin (2007).

based on the pooled probit, or GLMM-AR(0) model are spurious and misleading. The solution could be adding a lagged response variable as a regressor in addition to the autoregressive errors or finding other time-varying explanatory variables to achieve cointegration. This example shows that the GLMM-AR(p) model can be applied for the purpose of cointegration testing and avoiding spurious regressions in discrete TSCS analysis.

8 Conclusions

In TSCS analysis, modeling heterogeneity in both the time and the spatial dimensions while adequately correcting serial correlation is required by the generic characteristics of the TSCS data structure. However, for categorical responses, this is especially challenging because of nonlinearity and the complex error structure. Moreover, because TSCS models analyze multiple time series, cointegration testing is necessary and important. This paper proposes the GLMM-AR(p) model as a comprehensive solution and develops an MCMC algorithm for model estimation. The error-orthogonalizing method facilitates the construction of the MCMC algorithm and achieves the goal of conducting data augmentation in a single block. It also simplifies BF computation, making Bayesian model comparison easy and convenient. The application of the PGM-MGMC method further improves simulation efficiency. An R package GLMMarp is available on CRAN for estimating various specifications within the GLMM-AR(p) framework.[18] A possible future extension of the model could be to relax the assumption that time-specific common shocks have the same impact on all units. This extended model may be a multifactor residual model with a hierarchical component and serially correlated errors.

Funding

References

Achen, Christopher H. 2001. *Why lagged dependent variables can suppress the explanatory power of other independent variables.* Working paper.

Albert, James A., and Siddhartha Chib. 1993. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 88:669–79.

Alston, Clair, Petra Kuhnert, Low S. Choy, R. McVinish, and K. Mengersen. 2005. *Bayesian model comparison: Review and discussion.* International Statistical Insitute, 55th session.

Andrews, Donald W. K. 1991. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica* 59:817–58.

Aschbacher, Michael. 2000. *Finite group theory.* Cambridge: Cambridge University Press.

Bai, Jushan, and Serena Ng. 2004. A panic attack on unit roots and cointegration. *Econometrica* 72:1127–77.

Beck, Nathaniel. 1993. The methodology of cointegration. *Political Analysis* 4:237–48.

Beck, Nathaniel, David Epstein, Simon Jackman, and Sharyn O'Halloran. 2002. *Alternative models of dynamics in binary time-series-cross-section models: The example of state failure.* Working paper.

Beck, Nathaniel, and Jonathan N. Katz. 1995. What to do (and not to do) with time-series cross-section data. *American Political Science Review* 89:634–47.

———. 1996. Nuisance vs. substance: Specifying and estimating time-series-cross-section models. *Political Analysis* 6:1–36.

———. 2007. Random coefficient models for time-series-cross-section data: Monte Carlo experiments. *Political Analysis* 15:182–95.

---

[18]The GLMMarp package can be downloaded from http://cran.r-project.org/web/packages/GLMMarp/index.html.

———. 2009. *Modeling dynamics in time-series-cross-section political economy data.* Working paper.

Beck, Nathaniel, Jonathan N. Katz, and Richard Tucker. 1998. Taking time seriously: Time-series-cross-section analysis with a binary dependent variable. *American Journal of Political Science* 42:1260–88.

Bogopolski, Oleg. 2008. *Introduction to group theory.* Zürish, Switzerland: European Mathematical Society.

Borsch-Supan, A., and V. Hajivassiliou. 1993. Smooth unbiased multivariate probability simulators for maximum likelihood estimation of limited dependent variable models. *Journal of Econometrics* 58:347–68.

Box, George E. P., Gwilym M. Jenkins, and Gregory C. Reinsel. 1994. *Time series analysis: Forecasting and control.* 3rd ed. Englewood Cliffs, NJ: Prentice Hall.

Box-Steffensmeier, Janet M., and Andrew R. Tomlinson. 2000. Fractional integration methods in political science. *Electoral Studies* 19:63–76.

Breslow, Norman E. 1996. Statistics in epidemiology: The case–control study. *Journal of the American Statistical Association* 91:14–28.

Briggs, William L. 1987. *A multigrid tutorial.* Philadelphia, PA: Society for Industrial and Applied Mathematics.

