Universität Basel

Wirtschaftswissenschaftliche Fakultät

WWZ forum

March 2018

Estimating Interdependence Across Space, Time and Outcomes in Binary Choice Models Using Pseudo Maximum Likelihood Estimators

WWZ Working Paper 2018/11

Julian Wucherpfennig, Aya Kachi, Nils-Christian Bormann, Philipp Hunziker

A publication of the Center of Business and Economics (WWZ), University of Basel.

Universität Basel
Peter Merian-Weg 6
4052 Basel, Switzerland
wwz.unibas.ch

Corresponding Author:
Prof. Dr. Aya Kachi
Tel: +41 (0) 61 207 28 41
Mail: aya.kachi@unibas.ch

Estimating Interdependence Across Space, Time and Outcomes in Binary Choice Models Using Pseudo Maximum Likelihood Estimators

Julian Wucherpfennig[1], Aya Kachi[2], Nils-Christian Bormann[3], and Philipp Hunziker*[4]

[1]*Hertie School of Governance*

[2]*Faculty of Business and Economics, University of Basel*

[3]*Department of Politics, University of Exeter*

[4]*College of Computer and Information Science & Department of Political Science, Northeastern University*

March 30, 2018

PLEASE DO NOT CITE WITHOUT PERMISSION OF THE AUTHORS

Binary outcome models are frequently used in Political Science. However, such models have proven particularly difficult in dealing with interdependent data structures, including spatial autocorrelation, temporal autocorrelation, as well as simultaneity arising from endogenous binary regressors. In each of these cases, the primary source of the estimation challenge is the fact that jointly determined error terms in the reduced-form specification are analytically intractable due to a high-dimensional integral. To deal with this problem, simulation approaches have been proposed, but these are computationally intensive and impractical for datasets with thousands of observations. As a way forward, in this paper we demonstrate how to reduce the computational burder significantly by (i) introducing analytically tractable pseudo maximum likelihood estimators for latent binary choice models that exhibit interdependence across space, time and/or outcomes, and by (ii) proposing an implementation strategy that increases computational efficiency considerably. Monte-Carlo experiments demonstrate that our estimators perform similarly to existing alternatives in terms of error, but require only a fraction of the computational cost.

*We would like to thank Frederick Boehmke, Lars-Erik Cederman, Robert Franzese Jr., Dennis Quinn, Andrea Ruggeri, Emily Schilling, Martin Steinwand, Oliver Westerwinter, Michael Ward and Christopher Zorn for their useful comments.

1 Introduction

War vs. peace, vote "yes" vs. "no"—binary data are ubiquitous in political science. Thus, it is little surprising that binary choice models—e.g. probit and logit models—are amongst the most common statistical tools for empirical analyses. However, as is true for any other type of data, political scientists are becoming increasingly aware that observational data generated through social and political processes exhibits various forms of interdependence.

Take the simple example of international treaty ratification and compliance. A burgeoning literature that makes the case that countries are more likely to ratify international treaties, such as the Kyoto protocol to limit greenhouse gas emissions or the Ottawa convention on the ban of landmines, if they expect other countries to do the same. This pattern is often explained by international norms and the logic of public-goods provision (e.g., Nordhaus 1999; Simmons and Elkins 2004). Accordingly, one country's decision to ratify is (at least in part) dependent on decisions of other countries, and vice-versa. Following a large literature, we refer to this form of interdependence as "spatial interdependence".

Because such treaties are frequently the result of long-lasting negotiations, a related question pertains to the timing of ratification. Indeed, countries with a greater proclivity are generally expected to ratify sooner (such as Germany which is among the environmental frontrunners, Weidner and Mez (2008)), while an abstaining countries (such as the U.S. in both examples mentioned above) is likely to continue this behavior from one year to the next. Thus, we refer to this type of dependence as "temporal dependence".

Sceptics of international treaties have argued that in many instances, ratification does not necessarily occur because countries are willing to accept the constraint on the particular forms of behavior that is implied by the treaty (e.g., refrain from using landmines). Instead, countries could ratify if they were to comply even in the absence of the treaty. As such, treaties merely screen, rather than constrain (see e.g., Von Stein 2005). In this case, the causal arrow runs from compliance to ratification, rather than the other way around. Put differently, ratification and compliance are interdependent outcomes, making it difficult to

evaluate the actual effectiveness of treaties. In this paper, we refer to this pattern as "outcome interdependence".

All three forms of (inter-)dependence outlined above— across space, time and outcomes— suggest that individual observations violate the basic assumption of conventional statistical tools, including i.i.d. error terms for regression based estimators, or SUTVA for matching techniques. This violation poses a significant challenge when it comes to valid causal inference. While existing methods to tackle the challenge are relatively well developed for continuous data, there is a clear gap when it comes to binary data. As we show below, the main problem stems from a mathematically intractable error term distribution which prevents researchers from deriving closed form solutions for appropriate likelihood functions. While some simulation-based techniques exist, these are generally extremely computationally burdensome, even for small datasets, and thus often impracticable for applied researchers. Moreover, different forms of (inter-)dependence have generally been treated in isolation, even though typical political science data is likely to be affected by multiple forms at the same time.

Therefore the primary goals of our paper is to develop a general toolkit for applied researchers studying binary data featuring (inter-)dependencies. We do so with an eye on reducing computational burdens significantly in order to accommodate large datasets of several thousand observations by developing a series of pseudo maximum likelihood estimators. Our analytical point of departure is a pseudo maximum likelihood estimator for binary spatial models due to Smirnov (2010), for which the remaining computational burden amounts to inverting an N-dimensional matrix we refer to as the "interdependence multiplier." We then show that this estimation framework lends itself suitable to accommodate other types of interdependence as well, including temporal and outcome interdependence across simultaneous equations. Finally, we further reduce the estimation costs by proposing an implementation strategy that avoids direct matrix inversion, and instead relies on a combination of iterative gradient procedures and approximations that yield an estimation algorithm with almost

linear complexity in N.

The paper proceeds as follows. First, we provide a technical description of the mathematical problem associated with the three sources of interdependence in binary dependent variables. Then we introduce a pseudo maximum likelihood estimator (PMLE) for spatial models according to Smirnov (2010). Next we derive equivalent pseudo maximum likelihood estimators for accommodating dependence across time, and interdependence between outcomes. We then discuss our implementation strategy, followed by an evaluation of our estimators via Monte Carlo simulation. We conclude with a plan for further research.

