A Fast Estimator for Binary Choice Models with Spatial, Temporal, and Spatio-Temporal Interdependence

Julian Wucherpfennig[a], Aya Kachi[b,*], Nils-Christian Bormann[c], Philipp Hunziker[d]

[a]*Centre for International Security, Hertie School*
[b]*Faculty of Business and Economics, University of Basel*
[c]*Department of Philosophy, Politics & Economics, Witten/Herdecke University*
[d]*Network Science Institute, Northeastern University (Now at Google)*

Abstract

Binary outcome models are frequently used in the social sciences and economics. However, such models are difficult to estimate with interdependent data structures, including spatial, temporal, and spatio-temporal autocorrelation because jointly determined error terms in the reduced-form specification are generally analytically intractable. To deal with this problem, simulation-based approaches have been proposed. However, these approaches (i) are computationally intensive and impractical for sizable datasets commonly used in contemporary research, and (ii) rarely address temporal interdependence. As a way forward, we demonstrate how to reduce the computational burden significantly by (i) introducing analytically-tractable pseudo maximum likelihood estimators (PMLE) for latent binary choice models that exhibit interdependence across space *and* time and by (ii) proposing an implementation strategy that increases computational efficiency considerably. Monte Carlo experiments show that our estimators recover the parameter values as good as commonly-used estimation alternatives and require only a fraction of the computational cost.

1. Introduction

Modeling binary outcomes—such as war, regime transitions, or policy adaption—poses considerable methodological challenges in the presence of spatial and/or temporal autocorrelation resulting from interdependent outcomes across and within units. The methodological difficulty stems from the

☆This is a preprint version (October 2020). We would like to thank Frederick Boehmke, Lars-Erik Cederman, Scott Cook, Olga V. Chyzh, Martin Elff, Jos Elkink, Robert J. Franzese Jr., Christian Kleiber, Dennis Quinn, Andrea Ruggeri, Emily Schilling, Martin Steinwand, Oliver Westerwinter, Michael Ward, Christopher Zorn as well as audiences at various conferences for their useful comments.

*Corresponding author: Aya Kachi (aya.kachi@unibas.ch).

likelihood function that involves an analytically intractable NT-dimensional integral.[1] Simulation-based estimation strategies including Gibbs sampling [14] and recursive importance sampling (RIS) [4] have been proposed to overcome this challenge. While these techniques promise to provide reliable estimates of spatial, and more recently spatio-temporal interdependence [9], they are computationally burdensome [see 6].[2] As social scientists implement research designs at increasing resolutions (e.g., at the grid-cell level) and with increasingly large datasets, simulation-based approaches quickly become infeasible.

We provide a new estimator to address spatial, temporal, and spatio-temporal forms of interdependence embedded in binary outcome data. We build on a pseudo maximum likelihood estimator (PMLE) for binary spatially autoregressive models proposed by Smirnov [15], and extend it to cases of temporal and spatio-temporal interdependence. So far, only Franzese et al.'s (2016) RIS estimator has addressed the spatio-temporal case. In addition, we reduce the estimation costs by proposing an implementation strategy that avoids direct matrix inversion (of large "*interdependence multipliers*"), and instead relies on a combination of iterative gradient procedures and approximations that yield an estimation algorithm with almost (only) linear complexity in N.

Monte Carlo experiments demonstrate that the PMLE recovers the parameter values as good as commonly-used estimation alternatives—including Bayes, GMM, RIS, and naïve probit—in a fraction of the time that simulation-based methods require. Yet, our analyses also accentuate important methodological issues that await solutions in improving the two existing estimators for spatio-temporal models (RIS and PMLE). First, both estimators generate seemingly biased standard errors. Second, we find that the performance of the RIS estimator is sensitive to the choice of data generating process (DGP). The conclusion section elaborates on these points.

2. Binary choice models with spatio-temporal interdependence

This section specifies a binary choice model with spatio-temporal interdependence, for which we then develop a pseudo maximum likelihood estimators. We focus on the spatio-temporal case, which is applicable to cross-sectional-time-series data, noting that a purely spatial model for cross-

[1]Spatial probit is a special case where the marginal probability has a closed form but the likelihood function requires an evaluation of a multivariate normal distribution, which again cannot be computed exactly [4].

[2]Franzese et al. [9] and Calabrese and Elkink [6] provide extensive reviews of the spatial probit literature, and useful comparisons among simulation-based estimation methods. Similarly, [3] discuss a Bayesian estimation strategy for the binary temporal autoregressive model.

sectional data and a purely temporal model for time-series data are nested herein. Full derivations are given in the Online Appendix.

Our analytical point of departure is a discrete-choice spatio-temporal autoregressive (STAR) model as the conventional latent variable formulation:

$$y_{it}^* = \rho \sum_{j=1}^{N} w_{ij,t} y_{jt}^* + \gamma y_{i,t-1}^* + \mathbf{x}_{it}\boldsymbol{\beta} + u_{it} \tag{1}$$

or in matrix form:

$$\mathbf{y}_{(N\times T)}^* = \rho \mathbf{W}\mathbf{y}^* + \gamma \mathbf{T}\mathbf{y}^* + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}. \tag{2}$$

Here $\mathbf{y}^*$ is our latent outcome variable for which we observe realizations $\mathbf{y}$, such that $y_{it} = 1$ if $y_{it}^* > 0$ and $y_{it} = 0$ otherwise. The spatial connectivity matrix $\mathbf{W}$ captures dependency between units across space,[3] $\mathbf{T}$ is a temporal connectivity matrix that links the unit's current outcomes to past realizations thereof, $\mathbf{X}\boldsymbol{\beta}$ is a vector of covariates with corresponding parameters, and $\mathbf{u}$ is an error term (zero-mean, iid, on the individual unit level). More specifically, $\mathbf{W}_{NT\times NT}$ is block-diagonal with blocks of $\mathbf{W}_{N\times N}^*$ for each time period t, while $\mathbf{T}$ is an $NT \times NT$ matrix full of zeros except for the identity matrices of size N on the lower first-minor (block) diagonal. The reduced form is:

$$\mathbf{y}^* = (\mathbf{I} - \rho\mathbf{W} - \gamma\mathbf{T})^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \rho\mathbf{W} - \gamma\mathbf{T})^{-1}\mathbf{u}. \tag{3}$$

Deriving the likelihood function requires the computation of $P(y|\boldsymbol{\beta}, \rho, \gamma, \mathbf{X})$, the joint probability for the observed random variable Y given the model parameters and regressors, which then requires the marginal CDF of the reduced-form error term, $(\mathbf{I} - \rho\mathbf{W} - \gamma\mathbf{T})^{-1}\mathbf{u}$. This computation is analytically intractable (as long as $\rho \neq 0$) due to the *interdependence multiplier*, $(\mathbf{I} - \rho\mathbf{W} - \gamma\mathbf{T})^{-1}$ [1].[4]

3. A PMLE for binary spatio-temporal autoregressive (STAR) models

To circumvent this problem, we turn to a pseudo maximum likelihood method. We build on Smirnov's (2010) spatial PMLE and extend it to cases of temporal and spatio-temporal interdependence. The remainder of the section illustrates the gist of this derivation. Here, we maintain general mathematical expressions without assuming a specific marginal distribution

[3]As is convention, we row-standardize $\mathbf{W}$ to ensure stationarity for $|\rho| < 1$ [11].