Carlin, Bradley P. 1996. Hierarchical longitudinal modeling. In *Markov chain Monte Carlo in practice*, ed. S. Richardson, W. R. Gilks, and D. J. Spiegelharlter, 303–19. London: Chapman and Hall.

Cederman, Lars-Erik, and Luc Girardin. 2007. Beyond fractionalization: Mapping ethnicity onto nationalist insurgencies. *American Political Science Review* 101:173–85.

Chib, Siddhartha. 1993. Bayes regression with autoregressive errors: A Gibbs sampling approach. *Journal of Econometrics* 58:275–94.

———. 1995. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* 90:1313–21.

Chib, Siddhartha, and Edward Greenberg. 1994. Bayesian inference in regression models with ARMA (p, q) errors. *Journal of Econometrics* 64:183–206.

Chib, Siddhartha, and Ivan Jeliazkov. 2001. Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association* 96:270–81.

———. 2006. Inference in semiparametric dynamic models for binary longitudinal data. *Journal of the American Statistical Association* 101:685–700.

Choi, In. 2001. Unit root tests for panel data. *Journal of International Money and Finance* 20:249–72.

Collier, Paul, and Anke Hoeffler. 2004. Greed and grievance in civil war. *Oxford Economic papers* 56:563–95.

Collier, Paul, Anke Hoeffler, and Mans Soderbom. 2004. On the duration of civil war. *Journal of Peace Research* 41:253–73.

Cowles, Mary K., Bradley P. Carlin, and John E. Connett. 1996. Bayesian tobit modeling of longitudinal ordinal clinical trial compliance data with nonignorable missingness. *Journal of the American Statistical Association* 91:86–98.

DeBoef, Suzanna. 2001. Modeling equilibrium relationships: Error correction models with strongly autoregressive data. *Political Analysis* 9:78–94.

Doyle, Michael W., and Nicholas Sambanix. 2000. International peacebuilding: A theoretical and quantitative analysis. *American Political Science Review* 94:779–802.

Durr, Robert. 1993. An essay on cointegration and error correction models. *Political Analysis* 4:185–228.

Engle, Robert F., and Clive W. J. Granger. 1987. Cointegration and error correction: Representation, estimation and testing. *Econometrica* 55:251–76.

Fearon, James D. 2004. Why do some civil wars last so much longer than others. *Journal of Peace Research* 41:275–301.

Fearon, James D., Kmuli Kasara, and David D. Laitin. 2007. Ethnic minority rule and civil war onset. *American Political Science Review* 101:187–93.

Fearon, James D., and David D. Laitin. 2003. Ethnicity, insurgency, and civil war. *American Political Science Review* 97:75–90.

Franzse, Robert J., and Jude C. Hays. 2007. Spatial econometric models of cross-sectional interdependence in political science panel and time-series-cross-section data. *Political Analysis* 15:140–64.

———. 2008a. Empirical models of spatial interdependence. In *Oxford handbook of political ethodology*, ed. J. Box-Steffensmeier, H. Brady, and D. Dollier, 570–604. Oxford: Oxford University Press.

———. 2008b. Empirical modeling of spatial interdependence in time-series cross-sections. In *Methods of comparative political and social science: New developments & applications*, ed. S. Pickel, G. Pickel, H.-J. Lauth, and D. Jahn. Wiesbaden: Westdeutscher Verlag.

Garrett, Geoffrey. 1998. Global markets and national politics: Collision course or virtuous circle? *International Organization* 52:787–824.

Gelman, Andrew, John B. Carlin, S. HalStern, and Donald B. Rubin. 1995. *Bayesian data analysis.* New York: Chapman and Hall.

Gelman, Andrew, and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.

Geweke, John. 1991. Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints. In *Computing Science and Statistics: Proceedings of the Twenty Third Symposium on the Interface*, ed. E. M. Keramidas, 571–8. Fairfax, VA: Interface Foundation of North America.

———. 1996. Bayesian inference for linear models subject to linear inequality constraints. In *Modeling and prediction: Honouring Seymour Geisser*, ed. W. O. Johnson, J. C. Lee, and A. Zellner. New York: Springer.

Gill, Jeff. 2007. *Bayesian methods: A social and behavioral sciences approach*. 2nd ed. Boca Raton, FL: Chapman and Hall.