2 The challenge: The intractable reduced-form error specification

In this section we set out to demonstrate two things. First, we use the binary spatial model to show why fitting models for binary spatial data is challenging. At its core, the problem is that the likelihood function for the binary spatial model involves an analytically intractable integral. Second, we show that this result generalizes to other sources of (inter-)dependence, namely across time and between outcomes. In fact, it turns out that spatial, temporal, and outcome (inter-)dependence all give rise to the same reduced form specification for the latent outcome vector. This result is a double edged sword. On the one hand, it implies that fitting models on binary data featuring any of these types of (inter-)dependencies will be challenging. On the other hand, it suggests that solving the estimation challenge for one type of model will be useful for addressing the other types of (inter-)dependencies as well.

2.1 Spatial interdependence

Suppose one is interested in the following model

$$y_i^* = \rho \sum_{j=1}^{N} w_{ij} y_j^* + \mathbf{x}_i \boldsymbol{\beta} + u_i \tag{1}$$

$$y_i = \begin{cases} 1 \text{ if } y_i^* > 0 \\ 0 \text{ otherwise} \end{cases} \tag{2}$$

whereas y_i^* is a continuous latent outcome variable, w_{ij} is a spatial lag between unit i and j indicating how closely the two units are connected in a given space (e.g, geographical proximity, membership in the same organizations etc.), $\mathbf{x}_i$ is a $1 \times k$ vector of covariates with parameter vector $\boldsymbol{\beta}$, and u_i is a zero-mean iid error term with fixed variance. We call this specification the binary spatial model. Note that in this specification, spatial dependence occurs on the level of the latent (i.e. not observed) outcome y_i^*. This specification follows Franzese et al. (2016), and it is suitable for cases where actors of our interest can observe or know more or less what others' latent characteristics are, and not only their revealed binary actions.

It is useful to write the latent equation in matrix notation, yielding

$$\mathbf{y}^*_{(N\times 1)} = \rho \mathbf{W}\mathbf{y}^* + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{3}$$

with

$$\mathbf{W}_{N\times N} = \begin{pmatrix} 0 & w_{12} & \cdots & w_{1N} \\ w_{21} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & w_{N-1,N} \\ w_{N1} & \cdots & w_{N,N-1} & 0 \end{pmatrix}. \tag{4}$$

$\mathbf{W}$ is commonly referred to as the *spatial weights matrix*. Throughout the paper we assume that $\mathbf{W}$ is row-standardized. Doing so ensures that the spatial process defined in (3) is stationary as long as $|\rho| < 1$ (Kelejian and Prucha 2010). Given (3) we can derive the reduced form as

$$\begin{aligned} \mathbf{y}^* &= (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{u} \\ &= (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{v}, \end{aligned} \tag{5}$$

where vector $\mathbf{v}$ contains the reduced-form error terms with non-spherical covariance matrix structure due to the multiplier $(\mathbf{I} - \rho\mathbf{W})^{-1}$.

Let us now derive the likelihood of the binary spatial model. Given y_i is binary, the

likelihood function assumes the following general form:

$$\begin{aligned} L(\rho, \boldsymbol{\beta}|\mathbf{X}, \mathbf{y}^*) &= \Big[\prod_{i=1}^{N} P(y_i = 1)\Big]^{y_i} \Big[\prod_{i=1}^{N} P(y_i = 0)\Big]^{(1-y_i)} \\ &= \Big[\prod_{i=1}^{N} P(y_i = 1)\Big]^{y_i} \Big[\prod_{i=1}^{N} \Big(1 - P(y_i = 1)\Big)\Big]^{(1-y_i)}, \end{aligned} \tag{6}$$

The main component of the likelihood function – the marginal probability that i takes 1 given $\mathbf{X}$ and the parameters – is,

$$\begin{aligned} P(y = 1) &= P(y_i^* > 0) \\ &= P\Big(\Big[(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}\Big]_i + v_i > 0\Big) \\ &= P\Big(v_i > -\Big[(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}\Big]_i\Big) \\ &= 1 - P\Big(v_i \leq -\Big[(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}\Big]_i\Big) \\ &= 1 - F_{V_i}\Big(-\Big[(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}\Big]_i\Big). \end{aligned} \tag{7}$$

where $[\cdot]_i$ indicates the i'th element of vector $[\cdot]$. $F_{V_i}(\cdot)$ is the marginal cdf of random variable V_i (the reduced form error term for unit i). Therefore, expression $F_{V_i}\Big(-\Big[(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}\Big]_i\Big)$ is the marginal cdf of V_i evaluated at $-\Big[(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}\Big]_i$. In theory, one can derive the marginal cdf of V_i based on the definition of a marginal cdf as follows

$$\begin{aligned} &F_{V_i}(-[(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}]_i) \\ &\quad = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{-\infty}^{-[(\mathbf{I}-\rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}]_i} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{V}}(s_1, \cdots, s_i, \cdots, s_N) ds_1 \cdots ds_i \cdots ds_N, \end{aligned} \tag{8}$$

where $f_{\mathbf{V}}(s_1, \cdots, s_N)$ is the joint pdf of the reduced-form error. The estimation challenge arises because evaluating F_{V_i} is generally analytically intractable as long as $\rho \neq 0$. As a consequence, direct maximum likelihood estimation of β and ρ is generally infeasible.

2.2 Temporal dependence

Consider the following time-series model

$$y_t^* = \mathbf{X}_t\boldsymbol{\beta} + \gamma y_{t-1}^* + u_t \tag{9}$$

$$y_t = \begin{cases} 1 \text{ if } y_t^* > 0 \\ 0 \text{ otherwise} \end{cases} \tag{10}$$

where the latent outcome exhibits a first-order temporal autoregressive process, governed by the correlation parameter γ with $|\gamma| < 1$.[1] We refer to this model as the binary temporal autoregressive model. For a discussion of this model in a political science context, see Beck et al. (2001). Note that we can rewrite the model in matrix notation as follows

$$\mathbf{y}^*_{(T\times 1)} = \mathbf{X}\boldsymbol{\beta} + \gamma\mathbf{T}\mathbf{y}^* + \mathbf{u}, \tag{11}$$

where $\mathbf{T}$ is defined as

$$\mathbf{T} = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix} \tag{12}$$

It is evident that this model is mathematically almost equivalent to the binary spatial model from the past section, the sole difference being that now we impose a weights matrix where the first minor diagonal (all the 1's) maps y_{t-1}^* to y_t^*. The reduced form of the autoregressive model is given by

$$\mathbf{y}^*_{(T\times 1)} = (\mathbf{I} - \gamma\mathbf{T})^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \gamma\mathbf{T})^{-1}\mathbf{u}. \tag{13}$$

Given this result, it is clear that this model gives rise to the same difficulties in ML estimation as the binary spatial model.