[4]Spatial probit is a special case where the marginal probability has a closed form but the likelihood function requires an evaluation of a multivariate normal distribution, which again cannot be computed exactly [4].

(e.g., logistic vs. normal). In fact, the PMLE's feasibility regardless of the error-term distribution is one of the strengths of this approach.[5]

Let $\mathbf{Z}_{NT\times NT} = (\mathbf{I} - \rho\mathbf{W} - \gamma\mathbf{T})^{-1}$ and define $\mathbf{D}$ as a corresponding diagonal matrix that contains only the diagonal elements of $\mathbf{Z}$, with all off-diagonal elements being zeros. This allows us to rewrite the model reduced form as:

$$\mathbf{y}^*_{(NT\times 1)} = \mathbf{ZX}\boldsymbol{\beta} + \underbrace{(\mathbf{Z}-\mathbf{D})\mathbf{u}}_{\text{higher-order effects}} + \underbrace{\mathbf{Du}}_{\text{zero-order effects}}. \tag{4}$$

Decomposing the spatial multiplier this way allows us to distinguish between zero-order and higher-order effects of an external shock as it affects observation *it*. More specifically, zero-order effects capture those shocks that the unit experiences directly. Higher-order effects are spillovers of external shocks which are transmitted either spatially through other units', or temporally across multiple time-periods. Our decomposition aggregates across both dimensions.

In order to allow for an analytical formulation of a (pseudo) likelihood, we assume that higher-order effects can be "ignored". Behaviorally, this means that observations may simplify their choice by neglecting aggregate spatial effects of a random shock that are experienced by other (connected) observations. That is, mathematically, we do not expect a systematic effect of a random shock on unit i that is carried through the off-diagonal elements of the spatial multiplier; i.e., it does not affect the choice probability systematically. The assumption is warranted because u_{it} are i.i.d with mean 0. This dramatically simplifies the stochastic element of the choice probability:

$$P(y_{it}=1) = P(y^*_{it}\geq 0) = F_u\left(\frac{\sum_s\sum_j \beta z_{ijs}x_{js}}{d_{it}}\right), \tag{5}$$

which is now the cdf of the *univariate* distribution of u_{it} ($F_u(.)$), such as the standard normal (Probit) or a standard logistic (Logit). This allows us to write down a (pseudo) likelihood function, which is now in closed form. For any binomial link function $g(\cdot)$, we have

$$PL(\rho,\gamma,\boldsymbol{\beta}|\mathbf{X},\mathbf{y}) \propto \prod_{i=1}^{N}\prod_{t=1}^{T}\left[g^{-1}\left(\frac{[\mathbf{ZX}\beta]_{ij,s}}{d_{it}}\right)^{y_{it}}\left[1-g^{-1}\left(\frac{[\mathbf{ZX}\beta]_{ij,s}}{d_{it}}\right)\right]^{(1-y_{it})}\right]. \tag{6}$$

Note that this expression requires an estimate for the values for $\mathbf{y}^*_{i0}$, i.e. the values *preceding* the first observed period in order to calculate the first period

[5]In our view, this strength goes beyond these distributions. This is useful when one might develop an estimator, for instance, for a hybrid of a binary spatial model and another model from a different model class such as duration and count.

$\mathbf{y}^*_{i1}$ (see Equation 1).[6] Assuming mean stationarity, we draw on Kauppi and Saikkonen [10] and use what can be viewed as the unconditional expectation of $\mathbf{y}^*$ across all time period (and units): $E[\mathbf{y}^*] = (I - \rho\mathbf{W} - \gamma)^{-1}\bar{\mathbf{X}}\beta$, where $\bar{\mathbf{X}}$ are the sample means.

4. Speeding up computation further

Naive implementations of the proposed PMLE may still be costly to run. 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. (Appendix Appendix C.1 clarifies why and how.)

5. Monte Carlo Simulations

5.1. Main results: Estimator comparisons for the binary spatio-temporal model

We ran MCs for spatial, temporal, and spatio-temporal PMLE, comparing estimates to those of alternative estimators. Here our discussion focuses on spatio-temporal estimators. Other results are presented in the Online Appendix. The DGP is a spatio-temporal probit as specified in (1), with $u_i \sim N(0, 1)$. We ran MCs for probit because most other existing tools are for probit models.[7] Throughout the experiments we set β_0 to -0.5 and β_1 to 1. The covariate vector $\mathbf{x}$ is drawn from the standard normal and the spatial weights matrix $\mathbf{W}$ captures queen-neighborhoods on a square lattice (row-standardized). We repeat the experiments for sample sizes of $N \times T = \{64 \times 4;\ 64 \times 16;\ 256 \times 16\}$ with three combinations of spatial and temporal autocorrelation: $\{\rho = .25\ \&\ \gamma = .25;\ \rho = .25,\ \&\ \gamma = .5;\ \rho = .5\ \&\ \gamma = .25\}$.

Table E.5 summarizes the results. Overall, the PMLE recovers the parameters accurately and with reasonable precision. However, the PMLE is overconfident in some cases, suggesting that the standard errors (obtained via the Hessian) are too small. Less than one percent of runs do not converge (see Tables E.8 and E.9 in the Online Appendix).

We contend that our estimator is nevertheless useful for applied researchers. Unlike most existing spatial estimators, it can simultaneously account for both spatial and temporal autocorrelation. So far, only the RIS approach proposed by Franzese et al. [9] is able to estimate spatio-temporal processes for binary data. However, our PMLE is several orders of magnitude faster: on a standard PC a single run for $N \times T = 64 \times 16$ takes seven

[6] Because $\mathbf{y}^*$ is latent, dropping the first period from the likelihood merely shifts the problem to the next period, rather than solving it [cf. 9].

[7] As is in equation (B.17), our PMLEs can host other link functions if one wishes.

Table 1: Simulation results for **spatio-temporal** PMLE (500 iterations; y_0^* is estimated by $E(y_t^*)$)

	N=64 & T=4				N=64 & T=16				N=256 & T=16			
	$\beta_0 = -0.5$	$\beta_1 = 1$	γ	ρ	$\beta_0 = -0.5$	$\beta_1 = 1$	γ	ρ	$\beta_0 = -0.5$	$\beta_1 = 1$	γ	ρ
Experiment #1: $\gamma = 0.25$, $\rho = 0.25$												
Mean Coefficient Estimate	-0.523	1.001	0.249	0.229	-0.484	0.970	0.251	0.249	-0.483	0.964	0.249	0.249
Mean Bias	0.171	0.122	0.080	0.154	0.083	0.057	0.036	0.080	0.044	0.039	0.019	0.042
RMSE	0.237	0.155	0.100	0.202	0.105	0.071	0.046	0.099	0.056	0.047	0.024	0.053
Actual SD of estimates	0.236	0.156	0.100	0.201	0.104	0.064	0.046	0.099	0.053	0.031	0.024	0.053
Overconfidence	1.302	1.104	1.086	1.203	1.192	0.971	1.011	1.166	1.161	0.952	1.027	1.167
Experiment #2: $\gamma = 0.25$, $\rho = 0.5$												
Mean Coefficient Estimate	-0.614	0.997	0.233	0.459	-0.469	0.930	0.250	0.497	-0.464	0.925	0.247	0.501
Mean Bias	0.318	0.162	0.102	0.160	0.136	0.093	0.046	0.075	0.071	0.077	0.023	0.037
RMSE	0.580	0.206	0.141	0.231	0.176	0.113	0.059	0.099	0.087	0.086	0.029	0.047
Actual SD of estimates	0.570	0.206	0.140	0.228	0.174	0.089	0.059	0.099	0.079	0.042	0.029	0.047
Overconfidence	2.027	1.092	1.371	1.474	1.548	1.067	1.210	1.394	1.392	1.011	1.137	1.283
Experiment #3: $\gamma = 0.5$, $\rho = 0.25$												
Mean Coefficient Estimate	-0.530	0.937	0.493	0.224	-0.451	0.876	0.496	0.245	-0.438	0.860	0.499	0.246
Mean Bias	0.229	0.169	0.079	0.128	0.143	0.130	0.042	0.079	0.086	0.140	0.022	0.039
RMSE	0.331	0.210	0.104	0.176	0.179	0.149	0.053	0.101	0.103	0.146	0.027	0.049
Actual SD of estimates	0.330	0.200	0.103	0.174	0.173	0.083	0.053	0.101	0.082	0.041	0.027	0.049
Overconfidence	1.464	1.069	1.125	1.354	1.514	1.054	1.142	1.393	1.458	1.058	1.141	1.338