Goldstone, Jack A., Ted Robert Gurr, Barbara Harff, Marc A. Levy, Monty G. Marshall, Robert H. Bates, David L. Epstein, Colin H. Kahl, Pamela T. Surko, John C. Ulfelder, and Alan U. Unger. 2000. *State failure task force report: Phase III findings*. McLean, VA: Science Applications International Corporation.

Goodman, Jonathan, and Alan D. Sokal. 1989. Multigrid Monte Carlo method: Conceptual foundations. *Physical Review D* 40(6):2035–72.

Gourieroux, Christian, A. Monfort, and A. Trognon. 1984. Estimation and test in probit models with serial correlation. In *Alternative approaches to time series analysis*, ed. J. P. Florens, M. Mouchart, J. P. Raoult, and L. Simar. Brussels: Publications des Facultes Universitaires Saint-Louis.

———. 1985. A general approach to serial correlation. *Econometric Theory* 1:315–40.

Hagenaars, Jacques A. 1990. *Categorical longitudinal data: Log-linear analysis of panel, trend and cohort data*. London: Sage.

Hamilton, James Douglas 1994. *Time series analysis*. Princeton, NJ: Princeton University Press.

Han, Cong, and Bradley Carlin. 2001. Markov chain Monte Carlo methods for computing Bayes factors: A comprehensive review. *Journal of the American Statistical Association* 96:1122–32.

Heckman, James. 1981. Heterogeneity and state dependence. In *Labor markets*, ed. S. Rosen, 91–131. Chicago, IL: University of Chicago Press.

Hubrich, Kirstin, Helmut Luetkepohl, and Pentti Saikkonen. 2001. A review of systems cointegration tests. *Econometric Reviews* 20:247–318.

Ibrahim, Joseph G., and Kenneth Klainman. 1998. Bayesian inference for random effect models. In *Practical nonparametric and semiparametric Bayesian statistics*, ed. D. Dey, P. Mueller, and D. Sinha. New York: Springer.

Im, Kyung So, M. Hashem Pesaran, and Yongcheol Shin. 2003. Testing for unit roots in heterogeous panels. *Journal of Econometrics* 115:53–74.

Kao, Chihwa. 1999. Spurious regression and residual-based tests for cointegration in panel data. *Journal of Econometrics* 90:1–44.

Keane, Michael P. 1994. A computational practical simulation estimator for panel data. *Econometrica* 62:95–116.

King, Gary, and Langche Zeng. 2001a. Explaining rare events in international relations. *International Organization* 55:693–715.

———. 2001b. Improving forecasts of state failure. *World Politics* 53:623–58.

———. 2001c. Logistic regression in rare events data. *Political Analysis* 9:137–63.

Liu, Jun S., and Chiara Sabatti. 2000. Generalised Gibbs sampler and multigrid Monte Carlo for Bayesian computation. *Biometrika* 87:353–69.

Liu, Jun S., and Ying Nian Wu. 1999. Parameter expansion for data augmentation. *Journal of American Statistical Association* 94:1264–74.

Lumley, Thomas, and Patrick Heagerty. 1999. Weighted empirical adaptive variance estimators for correlated data regression. *Journal of the Royal Statistical Society: Series B* 61:459–77.

Miguel, Edward, Shanker Satyanath, and Ernest Sergenti. 2004. Economic shocks and civil conflict: An instrumental variables approach. *Journal of Political Economy* 112:725–53.

Molenberghs, Geert, and Geert Verbeke. 2005. *Models for discrete longitudinal data*. New York: Spriner.

Mueller, Gernot, and Claudia Czado. 2005. An autoregressive ordered probit model with application to high-frequency financial data. *Journal of Computational & Graphical Statistics* 14:320–338.

Ng, Edmond S. W., James R. Carpenter, Harvey Goldstein, and Jon Rasbash. 2006. Estimation in generalized linear mixed models with binary outcomes by simulated maximum likelihood. *Statistical Modelling* 6: 23–42.

Olsen, Karen K., and Joseph L. Schafer. 2001. A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association* 96:730–45.

Pang, Xun. 2008. *Binary time series with AR(p) errors: Bayes factor for lag order determination and model selection.* Working paper.