2.3 Dependence across outcomes

Finally, consider the following simultaneous equation model with two binary-choice processes,

$$y_{i1}^* = \mathbf{x}_{i1}\boldsymbol{\beta_1} + \lambda y_{i2}^* + u_{i1} \tag{14}$$

$$y_{i2}^* = \mathbf{x}_{i2}\boldsymbol{\beta_2} + \lambda y_{i1}^* + u_{i2} \tag{15}$$

$$y_{il} = \begin{cases} 1 \text{ if } y_{il}^* > 0 \\ 0 \text{ otherwise} \end{cases} \quad \text{for } l = 1, 2, \tag{16}$$

with interdependence parameters $|\lambda| < 1$. We can write this model in the now-familiar matrix form

$$\mathbf{y}_{(2N\times 1)}^* = \mathbf{A}\mathbf{y}^* + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{17}$$

whereas $\mathbf{y}^* = [y_{11}^*,\ \ldots,\ y_{N1}^*,\ y_{12}^*,\ \ldots,\ y_{N2}^*]'$, $\boldsymbol{\beta} = [\boldsymbol{\beta_1},\ \boldsymbol{\beta_2}]'$, and

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{O} \\ \mathbf{O} & \mathbf{X}_2 \end{pmatrix}. \tag{18}$$

The weights matrix is now given as

$$\mathbf{A}_{(2N\times 2N)} = \left(\begin{array}{ccc|ccc} 0 & 0 & 0 & \lambda & 0 & 0 \\ 0 & \ddots & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & \lambda \\ \hline \lambda & 0 & 0 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 & \ddots & 0 \\ 0 & 0 & \lambda & 0 & 0 & 0 \end{array}\right) = \left(\begin{array}{c|c} \mathbf{A}_1 & \mathbf{A}_2 \\ \text{“1st region”} & \text{“2nd region”} \\ \hline \mathbf{A}_3 & \mathbf{A}_4 \\ \text{“3rd region”} & \text{“4th region”} \end{array}\right). \tag{19}$$

The reduced form follows as

$$\mathbf{y}^* = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{A})^{-1}\mathbf{u} \tag{20}$$

Again, it is clear that the structure of this specification is essentially the equal to the one of the binary spatial model, and lead to the same estimation challenge.

2.4 Combining dependencies

So far, we have established that spatial, temporal, and outcome (inter-)dependence all give rise to the same reduced form specification for the latent outcome vector $\mathbf{y}^*$, and are thus all impossible to fit via direct ML estimation. However, the similarity in functional form also means that it is very straightforward to combine different dependency structures, yielding models exhibiting multiple types of dependencies among observations. In the following we give three examples of hybrid models that are potentially useful in applied research.

The first hybrid model we consider is the binary spatio-temporal autoregressive model (STAR), which combines the binary spatial model with the temporal autoregressive binary model, yielding a panel setup (see e.g. Franzese et al. 2016). The binary STAR model is given by

$$\mathbf{y}^*_{(NT\times 1)} = \mathbf{Q}\mathbf{y}^* + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{21}$$

where $\mathbf{y}^* = [y^*_{\cdot 1},\ \ldots,\ y^*_{\cdot T}]'$ and $y^*_{\cdot t} = [y^*_{1t},\ \ldots,\ y^*_{Nt}]'$. Hence, the cross-sectional $y^*_{\cdot t}$ vectors are stacked "on top of each other". The X matrix is constructed analogously. The weights matrix $\mathbf{Q}$ is given by

$$\mathbf{Q}_{NT\times NT} = \rho \begin{pmatrix} \mathbf{W} & 0 & 0 & \cdots & 0 \\ 0 & \mathbf{W} & 0 & \cdots & 0 \\ 0 & 0 & \mathbf{W} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{W} \end{pmatrix} + \gamma \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ I_N & 0 & 0 & \cdots & 0 \\ 0 & I_N & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}, \tag{22}$$

where $\mathbf{W}$ is the $N \times N$ spatial weights matrix and I_N is the $N \times N$ identity matrix.

Another potentially useful model is the binary simultaneous equation panel model,

where two related binary choice processes are repeated over time. This model is given by

$$\mathbf{y}^*_{(2NT\times 1)} = \mathbf{Q}\mathbf{y}^* + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{23}$$

where $\mathbf{y}^* = [y^*_{\cdot 11},\ y^*_{\cdot 21}\ \ldots,\ y^*_{\cdot 1T},\ y^*_{\cdot 2T}]'$ and $y^*_{\cdot lt} = [y^*_{1lt},\ \ldots,\ y^*_{NlT}]'$. Hence, the cross-sectional latent outcomes are first stacked by choice type, and then by time period. Here, the weights matrix is block diagonal:

$$\mathbf{Q}_{2NT\times 2NT} = \begin{pmatrix} \mathbf{A} & 0 & \cdots & 0 \\ 0 & \mathbf{A} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{A} \end{pmatrix}, \tag{24}$$

with $\mathbf{A}$ as defined in the previous section.

Finally, we consider the binary simultaneous equation spatial model, where two related binary outcomes are each spatially interdependent. Here we have

$$\mathbf{y}^*_{(2N\times 1)} = \mathbf{Q}\mathbf{y}^* + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{25}$$

with $\mathbf{y}^*$ defined as in the previous section. The weights matrix now assumes the form

$$\mathbf{Q}_{2N\times 2N} = \left(\begin{array}{ccc|ccc} & & & \lambda & 0 & 0 \\ & \rho_1\mathbf{W} & & 0 & \ddots & 0 \\ & & & 0 & 0 & \lambda \\ \hline \lambda & 0 & 0 & & & \\ 0 & \ddots & 0 & & \rho_2\mathbf{W} & \\ 0 & 0 & \lambda & & & \end{array}\right) \tag{26}$$

whereas $\mathbf{W}$ is the spatial weights matrix.