Overconfidence is the standard deviation of the estimated parameter divided by the mean of its estimated standard error.

seconds for the PMLE, and nearly three hours for the RIS (see Figure E.4 for a summary of estimation times); for $N \times T = 256 \times 16$ the PMLE takes nine seconds, while we estimate the RIS to take almost two weeks if estimation time increases linearly (not executed).

More concernedly, our experiments show that the RIS's unbiasedness hinges on the choice of DGP, especially as true values of γ and ρ increase [c.f. 7, 166]. For instance, the biases are not prominent under the DGPs chosen in Franzese et al. [9], Calabrese and Elkink [6] but they are under the DGP we selected. (see Figure 1.)[8]

5.2. *Other results: Estimator comparisons for the binary spatial* or *temporal model*

For a purely spatial DGP, we compared the Bayesian spatial probit model proposed by LeSage [14] and implemented by Wilhelm and Godinho de Matos [19], a linearized spatial GMM [12], RIS [9], a naive probit with an observed spatial lag, and our PMLE. Our experiments suggest that a Bayesian approach is preferable (see Tables E.2, E.6, Figure E.2, and Calabrese and Elkink [6]). However, given its speed, PMLE potentially provides useful starting values for the Bayesian approach. Finally, we also examined a setup with just temporal autocorrelation, comparing the RIS and PMLE approach. Again the PMLE outperforms the RIS (see Tables E.3, E.7 and Figure E.3).

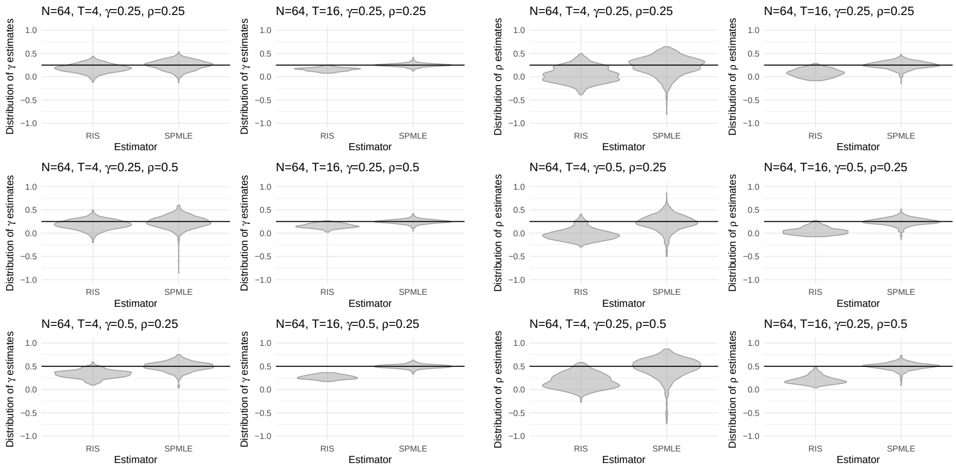
6. Conclusion

In this paper, we (i) introduced an analytically tractable pseudo maximum likelihood estimator for binary choice models that exhibit interdependence across space and/or time and (ii) proposed an implementation strategy that increases computational efficiency considerably. Our Monte Carlo experiments demonstrate that the estimators are able to recover the parameters of the DGP, and requires only a fraction of the computational cost of simulation-based methods. For spatio-temporal and temporal models, the PMLE estimator outperforms the only available alternative, the RIS implementation by [9]. By contrast, for purely spatial applications the Bayesian approach by Beron and Vijverberg [4] appears to perform best.

However, our PMLE approach comes with one drawback: its standard errors are potentially biased (but apparently less so than those given by the

[8] We were able to retrieve virtually identical RIS estimates to those presented in Franzese et al. [9] when using their DGP and contiguity matrix (see Table E.10). Moreover, this bias is also present in our purely spatial experiments, for which several other estimators are able to retrieve far better, if not unbiased, estimates. This provides confidence that our R implementation of Franzese et al.'s (2016) Matlab code (which is a direct translation) is correct.

Figure 1: Distribution of γ and ρ estimates from Monte Carlo simulations for Recursive Importance Sampler and Pseudo-Maximum-Likelihood estimator.



(a) Plots for γ estimates

(b) Plots for ρ estimates

RIS). In the broader context of composite maximum likelihood approaches, Varin et al. [18] provide an extensive review of this property. In short, the standard errors obtained via the Hessian of the PMLE tend to be underestimated for certain parameters. Generally speaking, the literature finds that the bias is greater when NT is not sufficiently large compared to the variables included in the model. Our own first-cut Monte Carlo simulations (only with a single configuration of parameter values) indicate that this bias can emerge especially in the standard error estimate for the spatial parameter ρ. One approach would be a parametric bootstrap. Another approach would be an approximation, such as the use of a sandwich estimator. In this realm, for PMLEs in particular, a sandwich estimator using the Godambe information matrix appears promising [18].

Appendix A. Binary choice models with spatial, temporal, and spatio-temporal interdependence (Full description)

This section specifies and derives a mathematical expression of binary choice models, for which we develop a pseudo maximum likelihood estimator later. We do so for a model with spatial, temporal, and spatio-temporal interdependence, respectively. Note that, in this specification, we try to maintain general mathematical expressions without assuming a specific marginal distribution (such as logistic vs. normal). In fact, the PMLE's estimation feasibility *regardless* of the error-term probability distribution is one of the strengths of this approach. In our view, this strength goes beyond the probit-vs.-logit consideration. This can become useful when one might need to develop an estimator, for instance, for a hybrid of a binary spatial model and another model from a different model class such as duration and count.