Pang, Xun, and Jeff Gill. 2010. *Spike and slab prior distributions for simultaneous Bayesian hypothesis testing, model selection, and prediction, of nonlinear outcomes.* Working paper.

Pedroni, Peter. 1999. Critical values for cointegration tests in heterogeneous panels with multiple regressors. *Oxford Bulletin of Economics and Statistics* 61:653–70.

———. 2004. Panel cointegration: Asymptotic and finite sample properties of pooled time series tests with an application to the PPP hypothesis. *Econometric Theory* 3:579–625.

Peters, B. Guy, Jon Pierre, and Desmond S. King. 2005. The politics of path dependency: Political conflict in historical institutionalism. *The Journal of Politics* 67:1275–300.

Philips, Peter C. B., and Donggyu Sul. 2003. Dynamic panel estimation and homogeneity testing under cross section dependence. *Econometrics Journal* 6:217–59.

Pierson, Paul, and Theda Skocpol. 2002. Historical insitutionalism in contemporary political science. In *Political science: State of the discipline*, ed. Ira Katznelson, and V. Helen Milner. 692–721. New York: W.W. Norton.

Poirier, Dale J., and Paul A. Ruud. 1988. Probit with dependent observations. *The Review of Economic Studies* 55:593–614.

Renard, Didier, Geert Molenberghs, and Helena Geys. 2004. A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics & Data Analysis* 44:649–67.

Rodriguez-Yam, Gabriel, Richard A. Davis, and Louis L. Scharf. 2004. *Efficient Gibbs sampling of truncated multivariate normal with application to constrained inear regression*. Unpublished manuscript, Colorado State University.

Rudra, Nita. 2002. Globalization and the decline of the welfare state in less-developed countries. *International Organization* 56:411–45.

Sambanis, Nicholas. 2001. Do ethnic and nonethnic civil wars have the same causes?: A theoretical and empirical inquiry (Part I). *The Journal of Conflict Resolution* 45:259–82.

———. 2002. A review of recent advances and future directions in the quantitative literature on civil war. *Defence and Peace Economics* 13:215–43.

Sandor, Zsolt, and Peter Andras. 2004. Alternative sampling methods for estimating multivariate normal probabilities. *Journal of Econometrics* 120:207–34.

Schafer, Joseph L., and Recai M. Yucel. 2002. Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational & Graphical Statistics* 11:437–57.

Shor, Boris, Joseph Bafumi, Luke Keele, and David Park. 2007. A Bayesian multilevel modeling approach to time-series cross-sectional data. *Political Analysis* 15:165–81.

Singer, Judith D., and John B. Willett. 2003. *Applied longitudinal data analysis: Modelling change and event occurrence*. New York: Oxford University Press.

Skrondal, Anders, and Sophia Rabe-Hesketh. 2004. *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. New York: Chapman and Hall.

———. 2008. Multilevel and related models for longitudinal data. In *Handbook of multilevel analysis*, ed. Jan de Leeuw and Erik Meijer, 275–300. New York: Springer.

Smith, Robert. 1993. Error correction, attractions, and cointegration: Substantive and methodological issues. *Political Analysis* 4:249–54.

Sul, Donggyu. 2009. Panel unit root tests under cross section dependence with recursive mean adjustment. *Economics Letter* 105(1):123–6.

Thelen, Kathleen. 1999. Historical institutionalism in comparative politics. *Annual Review of Political Science* 2:369–404.

Williams, John. 1993. What goes around, comes around: Unit root tests and cointegration. *Political Analysis* 4:229–36.

Wilson, Sven E., and Daniel M. Butler. 2007. A lot more to do: The sensitivity of time-series-cross-section analyses to simple alternative specifications. *Political Analysis* 15:101–23.

Woods, Ngaire. 2001. International political economy in an age of globalization. In *The globalization of world politics*, ed. John Baylis and Steve Smith. New York: Oxford University Press.

Yang, Yang, Wenjiang Fu, and Kenneth C. Land. 2004. A methodological comparison of age-period-cohort models: The intrinsic estimator and conventional generalized linear models. *Sociological Methodology* 34(1):75–110.

Zeileis, Achim. 2004. Econometric computing with HC and HAC covariance matix estimators. *Journal of Statistical Software* 11(i10):1–17.