3 A pseudo maximum likelihood estimator for (inter-)dependent binary outcomes

Having established that direct ML estimation is infeasible for binary models featuring (inter-)dependencies, it is clear that we require an alternative approach. One option is simulation. Franzese et al. (2016) and Calabrese and Elkink (2014b) provide extensive reviews of the spatial probit literature, and useful comparisons of several simulation-based estimation methods like recursive-importance-sampling (RIS) and Bayesian MCMC approaches (see also Calabrese and Elkink (2014a) for cases with asymmetric link functions accommodating rare events). Similarly, Beck et al. (2001) discuss a Bayesian estimation strategy for the binary temporal autoregressive model. The main drawback of simulation-based approaches is that they are computationally intensive, and often require tedious hyperparameter tuning (especially in the MCMC case). For this reason, this section introduces a pseudo maximum likelihood (PML) method as a feasible way to overcome technical challenges associated with estimating binary (inter-)dependence models. The method was originally proposed for binary spatial models by (Smirnov 2010), but we show in this section that it is applicable for all three types of (inter)-dependence.

3.1 PMLE for the binary spatial model

Recall the reduced form for the binary spatial model is given by

$$\begin{aligned} \mathbf{y}^* &= (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{u} \\ &= (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{v}. \end{aligned} \tag{27}$$

Denote the spatial multiplier by $\mathbf{Z}$,

$$\mathbf{Z}_{(N\times N)} = (\mathbf{I} - \rho\mathbf{W})^{-1}, \tag{28}$$

and, by $\mathbf{D}$, an $N \times N$ matrix that contains diagonal elements of $\mathbf{Z}$. All off-diagonal elements of $\mathbf{D}$ are zero. The spatial multiplier indicates the degree of local and global spillovers of an exogenous shock that unit i receives (Anselin 2003); in other words, $z_{ij} = \frac{\partial y_i^*}{\partial u_j}$, where z_{ij} is the ijth element of $\mathbf{Z}$. The diagonal matrix $\mathbf{D}$ indicates "private effects," borrowing Smirnov's (2010) term, of exogenous shocks on the individual latent outcomes. The relative effect captured by $\mathbf{D}$ is "private" in that it indicates the magnitude of the effect that unit i receives from an exogenous shock that occurred to unit i itself; in other words, $d_i = \frac{\partial y_i^*}{\partial u_i}$.

On the other hand, the off-diagonal elements of $\mathbf{Z}$, i.e. $\mathbf{Z} - \mathbf{D}$, represent "aggregate spatial effects" of an exogenous shock. Note that all diagonal elements of $\mathbf{Z} - \mathbf{D}$ are zero. One could interpret it as an aggregate spillover effects that unit i receives from an exogenous shock through all the other units.

The reduced form can now be re-written as

$$\mathbf{y}^*_{(N\times 1)} = \mathbf{ZX}\boldsymbol{\beta} + \underbrace{(\mathbf{Z}-\mathbf{D})\mathbf{u}}_{\text{"Social effects"}} + \underbrace{\mathbf{Du}}_{\text{"Private effects"}}, \tag{29}$$

or, for each unit i,

$$y_i^* = \sum_j \beta z_{ij} x_j + \sum_j [\mathbf{Z}-\mathbf{D}]_{ij} u_j + d_i u_i. \tag{30}$$

We can now rewrite the probability of unit i seeing a positive outcome as

$$\begin{aligned} P(y_i = 1) &= P(y_i^* \geq 0) \\ &= P(\sum_j \beta z_{ij} x_j + \sum_j [\mathbf{Z}-\mathbf{D}]_{ij} u_j + d_i u_i \geq 0) \\ &= P\left(u_{it} \leq \frac{\sum_j \beta z_{ij} x_j}{|d_i|} + \frac{\sum_j [\mathbf{Z}-\mathbf{D}]_{ij} u_j}{|d_i|}\right). \end{aligned} \tag{31}$$

The above expression holds regardless of the sign of d_i, a diagonal element of the spatial multiplier, as long as the distribution for u_i is symmetric.

Note that there is a stochastic element left in the argument of the probability in the

above expression: $\sum_j [\mathbf{Z} - \mathbf{D}]_{ij} u_j$. Because $E(u_i) = 0$ by assumption, we can write

$$E\left[\sum_j [\mathbf{Z} - \mathbf{D}]_{ijt} u_j\right] = 0. \tag{32}$$

This implies that the effects of exogenous shocks that unit i receives via the spatial multiplier component $z_{ij},\ i \neq j$ are not systematic, and have no systematic effect on the choice probability $P(y_i = 1)$. Smirnov's (2010) key proposal is to approximate $\sum_j [\mathbf{Z} - \mathbf{D}]_{ij} u_j$ in (31) by its expectation, i.e., zero. This step massively simplifies the likelihood function. To see why, note that $P(y_{it} = 1)$ can now be written as follows:

$$\begin{aligned} P(y_i = 1) &= P(y_i^* \geq 0) \\ &= P\left(u_i \leq \frac{\sum_j \beta z_{ij} x_j}{|d_i|}\right) \\ &= F_u\left(\frac{\sum_j \beta z_{ij} x_j}{|d_i|}\right), \end{aligned} \tag{33}$$

where $F_u(.)$ is the cdf of the *univariate* distribution of u_i, which is typically the standard normal (yielding a Probit model) or a standard logistic (yielding a Logit model).