Appendix A.1. Spatial interdependence

We consider the following model

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

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

where 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.), ρ is the spatial autocorrelation parameter, $\mathbf{x}_i$ 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 autoregressive model (or binary spatial model as we sometimes mention interchangeably). 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. [9], implying 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

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

where

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

$\mathbf{W}$ 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 (A.3) is stationary as long as $|\rho| < 1$ [11]. Given (A.3) we can derive the reduced form as

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

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

The main component of the (pseudo) likelihood function of our interest will be the joint probability for the observed random variable Y given the model parameters and regressors. This leads to the following expression:

$$\begin{aligned} P(y = 1) &= P(y_i^* > 0) \\ &= P\Big(\Big[(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}\Big]_i + v_i > 0\Big) \\ &= P\Big(v_i > -\Big[(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}\Big]_i\Big) \\ &= 1 - P\Big(v_i \leq -\Big[(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}\Big]_i\Big) \\ &= 1 - F_{V_i}\Big(-\Big[(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}\Big]_i\Big). \end{aligned} \tag{A.6}$$

where $[\cdot]_i$ indicates the i'th element of the vector $[\cdot]$. $F_{V_i}(\cdot)$ is the marginal CDF of the 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 $-\Big[(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}\Big]_i$. By definition the marginal CDF of V_i is

$$\begin{aligned} &F_{V_i}(-[(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}]_i) \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{-\infty}^{-[(\mathbf{I}-\rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}]_i} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{\mathbf{V}}(s_1, \cdots, s_i, \cdots, s_N) ds_1 \cdots ds_i \cdots ds_N, \end{aligned} \tag{A.7}$$

where $f_{\mathbf{V}}(s_1, \cdots, s_N)$ is the joint PDF of the reduced-form error. The estimation challenge for binary choice models arises when evaluating F_{V_i} at $-\Big[(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}\Big]_i$ is analytically intractable (as long as $\rho \neq 0$) [1]. As a consequence, direct maximum likelihood estimation of β and ρ is not always feasible. Of course, one common exception is spatial probit, where the marginal probability has a closed-form expression.

Using this expression for the choice probability, $P(y = 1)$, we have the following expression that is proportional to the (pseudo) likelihood function for a binary spatial autoregressive model.

$$
\begin{aligned}
L(\rho, \boldsymbol{\beta} | \mathbf{X}, \mathbf{y}) &= \Big[\prod_{i=1}^{N} P(y_i = 1)^{y_i} \Big] \Big[\prod_{i=1}^{N} P(y_i = 0)^{(1-y_i)} \Big] \\
&= \Big[\prod_{i=1}^{N} P(y_i = 1)^{y_i} \Big] \Big[\prod_{i=1}^{N} \Big(1 - P(y_i = 1) \Big)^{(1-y_i)} \Big].
\end{aligned} \tag{A.8}
$$

Appendix A.2. Temporal dependence

As an intermediate step toward the binary spatio-temporal model—for which our proposed estimator would eventually be useful—we first illustrate a binary temporal autoregressive model, where the latent outcome exhibits a first-order temporal autoregressive process governed by the temporal autocorrelation parameter γ with $|\gamma| < 1$. The structural form error term u_t is a zero-mean iid error term with fixed variance.[9]

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

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

As it falls out of the main contribution of this paper, we are grossly skipping over the rich time-series methods literature here and we are aware of it. For a discussion of this class of models in a political science context, see [3], for example.

Next, note that we can rewrite the model in matrix notation as follows (equation(A.11)). One might argue that matrix notation of a time-series model is not the most useful expression in terms of estimating model parameters; and yet, as a stepping stone toward the binary spatio-temporal model, it is an analytically appealing expression.

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

where $\mathbf{T}$, called the *temporal weights matrix*, is defined as

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

It is evident that this model is mathematically comparable to the binary spatial model, the sole difference being that now we impose a weights matrix

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

where the first subdiagonal (all the 1's) maps y^*_{t-1} to y^*_t. The reduced form of the autoregressive model is given by

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

As one might see it already, this gives rise to a similar difficulties in ML estimation as the binary spatial model described above.

Appendix A.3. Spatio-temporal interdependence

So far, we have illustrated that spatial and temporal interdependence give rise to the same reduced form expression for the latent outcome vector $\mathbf{y}^*$, and are thus all subject to the same estimation challenge whenever the joint probability $P(y = 1)$ does not have a closed-form expression. This similarity in functional form allows us to combine different dependency structures relatively straightforwardly, yielding models exhibiting multiple types of dependencies among observations. In the following, we consider the binary spatio-temporal autoregressive model (STAR), which combines the binary spatial autoregressive model with the temporal autoregressive model, yielding a panel setup [see e.g. 9]. The binary STAR model is given by

$$\mathbf{y}^*_{(NT\times 1)} = \mathbf{Q}\mathbf{y}^* + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{A.14}$$

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

$$\mathbf{Q}_{NT\times NT} = \rho\mathbf{W}^* + \gamma\mathbf{T}^*, \tag{A.15}$$

where $\mathbf{W}^*$ is the block-diagonal *panel spatial weights matrix* given by

$$\mathbf{W}^* = \begin{pmatrix} \mathbf{W} & 0 & 0 & \cdots & 0 \\ 0 & \mathbf{W} & 0 & \cdots & 0 \\ 0 & 0 & \mathbf{W} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \mathbf{W} \end{pmatrix}, \tag{A.16}$$

and $\mathbf{T}^*$ is the *panel temporal weights matrix* given by

$$\mathbf{T}^* = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ I_N & 0 & 0 & \cdots & 0 \\ 0 & I_N & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix}, \tag{A.17}$$

where I_N is the $N \times N$ identity matrix.

The reduced form of the spatio-temporal autoregressive model is given by

$$\mathbf{y}^*_{(NT\times 1)} = (\mathbf{I} - \mathbf{Q})^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \mathbf{Q})^{-1}\mathbf{u}, \tag{A.18}$$

which again gives rise to the familiar complication.

Appendix B. A pseudo maximum likelihood estimator for interdependent binary outcomes (Full description)

This section describes the PMLE estimator to tackle spatial, temporal, and spatio-temporal forms of interdependence for binary outcome data. Our analytical point of departure is a pseudo maximum likelihood estimator (PMLE) for binary spatially autoregressive models described in Smirnov [15], for which the remaining computational burden amounts to inverting an N-dimensional matrix we refer to as the "interdependence multiplier." We extend the PMLE to cases of temporal and spatio-temporal interdependence, which is a tool that is so far only offered by the Franzese' et al.'s RIS estimator [2016]. 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. This additional procedure we propose will be detailed separately in the following section.

When direct ML estimation is infeasible for binary models featuring interdependence of the outcome variables (due to the lack of a closed-form cdf that goes into $P(y = 1)$), it is clear that we require an alternative approach. One option is simulation. Franzese et al. [9] and Calabrese and Elkink [6] provide extensive reviews of the spatial probit literature, and useful comparisons of several simulation-based estimation methods such as recursive-importance-sampling (RIS) and Bayesian MCMC approaches (see also Calabrese and Elkink [5] for cases with asymmetric link functions accommodating rare events). Similarly, [3] discuss a Bayesian estimation strategy for the binary temporal autoregressive model. However, simulation-based approaches place a number of burdens on the researchers. First, they are computationally intensive and it usually takes a long time to estimate them. Estimation time can be prohibitive if researchers work with big data and do not have access to high-performance computing clusters. Second, convergence problems in MCMC simulations often require tedious hyperparameter tuning and exacerbate the estimation-time problem. Third, and as perhaps the most broadly relevant point, currently, applied researchers do not have access to more than the most basic tools, for example, cross-sectional spatial probit estimators. For these reasons, we now introduce a pseudo maximum likelihood (PML) method as a feasible way to reduce estimation time, minimize convergence problems, and enable applied researchers to run models that more clearly address their research problems. Our estimator builds on Smirnov's [2010] spatial PML estimator and extends it to temporal and spatio-temporal interdependence.