With this approximation, we can write the pseudo likelihood in closed form. If u_i follows the standard logistic distribution, for instance, we have

$$\begin{aligned} PL(\rho, \boldsymbol{\beta} | \mathbf{X}, \mathbf{y}^*) &= \left[\prod^N P(y_i = 1)\right]^{y_i} \left[\prod^N (1 - P(y_i = 1))\right]^{(1-y_i)} \\ &\propto \left[\prod^N \frac{\exp\left((\sum_j \beta z_{ij} x_j)/|d_i|\right)}{1 + \exp\left((\sum_j \beta z_{ij} x_j)/|d_i|\right)}\right]^{y_i} \\ &\times \left[\prod^N \frac{1}{1 + \exp\left((\sum_j \beta z_{ij} x_j)/|d_i|\right)}\right]^{(1-y_i)}. \end{aligned} \tag{34}$$

3.2 PMLE for the temporal autoregressive model

Recall the reduced form for the binary temporal autoregressive model, given by

$$\mathbf{y}^*_{(T\times 1)} = (\mathbf{I} - \gamma\mathbf{T})^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \gamma\mathbf{T})^{-1}\mathbf{u}. \tag{35}$$

Next, let

$$\mathbf{Z}_{(T\times T)} = (\mathbf{I} - \gamma\mathbf{T})^{-1}, \tag{36}$$

denote the dependency multiplier. Applying the logic of the previous section, we can decompose the reduced-form error term into two parts

$$\begin{aligned} \mathbf{y}^* &= \mathbf{Z}\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} \\ &= \mathbf{Z}\mathbf{X}\boldsymbol{\beta} + \underbrace{(\mathbf{Z} - \mathbf{D})\mathbf{u}}_{distributed} + \underbrace{\mathbf{D}\mathbf{u}}_{contemporaneous}. \end{aligned} \tag{37}$$

The *distributed* effect captures the effect of exogenous shocks that occurred in the past and were carried over to the outcome of time t. These are distributed because this term focuses on the effect that is carried across multiple time periods ("neighbors" in time). On the other hand, the contemporaneous effects capture the effect of an exogenous shock that occurred in the current time period on the current outcome. Note that due to the lower-diagonal structure of $\mathbf{T}$, $\mathbf{D} = \mathbf{I}$, and thus $d_i = 1$. Again substituting $(\mathbf{Z} - \mathbf{D})\mathbf{u}$ with its expectation, we arrive at the following expression for the probability of a positive outcome:

$$\begin{aligned} &Pr(y_t = 1) \\ &= Pr(y^*_t > 0) \\ &= Pr\left(u_t < [\mathbf{Z}\mathbf{X}\beta]_t\right), \end{aligned} \tag{38}$$

and the pseudo likelihood function, for instance with a logit link function, is given by

$$\begin{aligned} L(\gamma, \boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) &= \left[\prod^{T} P(y_t = 1)\right]^{y_t} \left[\prod^{T} (1 - P(y_t = 1))\right]^{(1-y_t)} \\ &\propto \left[\prod^{T} 1 - \frac{1}{1 + \exp\left(-[\mathbf{ZX}\beta]_t\right)}\right]^{y_t} \\ &\times \left[\prod^{T} \frac{1}{1 + \exp\left(-[\mathbf{ZX}\beta]_t\right)}\right]^{(1-y_t)} . \end{aligned} \tag{39}$$

3.3 PMLE for the simultaneous outcome model

As established earlier, the reduced form for the binary simultaneous outcome model is given by

$$\mathbf{y}^* = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{A})^{-1}\mathbf{u} \tag{40}$$

Applying the same decomposition to the reduced form error as in the previous sections yields

$$\begin{aligned} \mathbf{y}^* &= (\mathbf{I} - \mathbf{A})^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{A})^{-1}\mathbf{u} \\ &= \mathbf{ZX}\boldsymbol{\beta} + \underbrace{(\mathbf{Z} - \mathbf{D})\mathbf{u}}_{\text{plural effects}} + \underbrace{\mathbf{Du}}_{\text{singular effects}} . \end{aligned} \tag{41}$$

The *singular* effect captures the effect of exogenous shocks on outcome type l directly. The *plural* effect captures exogenous shocks that apply to outcome type $\neg l$ and spill over to outcome type l. Given the error decomposition, the pseudo likelihood estimator may be derived in the exact same way as in the previous two sections. The same applies for any of the discussed hybrid models.

4 Speeding up computation

In the previous section, we have derived pseudo likelihood functions for binary (inter-)dependence models that can be evaluated directly, thus permitting a pseudo maximum likelihood (PML) strategy that does not require simulation. However, naive implementations of the proposed

PML estimator may still be prohibitively costly to run. To see why, let us assume that we attempt to fit a model with the following reduced form

$$\mathbf{y}^*_{N\times N} = (\mathbf{I}-\mathbf{Q})^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I}-\mathbf{Q})^{-1}\mathbf{u}, \tag{42}$$

which yields a pseudo likelihood function consisting of N terms of the following form

$$\begin{aligned} P(y_i = 1) &= P(y_i^* \geq 0) \\ &= F_u\left(\frac{[\mathbf{Z}\mathbf{X}\boldsymbol{\beta}]_i}{|d_i|}\right), \end{aligned} \tag{43}$$

with $\mathbf{Z} = (\mathbf{I}-\mathbf{Q})^{-1}$ and $d_i = [\mathbf{Z}]_{ii}$. Perhaps the most straightforward implementation of expression (43) is to invert $\mathbf{I}-\mathbf{Q}$ directly using a decomposition-based solver, then multiplying $\mathbf{Z}$ with $\mathbf{X}\boldsymbol{\beta}$, and dividing by $diag(\mathbf{Z})$. However, this strategy is typically very slow, as most decomposition-based solvers operate with almost cubic complexity in N.

We propose two (mutually compatible) strategies for avoiding the full inversion of $\mathbf{I}-\mathbf{Q}$. The first is trivial, but is worth spelling out nonetheless. If $\mathbf{Q}$ is block diagonal, which will be the case in any panel model *without* a temporal autoregressive component (e.g. the binary simultaneous equation panel model introduced earlier), then its inverse is the block diagonal matrix of block-wise inverses. More formally,

$$\text{if } \mathbf{Q} = \begin{pmatrix} \mathbf{B}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{B}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \mathbf{B}_T \end{pmatrix}, \text{ then } \mathbf{B}^{-1} = \begin{pmatrix} \mathbf{Q}_1^{-1} & 0 & \cdots & 0 \\ 0 & \mathbf{B}_2^{-1} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \mathbf{B}_T^{-1} \end{pmatrix}.$$

The second strategy is useful whenever $\mathbf{Q}$ can be decomposed as $\mathbf{Q} = \alpha\mathbf{M}$, where α is a scalar. Note that this is the case for all three models proposed earlier (spatial, temporal, simultaneous equation), as well as any panel model based on any of these. In this strategy, we completely avoid inverting $\mathbf{I}-\mathbf{Q}$, and instead compute $\mathbf{Z}\mathbf{X}\boldsymbol{\beta}$ and $\mathbf{D} = diag(\mathbf{Z})$ separately.