Appendix B.1. PMLE for the binary spatial model

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

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

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

$$\mathbf{Z}_{(N\times N)} = (\mathbf{I} - \rho\mathbf{W})^{-1}, \tag{B.2}$$

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 [2]; 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

$$\mathbf{y}^*_{(N\times 1)} = \mathbf{Z}\mathbf{X}\boldsymbol{\beta} + \underbrace{(\mathbf{Z} - \mathbf{D})\mathbf{u}}_{\text{"Social effects"}} + \underbrace{\mathbf{D}\mathbf{u}}_{\text{"Private effects"}}, \tag{B.3}$$

or, for each unit i,

$$y_i^* = \sum_j \beta z_{ij} x_j + \sum_j [\mathbf{Z} - \mathbf{D}]_{ij} u_j + d_i u_i. \tag{B.4}$$

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

$$\begin{aligned} P(y_i = 1) &= P(y_i^* \geq 0) \\ &= P(\sum_j \beta z_{ij} x_j + \sum_j [\mathbf{Z} - \mathbf{D}]_{ij} u_j + d_i u_i \geq 0) \\ &= P\left(u_i \leq \frac{\sum_j \beta z_{ij} x_j}{d_i} + \frac{\sum_j [\mathbf{Z} - \mathbf{D}]_{ij} u_j}{d_i}\right). \end{aligned} \tag{B.5}$$

Note that there is a stochastic element left in the argument of the probability in the above expression: $\sum_j [\mathbf{Z} - \mathbf{D}]_{ij} u_j$. In order to allow for an analytical formulation of a (pseudo) likelihood, we assume that higher-order effects can be "ignored". Behaviorally, this means that observations may simplify their choice by not worrying about aggregate spatial effects of a random shock that are experienced by other (connected) observations. That is, mathematically, we do not expect a systematic effect of a random shock on unit i that is carried through the off-diagonal elements of the spatial multiplier; i.e., it does not affect the choice probability systematically. The assumption is warranted because u_{it} are i.i.d with mean 0. Smirnov's (2010) key proposal is to approximate $\sum_j [\mathbf{Z}-\mathbf{D}]_{ij} u_j$ in (B.5) by its expectation, i.e., zero. This step simplifies the likelihood function. To see why, note that $P(y_{it}=1)$ can now be written as follows:

$$\begin{aligned} P(y_i = 1) &= P(y_i^* \geq 0) \\ &= P\left(u_i \leq \frac{\sum_j \beta z_{ij} x_j}{d_i}\right) \\ &= F_u\left(\frac{\sum_j \beta z_{ij} x_j}{d_i}\right), \end{aligned} \tag{B.6}$$

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

$$\begin{aligned} PL(\rho, \boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) &= \left[\prod^{N} P(y_i = 1)^{y_i}\right]\left[\prod^{N} (1 - P(y_i = 1))^{(1-y_i)}\right] \\ &\propto \left[\left(\prod^{N} \frac{\exp\left((\sum_j \beta z_{ij} x_j)/d_i\right)}{1 + \exp\left((\sum_j \beta z_{ij}^{y_i} x_j)/d_i\right)}\right)^{y_i}\right] \\ &\times \left[\left(\prod^{N} \frac{1}{1 + \exp\left((\sum_j \beta z_{ij} x_j)/d_i\right)}\right)^{(1-y_i)}\right]. \end{aligned} \tag{B.7}$$

Appendix B.2. PMLE for the temporal autoregressive model

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

$$\mathbf{y}^*_{(T\times 1)} = (\mathbf{I} - \gamma\mathbf{T})^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \gamma\mathbf{T})^{-1}\mathbf{u}. \tag{B.8}$$

Next, let

$$\mathbf{Z}_{(T\times T)} = (\mathbf{I} - \gamma\mathbf{T})^{-1}, \tag{B.9}$$

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

$$\begin{aligned} \mathbf{y}^* &= \mathbf{Z}\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} \\ &= \mathbf{Z}\mathbf{X}\boldsymbol{\beta} + \underbrace{(\mathbf{Z} - \mathbf{D})\mathbf{u}}_{\text{distributed}} + \underbrace{\mathbf{D}\mathbf{u}}_{\text{contemporaneous}}. \end{aligned} \tag{B.10}$$

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 and given that u is i.i.d., we arrive at the following expression for the probability of a positive outcome:

$$\begin{aligned} &Pr(y_t = 1) \\ &= Pr(y_t^* > 0) \\ &= Pr\left(u_t < [\mathbf{ZX}\beta]_t\right), \end{aligned} \tag{B.11}$$

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

$$\begin{aligned} PL(\gamma, \boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) &= \left[\prod^T P(y_t = 1)^{y_t}\right]\left[\prod^T (1 - P(y_t = 1))^{(1-y_t)}\right] \\ &\propto \left[\left(\prod^T 1 - \frac{1}{1 + \exp\left(-[\mathbf{ZX}\beta]_t/d_t\right)}\right)^{y_t}\right] \\ &\times \left[\left(\prod^T \frac{1}{1 + \exp\left(-[\mathbf{ZX}\beta]_t/d_t\right)}\right)^{(1-y_t)}\right]. \end{aligned} \tag{B.12}$$

Appendix B.3. PMLE for the spatio-temporal autoregressive model

Similarly to the above model, recall the reduced form:

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

We denote the spatio-temporal multiplier $(\mathbf{I} - \mathbf{Q})^{-1}$ again as $\mathbf{Z}_{(NT\times NT)}$ and define the matrix $\mathbf{D}_{NT\times NT}$ as a matrix that captures the diagonal elements of $\mathbf{Z}$ with all other elements being zeros.

$$\begin{aligned} \mathbf{y}^*_{(NT\times 1)} &= \mathbf{ZX}\boldsymbol{\beta} + \mathbf{Zu} \\ &= \mathbf{ZX}\boldsymbol{\beta} + \underbrace{(\mathbf{Z} - \mathbf{D})\mathbf{u}}_{\text{higher-order effects}} + \underbrace{\mathbf{Du}}_{\text{zero-order effects}}. \end{aligned} \tag{B.14}$$

Substituting $(\mathbf{Z} - \mathbf{D})\mathbf{u}$ with its expectation, we arrive at the following expression for the probability of a positive outcome:

$$\begin{aligned} &Pr(y_{it} = 1) \\ &= Pr(y_{it}^* > 0) \\ &= Pr\left(u_{it} < [\mathbf{ZX}\beta]_{it}\right) \end{aligned} \tag{B.15}$$

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

$$
\begin{aligned}
PL(\rho, \gamma, \boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) &= \left[\prod^{N}\prod^{T} P(y_{it}=1)^{y_{it}}\right]\left[\prod^{N}\prod^{T}(1-P(y_{it}=1))^{(1-y_{it})}\right] \\
&\propto \left[\left(\prod^{N}\prod^{T} 1 - \frac{1}{1+\exp\left(-[\mathbf{ZX}\beta]_{ij,t}/d_{it}\right)}\right)^{y_{it}}\right] \\
&\times \left[\left(\prod^{N}\prod^{T} \frac{1}{1+\exp\left(-[\mathbf{ZX}\beta]_{ij,t}/d_{it}\right)}\right)^{(1-y_{it})}\right].
\end{aligned}
\tag{B.16}
$$

Alternatively, for any binomial link function $g(\cdot)$, we have

$$
PL(\rho, \gamma, \boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) \propto \prod_{i=1}^{N}\prod_{t=1}^{T}\left[g^{-1}\left(\frac{[\mathbf{ZX}\beta]_{ij,s}}{d_{it}}\right)^{y_{it}}\left[1-g^{-1}\left(\frac{[\mathbf{ZX}\beta]_{ij,s}}{d_{it}}\right)\right]^{(1-y_{it})}\right].
\tag{B.17}
$$

Note that this expression requires an estimate for the values for $\mathbf{y}^*_{i0}$, i.e. the values *preceding* the first observed period in order to calculate the first period $\mathbf{y}^*_{i1}$.[10] Assuming mean stationarity, we draw on Kauppi and Saikkonen [10] and use what can be viewed as the unconditional expectation of $\mathbf{y}^*$ across all time period (and units): $E[\mathbf{y}^*] = (I - \rho\mathbf{W} - \gamma)^{-1}\bar{\mathbf{X}}\beta$, where $\bar{\mathbf{X}}$ are the sample means.