We compute $\mu = \mathbf{ZX}\boldsymbol{\beta}$ by solving the linear system $(\mathbf{I} - \mathbf{Q})\mu = \mathbf{X}\boldsymbol{\beta}$ for μ. This we can do without inversion by using gradient-based iterative procedures. A key advantage of doing so is that iterative procedures are especially efficient if $\mathbf{Q}$ is sparse, which is usually the case. We propose using the conjugate gradient method if $\mathbf{Q}$ is symmetric, and the slightly slower biconjugate gradient stabilized method otherwise (see Saad 2003). Both these methods have time complexity that is linear in the number of non-zero elements in $\mathbf{Q}$, and are thus far more efficient than directly inverting $\mathbf{I} - \mathbf{Q}$.

To obtain $\mathbf{D}$, we make use of the fact that $\mathbf{Z}$ can be written as a Neumann series (LeSage and Pace 2009, ch. 2)

$$\mathbf{Z} = \mathbf{I} + \sum_{l=1}^{\infty} (\alpha \mathbf{M})^l. \tag{44}$$

Thus, an approximation for $\mathbf{D}$ may be obtained via

$$\mathbf{D} \approx \widetilde{\mathbf{D}} = diag(\mathbf{I}) + \sum_{l=1}^{L} diag\left(\alpha^l \mathbf{M}^l\right). \tag{45}$$

Note that we can precompute the series $\{\mathbf{M}, \mathbf{M^2}, \ldots, \mathbf{M^L}\}$ prior to optimization. Thus, the time complexity of evaluating $\widetilde{\mathbf{D}}$ during optimization is linear in N.

5 Evaluation and comparison of estimation strategies

Next, we evaluate the performance of our proposed pseudo maximum likelihood (PML) estimator for a spatial probit model and a temporal autoregressive probit model. In both cases, we compare the performance of our estimator to that of some well-recognized alternatives.[2]

5.1 Monte Carlo simulation for the spatial probit model

In this section, we present simulation results for a spatial probit data generating process (DGP) as specified in (3), with $u_i \sim N(0, 1)$. Throughout the experiments we set the offset β_0 to 0.5 and β_1 to 1. The covariate vector $\mathbf{x}$ is drawn from the standard normal. The spatial autocorrelation parameter, ρs, is set to 0 (no spatial autocorrelation), 0.25 and

0.5, respectively in across experiments. The spatial weights matrix $\mathbf{W}$ is captures queen-neighborhoods on a square lattice and is row-normalized.

We compare our estimator (PMLE) with three alternatives: (1) A simple GLM probit model where the observed spatial lag $\mathbf{Wy}$ is added as a covariate. (2) The Bayesian spatial probit model proposed by LeSage (2000) and implemented by Wilhelm and Godinho de Matos (2013).[3] (3) The GMM-based spatial probit model by Klier and McMillen (2008).

We repeat the experiments for sample sizes of $N = \{256, 1024, 4096\}$. For each experiment, we record the root mean squared error (RMSE) associated with all three parameters for each model. We also keep track of estimation times, since one of our goals objectives in this paper is to reduce the computational burden associated with existing methods. We repeat each experiment 50 times.

Table 1 reports the RMSE estimates across all experiments. It is evident that our method, the Bayesian model, and the GMM method perform almost identically in terms of RMSE. As expected, the "naive" GLM model tends to incur relatively large errors. Perhaps more interestingly, Figure 1 summarizes the average estimation time for each experiment and model. It is on this metric where the benefits of using the PMLE become evident: While estimation times for the GMM and Bayesian models approach prohibitively high values for high N, the PMLE model is estimated almost instantly. In fact, we had to abort the estimation loop for the Bayesian models for $N = 4096$ because completing the experiments would have taken more than a day.

5.2 Monte Carlo simulation for the temporal autoregressive probit model

Next, we provide simulation results for the temporal autoregressive probit model, as defined in Section 2.2, with $u_i \sim N(0, 1)$. We set up the experiments equivalently to those for the spatial probit model in the previous section. We compare out estimator (PMLE) with two alternatives: (1) A simple GLM probit model that does not account for temporal autocorrelation. (2) A GLM probit model where the *observed* outcome is included as an additional regression