Appendix C. Speeding up computation further

Appendix C.1. Why still costly...

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 on data covering N units over T periods with reduced form

$$
\mathbf{y}^*_{NT\times NT} = \mathbf{ZX}\boldsymbol{\beta} + \mathbf{Zu},
\tag{C.1}
$$

where $\mathbf{Z} = \mathbf{A}^{-1} = (\mathbf{I} - \mathbf{Q})^{-1}$. This specification yields a pseudo likelihood function consisting of NT terms of the following form

$$
\begin{aligned}
P(y_j = 1) &= P(y^*_j \geq 0) \\
&= F_u\left(\frac{\mu_j}{d_j}\right),
\end{aligned}
\tag{C.2}
$$

[10] Because $\mathbf{y}^*$ is latent, dropping the first period from the likelihood merely shifts the problem to the next period, rather than solving it [cf. 9].

with $j \in \{1, 2, \ldots, NT\}$, $\mu = \mathbf{ZX}\boldsymbol{\beta}$, and $d_j = \mathbf{Z}_{jj}$. Perhaps the most straightforward implementation of expression (C.2) is to invert $\mathbf{A}$ directly using a decomposition-based solver, then multiplying $\mathbf{Z}$ with $\mathbf{X}\boldsymbol{\beta}$ to yield μ, and dividing by $diag(\mathbf{Z})$. However, this strategy is typically very slow, as most decomposition-based solvers operate with near cubic time complexity. Instead, we propose a strategy that avoids the full inversion of $\mathbf{A}$, but computes μ and d separately.

Appendix C.2. Computing μ

To compute μ we solve the linear system $\mathbf{A}\mu = \mathbf{X}\boldsymbol{\beta}$ for μ using an iterative method. In particular, we propose using the Biconjugate gradient stabilized method (Bi-CGSTAB), which yields similar performance to the more widely known conjugate gradient method, but is applicable even if $\mathbf{A}$ is not symmetric [17]. Doing so yields a substantial speed-up over decomposition-based solvers, especially when $\mathbf{A}$ is sparse, which will be the case as long as any spatial weights matrices entering $\mathbf{A}$ are neighborhood based.[11]

If $\mathbf{A}$ is block-diagonal, which is the case for all panel models that do not feature a temporal autoregressive term, then we can make use of the fact that the inverse of a block-diagonal matrix is the block diagonal matrix of block-wise inverses. In other words, instead of solving the full system, we can solve $\mathbf{A_t}\mu_t = \mathbf{X_t}\boldsymbol{\beta}$ for all $t \in \{1, 2, \ldots T\}$, whereas $\mathbf{A}_t$ represents a block in $\mathbf{A}$.

Appendix C.3. Computing d

First, we note that for panel models, d can always be computed in a period-wise fashion. This is obviously the case if $\mathbf{Q}$ does *not* include the panel temporal weights matrix $\mathbf{T}^*$, because then $\mathbf{A}$ is block-diagonal, and thus d is the concatenation of the period-wise diagonals $d_t = diag((\mathbf{A}_t)^{-1})$. Crucially, however, d can be computed analogously even if $\mathbf{Q}$ *does* include $\mathbf{T}^*$, and thus $\mathbf{A}$ is *not* block-diagonal. In the following, we provide a theorem to this end for the case where $\mathbf{Q}$ represents a spatial weights matrix; we note, however, that the results extends to the case where $\mathbf{Q}$ represents a panel outcome weights matrix (i.e. an $\mathbf{M}^*$ term), or a mixture of the two.

Theorem Appendix C.1. *Let $\mathbf{W}^*$ be a $NT \times NT$ block-diagonal panel spatial matrix as defined in* (A.16)*, and $\mathbf{T}^*$ be the panel temporal weights matrix as defined in* (A.17)*. Then $d = diag((\mathbf{I} - \gamma\mathbf{T}^* - \rho\mathbf{W}^*)^{-1}) = diag((\mathbf{I} - \rho\mathbf{W}^*)^{-1})$.*

[11] Note that the weights matrices for temporal dependence ($\mathbf{A}$) and outcome-interdependence ($\mathbf{M}$) are sparse by construction.

Define a *strictly lower block triangular* (SLBT) matrix as any square matrix with the following structure

$$\begin{pmatrix} \mathbf{0}_N & & & & \\ & \mathbf{0}_N & & \mathbf{0} & \\ & & \ddots & & \\ & \neq \mathbf{0} & & \mathbf{0}_N & \\ & & & & \mathbf{0}_N \end{pmatrix},$$

whereas $\mathbf{0}_N$ is the $N \times N$ matrix of zeros. It follows that $\gamma\mathbf{T}^*$ is SLBT. Note that $(\mathbf{I} - \gamma\mathbf{T}^* - \rho\mathbf{W}^*)^{-1}$ can be written as a Neumann series [13, ch. 2]:

$$\begin{aligned}(\mathbf{I} - \gamma\mathbf{T}^* - \rho\mathbf{W}^*)^{-1} &= \mathbf{I} + \sum_{l=1}^{L}(\gamma\mathbf{T}^* + \rho\mathbf{W}^*)^l \\ &= \mathbf{I} + (\gamma\mathbf{T}^* + \rho\mathbf{W}^*) \\ &\quad + (\gamma^2\mathbf{T}^{*2} + \gamma\rho\mathbf{T}^*\mathbf{W}^* + \gamma\rho\mathbf{W}^*\mathbf{T}^* + \rho^2\mathbf{W}^{*2}) \\ &\quad + \ldots\end{aligned} \tag{C.3}$$

Note that the product of two SLBT matrices is SLBT. Further note that the product of an SLBT matrix with a block-diagonal (BD) matrix is SLBT, regardless of the order of multiplication. It follows that the only terms in (C.3) with non-zero diagonals are of the form $\rho^l\mathbf{W}^{*l}$ for $l > 1$. Thus,

$$\begin{aligned}diag((\mathbf{I} - \gamma\mathbf{T}^* - \rho\mathbf{W}^*)^{-1}) &= diag(\mathbf{I}) + \sum_{l=1}^{L} diag((\gamma\mathbf{T}^* + \rho\mathbf{W}^*)^l) \\ &= diag(\mathbf{I}) + \sum_{l=1}^{L} diag((\rho\mathbf{W}^*)^l) \\ &= diag((\mathbf{I} - \rho\mathbf{W}^*)^{-1}).\end{aligned}$$