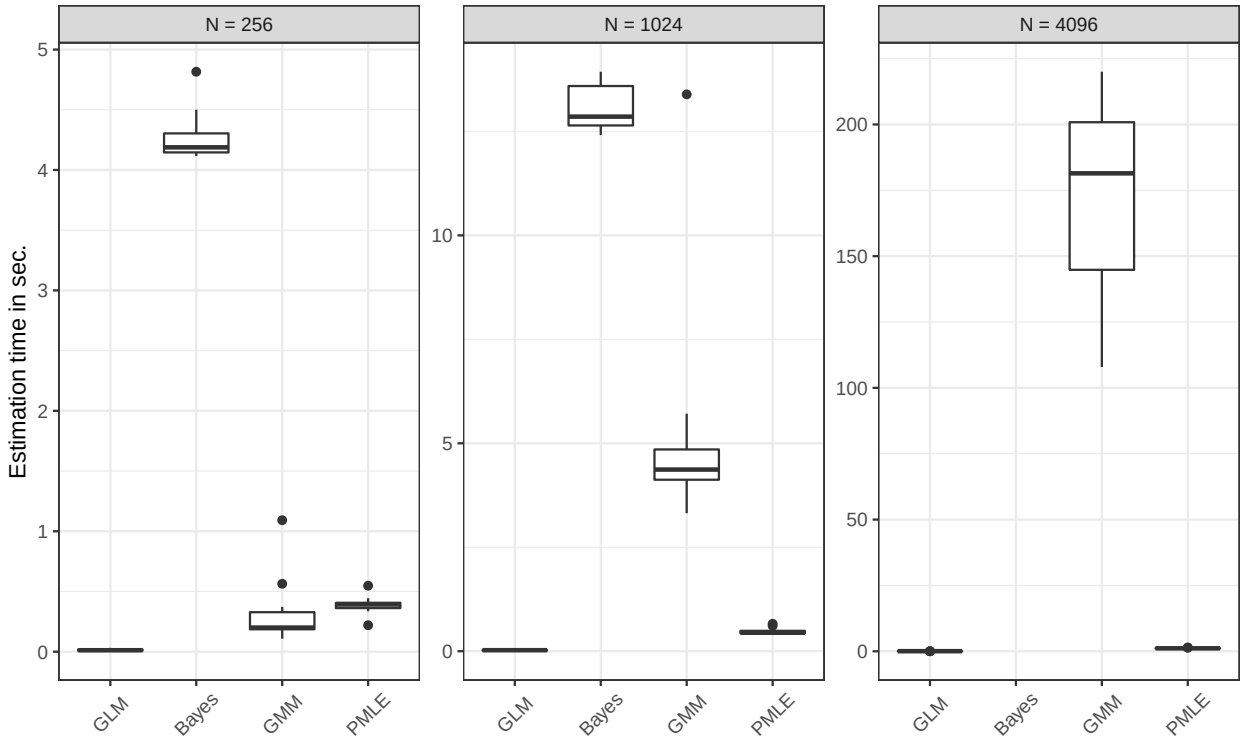

Figure 1: Estimation times (in seconds) for the spatial probit experiments.

Table 2 summarizes the RMSE estimates for the experiments. We find that our estimator exhibits the least error by far – not only for the temporal autocorrelation parameter (γ), but also for β_1. Figure 2 summarizes estimation times. Again, we find that our estimator performs extremely well, typically converging to a solution in under a second, even for $N = 4096$.

6 Conclusion

In this paper, we suggested pseudo maximum likelihood estimators (PMLE) as simpler solutions—particularly from applied the researcher's perspective—to the estimation problems that arise from various sources of interdependence in binary data. The primary source of the estimation challenge is the fact that the jointly determined error terms in the reduced-form specification are analytically intractable due to an N-dimensional integral; hence we cannot obtain an analytical form of the likelihood with respect to the structural-form errors,

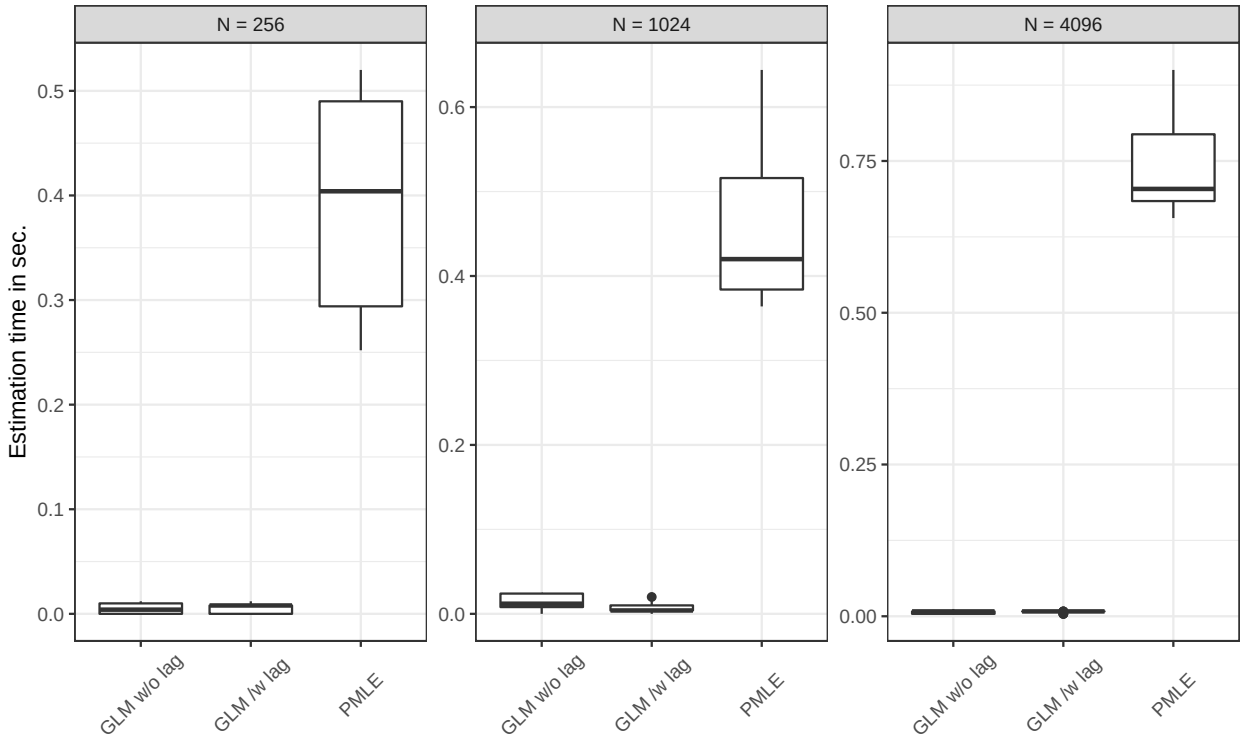

Figure 2: Estimation times (in seconds) for the temporal probit experiments.

for which we know the distributional characteristics. We discussed the problem as generally as possible, considering three sources of interdependence in data: (1) spatial interdependence, (2) temporal autocorrelation, and (3) simultaneity arising from endogenous binary regressors in simultaneous equations. To deal with this problem, the previous literature has proposed simulation approaches, but these are extremely computationally intensive and generally impractical for standard datasets of medium to large size.

As a way forward, we demonstrated how to reduce computational burdens significantly by (i) introducing analytically tractable pseudo maximum likelihood estimators (PMLE) for binary choice models that exhibit (inter-)dependence across space, time and/or outcomes, and by (ii) proposing an implementation strategy that increases computational efficiency considerably. Our first-cut Monte Carlo experiments demonstrate that (a) omitting interdependence induces bias in binary choice models, (b) our estimators are generally able to recover the parameters of the DGP, and (c) our estimators require only a fraction of the

computational cost of simulation-based methods.

That said, there are still at least two major tasks that need to be completed in future iterations of this paper:

- Monte Carlo analyses for the simultaneous equation model and the hybrid models
- An a pplication along the lines sketched in the introduction. This project was originally motivated by a set of complex theory in ethnic conflict; however we would like to illustrate the methods with more self-explanatory topics that incorporates all sources of interdependence.

References

Anselin, L. (2003). Spatial Externalities, Spatial Multipliers, and Spatial Econometrics. *International Regional Science Review*, 26(2):153–166.

Beck, N., Epstein, D., Jackman, S., and O'Halloran, S. (2001). Alternative models of dynamics in binary time-series-cross-section models: The example of state failure.

Calabrese, R. and Elkink, J. A. (2014a). Estimating Binary Spatial Autoregressive Models for Rare Events. *Working paper*.

Calabrese, R. and Elkink, J. A. (2014b). Estimators of binary spatial autoregressive models: A Monte Carlo study. *Journal of Regional Science*, 54(4):664–687.

Franzese, R. J., Hays, J. C., and Cook, S. J. (2016). Spatial-and spatiotemporal-autoregressive probit models of interdependent binary outcomes. *Political Science Research and Methods*, 4(1):151–173.

Kelejian, H. H. and Prucha, I. R. (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Econometrics*, 157(1):53–67.

Klier, T. and McMillen, D. P. (2008). Clustering of auto supplier plants in the united states: generalized method of moments spatial logit for large samples. *Journal of Business & Economic Statistics*, 26(4):460–471.

LeSage, J. and Pace, R. K. (2009). *Introduction to spatial econometrics*. Chapman and Hall/CRC.

LeSage, J. P. (2000). Bayesian estimation of limited dependent variable spatial autoregressive models. *Geographical Analysis*, 32(1):19–35.

Nordhaus, W. D. (1999). Global Public Goods and the Problem of Global Warming. The Institut d'Economie Industrielle (IDEI), Toulouse, France.

Saad, Y. (2003). *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, Philadelphia.

Simmons, B. A. and Elkins, Z. (2004). The globalization of liberalization: Policy diffusion in the international political economy. *American Political Science Review*, 98(1):171–189.

Smirnov, O. A. (2010). Modeling spatial discrete choice. *Regional Science and Urban Economics*, 40(5):292–298.

Von Stein, J. (2005). Do treaties constrain or screen? selection bias and treaty compliance. *American Political Science Review*, 99(04):611–622.

Weidner, H. and Mez, L. (2008). German climate change policy a success story with some flaws. *The Journal of Environment and Development*, 17(4):356–378.

Wilhelm, S. and Godinho de Matos, M. (2013). Estimating Spatial Probit Models in R. *The R Journal*, 5(1):130–143.

Notes

[1]The following results generalize trivially to higher-order processes.

[2]At the moment, we do not yet have simulation results for the simultaneous outcome model; they are forthcoming.

[3]We run the MCMC algorithm with default settings for 1000 iterations.

	N = 256				N = 1024				N = 4096			
True value	GLM	Bayes	GMM	PMLE	GLM	Bayes	GMM	PMLE	GLM	Bayes	GMM	PMLE
$\beta_0 = $ -0.5	0.2693	0.1554	0.1315	**0.1290**	0.1572	**0.0685**	0.0788	0.0795	**0.0441**		0.0477	0.0447
$\beta_1 = 1$	**0.1202**	0.1318	0.1210	0.1238	**0.0404**	0.0421	0.0416	0.0414	0.0369		**0.0364**	0.0366
$\rho = 0$	0.8527	**0.1836**	0.2015	0.1871	0.4050	**0.0805**	0.1350	0.1365	**0.0783**		0.0976	0.0923
$\beta_0 = $ -0.5	0.5940	0.1116	0.0842	**0.0739**	0.4423	**0.0266**	0.0350	0.0420	0.3934		**0.0493**	0.0537
$\beta_1 = 1$	0.1482	0.1800	0.1527	**0.1374**	0.0524	**0.0522**	0.0555	0.0533	0.0257		**0.0240**	0.0265
$\rho = 0.25$	1.1605	0.1718	0.1591	**0.1340**	0.6211	**0.0629**	0.0643	0.0703	0.5086		**0.0690**	0.0774
$\beta_0 = $ -0.5	0.8076	0.2703	**0.1631**	0.2010	0.9802	**0.0637**	0.1196	0.0946	0.9618		0.0398	**0.0366**
$\beta_1 = 1$	0.1683	**0.1563**	0.1639	0.1695	**0.0583**	0.0647	0.0797	0.0764	0.0470		**0.0306**	0.0477
$\rho = 0.5$	0.7214	0.2213	**0.1153**	0.1609	1.5294	**0.0588**	0.1258	0.0953	1.4694		0.0542	**0.0541**

Table 1: RMSE estimates for the spatial model. Best-performing estimator in bold.

	N = 256			N = 1024			N = 4096		
True value	GLM w/o lag	GLM /w lag	PMLE	GLM w/o lag	GLM /w lag	PMLE	GLM w/o lag	GLM /w lag	PMLE
$\beta_0 = -0.5$	**0.1450**	0.4995	0.1907	0.0483	0.4368	**0.0392**	0.0222	0.5192	**0.0220**
$\beta_1 = 1$	**0.1571**	1.5724	0.1577	**0.0563**	1.4917	0.0570	**0.0358**	1.5189	0.0362
$\gamma = 0$		1.0782	**0.1034**		1.0320	**0.0437**		1.0212	**0.0203**
$\beta_0 = -0.5$	0.1651	0.9845	**0.0692**	0.1047	1.0295	**0.0539**	0.1229	1.0556	**0.0214**
$\beta_1 = 1$	0.1316	1.8186	**0.1118**	0.1003	1.7958	**0.0780**	0.0497	1.8252	**0.0324**
$\gamma = 0.25$		0.6667	**0.1340**		0.6826	**0.0637**		0.7408	**0.0215**
$\beta_0 = -0.5$	0.4427	1.6373	**0.1025**	0.2670	1.6569	**0.0965**	0.2915	1.6714	**0.0639**
$\beta_1 = 1$	0.2506	2.2781	**0.2223**	0.2234	2.1772	**0.1432**	0.2240	2.2037	**0.1384**
$\gamma = 0.5$		0.3933	**0.0521**		0.4064	**0.0432**		0.3953	**0.0122**

Table 2: RMSE estimates for the temporal model. Best-performing estimator in bold.