Thanks to the above theorem, we now only need an efficient method for calculating $d_t = diag((\mathbf{A}_t)^{-1})$. Here we propose two approaches. The first (preferred) one applies whenever $\mathbf{A}_t$ is composed of only a single weights parameter and weights matrix, e.g. $\mathbf{A}_t = \mathbf{I} - \rho\mathbf{W}$ or $\mathbf{A}_t = \mathbf{I} - \lambda\mathbf{M}$. In this case, we make use of the fact that $(\mathbf{A}_t)^{-1}$ can be written as a Neumann series, e.g. for the spatial case

$$(\mathbf{A}_t)^{-1} = \mathbf{I} + \sum_{l=1}^{\infty}(\rho\mathbf{W})^l. \tag{C.4}$$

Thus, an approximation for d_t may be obtained via

$$d_t \approx \tilde{d}_t = diag(\mathbf{I}) + \sum_{l=1}^{L} diag\left(\rho^l\mathbf{W}^l\right), \tag{C.5}$$

with L suitably large; we use $L = 8$. Note that we can precompute the series $\{\mathbf{W}, \mathbf{W^2}, \ldots, \mathbf{W^L}\}$ prior to optimization. Thus, the time complexity of evaluating $\tilde{d}_t$ during optimization is linear in N.

The second approach for computing d_t comes into play when $\mathbf{A}_t$ is composed of multiple weights matrices and parameters, as for instance in the binary simultaneous equation spatial model discussed in Section Appendix A.3. In this case, we use the method of [16], and examined by [8], which relies on a recursive algorithm to calculate the diagonal of a matrix inverse. Importantly, using the Takahashi equations to calculate d_t is considerably faster than computing the full decomposition-based inverse if $\mathbf{A}_t$ is sparse, which will generally be the case as long as any spatial weights matrices entering into $\mathbf{A}_t$ are neighborhood-based.

Appendix D. Evaluation and comparison of estimation strategies

In the remainder of the Online Appendix, we present all the MC simulations referred to in the main text. In all cases, we compare the performance of our estimator to that of some well-recognized alternatives.

Appendix E. Notes on replication material

Tables A3 & A4 compare the spatio-temporal estimation results when using different approximations for y^* for the initial period. Table A4 is equivalent to Table 1 reported in the main paper, and we recommend using this estimation procedure. Replication results for Table A3 are available on request.

Further note that our replication material does not replicate Figure A3, which demonstrates estimation times. Unless the code is rerun on exactly the same hardware setup, estimation times will obviously differ.

Table E.2: Simulation results for **spatial** PMLE (500 iterations)

	N=256			N=1,024			N=4,096		
	$\beta_0 = -0.5$	$\beta_1 = 1$	ρ	$\beta_0 = -0.5$	$\beta_1 = 1$	ρ	$\beta_0 = -0.5$	$\beta_1 = 1$	ρ
Experiment #1: $\rho = 0$									
Mean Coefficient Estimate	-0.530	1.019	-0.038	-0.507	1.006	-0.010	-0.501	1.002	-0.001
Mean Bias	0.139	0.101	0.205	0.063	0.049	0.097	0.032	0.024	0.051
RMSE	0.185	0.127	0.269	0.078	0.061	0.122	0.040	0.031	0.065
Actual SD of estimates	0.183	0.126	0.266	0.078	0.061	0.121	0.040	0.031	0.065
Overconfidence	1.089	0.991	1.080	0.961	0.976	0.954	0.990	0.984	1.007
Experiment #2: $\rho = 0.25$									
Mean Coefficient Estimate	-0.534	1.024	0.212	-0.503	1.004	0.245	-0.498	0.996	0.249
Mean Bias	0.140	0.110	0.183	0.064	0.050	0.086	0.032	0.025	0.045
RMSE	0.197	0.140	0.244	0.079	0.062	0.108	0.040	0.031	0.056
Actual SD of estimates	0.194	0.138	0.241	0.079	0.062	0.108	0.040	0.030	0.056
Overconfidence	1.224	1.061	1.178	1.050	0.980	1.060	1.063	0.963	1.090
Experiment #3: $\rho = 0.5$									
Mean Coefficient Estimate	-0.523	1.000	0.472	-0.497	0.987	0.492	-0.488	0.975	0.498
Mean Bias	0.147	0.115	0.139	0.069	0.054	0.069	0.037	0.034	0.036
RMSE	0.201	0.143	0.188	0.086	0.067	0.089	0.046	0.041	0.046
Actual SD of estimates	0.199	0.143	0.186	0.086	0.066	0.088	0.044	0.033	0.046
Overconfidence	1.309	1.051	1.223	1.221	0.985	1.209	1.256	0.991	1.251

Overconfidence is the standard deviation of the estimated parameter divided by the mean of its estimated standard error.

Table E.3: Simulation results for **temporal** PMLE (500 iterations; y_0^* is estimated by $E(y_t^*)$)

	N=64 & T=4			N=64 & T=16			N=256 & T=16		
	$\beta_0 = -0.5$	$\beta_1 = 1$	γ	$\beta_0 = -0.5$	$\beta_1 = 1$	γ	$\beta_0 = -0.5$	$\beta_1 = 1$	γ
Experiment #1: $\gamma = 0$									
Mean Coefficient Estimate	-0.503	1.017	-0.001	-0.502	1.006	-0.001	-0.501	1.002	-0.000
Mean Bias	0.089	0.102	0.086	0.045	0.049	0.039	0.021	0.024	0.019
RMSE	0.113	0.128	0.106	0.057	0.061	0.048	0.026	0.031	0.024
Actual SD of estimates	0.113	0.127	0.106	0.057	0.061	0.048	0.026	0.031	0.024
Overconfidence	1.013	1.000	1.025	1.050	0.971	1.008	0.963	0.985	1.004
Experiment #2: $\gamma = 0.25$									
Mean Coefficient Estimate	-0.497	0.999	0.249	-0.488	0.980	0.249	-0.486	0.972	0.249
Mean Bias	0.083	0.108	0.075	0.045	0.052	0.035	0.023	0.034	0.018
RMSE	0.104	0.136	0.095	0.056	0.064	0.044	0.028	0.041	0.022
Actual SD of estimates	0.104	0.137	0.095	0.054	0.061	0.044	0.024	0.030	0.022
Overconfidence	1.049	1.038	1.036	1.126	0.974	0.975	1.005	0.970	0.965
Experiment #3: $\gamma = 0.5$									
Mean Coefficient Estimate	-0.470	0.936	0.495	-0.442	0.882	0.498	-0.440	0.878	0.499
Mean Bias	0.083	0.130	0.060	0.067	0.119	0.034	0.061	0.122	0.017
RMSE	0.101	0.159	0.076	0.079	0.133	0.043	0.065	0.126	0.021
Actual SD of estimates	0.097	0.146	0.076	0.054	0.062	0.043	0.025	0.031	0.021
Overconfidence	1.061	1.019	1.048	1.244	1.001	1.148	1.140	1.011	1.118

Overconfidence is the standard deviation of the estimated parameter divided by the mean of its estimated standard error.

Table E.4: Simulation results for **spatio-temporal** PMLE (500 iterations) when $y_1^* = X_1\beta + u_1$ is assumed

	N=64 & T=4				N=64 & T=16				N=256 & T=16			
	$\beta_0 = -0.5$	$\beta_1 = 1$	γ	ρ	$\beta_0 = -0.5$	$\beta_1 = 1$	γ	ρ	$\beta_0 = -0.5$	$\beta_1 = 1$	γ	ρ
Experiment #1: $\gamma = 0.25$, $\rho = 0.25$												
Mean Coefficient Estimate	-0.515	1.005	0.252	0.234	-0.483	0.971	0.250	0.250	-0.483	0.965	0.249	0.250
Mean Bias	0.150	0.112	0.087	0.159	0.079	0.056	0.035	0.078	0.042	0.039	0.019	0.040
RMSE	0.203	0.147	0.108	0.206	0.100	0.070	0.045	0.097	0.051	0.047	0.024	0.051
Actual SD of estimates	0.203	0.147	0.109	0.206	0.098	0.064	0.046	0.098	0.049	0.031	0.024	0.051
Overconfidence	1.266	1.097	1.130	1.171	1.210	0.972	1.030	1.161	1.144	0.960	1.068	1.134
Experiment #2: $\gamma = 0.25$, $\rho = 0.5$												
Mean Coefficient Estimate	-0.512	0.992	0.255	0.482	-0.478	0.935	0.248	0.495	-0.465	0.927	0.247	0.501
Mean Bias	0.176	0.133	0.088	0.134	0.111	0.088	0.041	0.070	0.059	0.075	0.022	0.035
RMSE	0.241	0.166	0.110	0.178	0.142	0.106	0.052	0.093	0.071	0.083	0.028	0.044
Actual SD of estimates	0.241	0.166	0.110	0.177	0.141	0.083	0.052	0.093	0.062	0.040	0.027	0.044
Overconfidence	1.490	1.076	1.246	1.292	1.653	1.059	1.262	1.361	1.445	1.020	1.276	1.247
Experiment #3: $\gamma = 0.5$, $\rho = 0.25$												
Mean Coefficient Estimate	-0.486	0.943	0.485	0.239	-0.451	0.880	0.493	0.246	-0.441	0.869	0.495	0.247
Mean Bias	0.154	0.122	0.092	0.142	0.105	0.125	0.039	0.069	0.070	0.131	0.021	0.035
RMSE	0.200	0.153	0.117	0.183	0.127	0.143	0.051	0.088	0.082	0.137	0.026	0.044
Actual SD of estimates	0.200	0.142	0.116	0.183	0.118	0.078	0.050	0.088	0.057	0.039	0.026	0.044
Overconfidence	1.387	1.049	1.128	1.255	1.575	1.103	1.178	1.354	1.541	1.102	1.190	1.306

Overconfidence is the standard deviation of the estimated parameter divided by the mean of its estimated standard error.

Table E.5: Simulation results for **spatio-temporal** PMLE (500 iterations; y_0^* is estimated by $E(y_t^*)$)

	N=64 & T=4				N=64 & T=16				N=256 & T=16			
	$\beta_0 = -0.5$	$\beta_1 = 1$	γ	ρ	$\beta_0 = -0.5$	$\beta_1 = 1$	γ	ρ	$\beta_0 = -0.5$	$\beta_1 = 1$	γ	ρ
Experiment #1: $\gamma = 0.25$, $\rho = 0.25$												
Mean Coefficient Estimate	-0.523	1.001	0.249	0.229	-0.484	0.970	0.251	0.249	-0.483	0.964	0.249	0.249
Mean Bias	0.171	0.122	0.080	0.154	0.083	0.057	0.036	0.080	0.044	0.039	0.019	0.042
RMSE	0.237	0.155	0.100	0.202	0.105	0.071	0.046	0.099	0.056	0.047	0.024	0.053
Actual SD of estimates	0.236	0.156	0.100	0.201	0.104	0.064	0.046	0.099	0.053	0.031	0.024	0.053
Overconfidence	1.302	1.104	1.086	1.203	1.192	0.971	1.011	1.166	1.161	0.952	1.027	1.167
Experiment #2: $\gamma = 0.25$, $\rho = 0.5$												
Mean Coefficient Estimate	-0.614	0.997	0.233	0.459	-0.469	0.930	0.250	0.497	-0.464	0.925	0.247	0.501
Mean Bias	0.318	0.162	0.102	0.160	0.136	0.093	0.046	0.075	0.071	0.077	0.023	0.037
RMSE	0.580	0.206	0.141	0.231	0.176	0.113	0.059	0.099	0.087	0.086	0.029	0.047
Actual SD of estimates	0.570	0.206	0.140	0.228	0.174	0.089	0.059	0.099	0.079	0.042	0.029	0.047
Overconfidence	2.027	1.092	1.371	1.474	1.548	1.067	1.210	1.394	1.392	1.011	1.137	1.283
Experiment #3: $\gamma = 0.5$, $\rho = 0.25$												
Mean Coefficient Estimate	-0.530	0.937	0.493	0.224	-0.451	0.876	0.496	0.245	-0.438	0.860	0.499	0.246
Mean Bias	0.229	0.169	0.079	0.128	0.143	0.130	0.042	0.079	0.086	0.140	0.022	0.039
RMSE	0.331	0.210	0.104	0.176	0.179	0.149	0.053	0.101	0.103	0.146	0.027	0.049
Actual SD of estimates	0.330	0.200	0.103	0.174	0.173	0.083	0.053	0.101	0.082	0.041	0.027	0.049
Overconfidence	1.464	1.069	1.125	1.354	1.514	1.054	1.142	1.393	1.458	1.058	1.141	1.338

Overconfidence is the standard deviation of the estimated parameter divided by the mean of its estimated standard error.

Table E.6: Summary statistics for ρ parameter in common spatial probit estimators from 500 Monte Carlo iterations (100 iterations for RIS with N=1,024).

		$\rho = 0$			$\rho = 0.25$			$\rho = 0.5$		
	N	**256**	**1,024**	**4,096**	**256**	**1,024**	**4,096**	**256**	**1,024**	**4,096**
Bayes	Mean Bias	0.154	0.075	0.038	0.139	0.062	0.032	0.111	0.050	0.025
	RMSE	0.197	0.094	0.048	0.183	0.080	0.040	0.149	0.064	0.030
	Overconfidence	1.002	0.985	1.026	1.062	0.974	1.005	1.011	0.955	0.931
	Non-convergence	0	0	0	0	0	0	0	0	0
GMM	Mean Bias	0.205	0.098		0.185	0.088		0.135	0.070	
	RMSE	0.267	0.122		0.243	0.110		0.188	0.088	
	Overconfidence	1.065	0.960		1.160	1.060		1.009	1.072	
	Non-convergence	0	0		2	0		76	16	
Naive Probit	Mean Bias	0.548	0.278	0.139	0.712	0.626	0.654	1.405	1.465	1.493
	RMSE	0.692	0.345	0.176	0.849	0.694	0.673	1.532	1.496	1.500
	Overconfidence	1.271	1.279	1.311	1.275	1.210	1.219	1.188	1.106	1.087
	Non-convergence	0	0	0	0	0	0	0	0	0
RIS	Mean Bias	0.099	0.046		0.147	0.121		0.255	0.229	
	RMSE	0.129	0.057		0.182	0.131		0.280	0.236	
	Overconfidence	3.141	1.468		5.051	4.213		10.756	12.421	
	Non-convergence	0	0		0	0		0	0	
SPMLE	Mean Bias	0.205	0.097	0.051	0.183	0.086	0.045	0.139	0.069	0.036
	RMSE	0.269	0.122	0.065	0.244	0.108	0.056	0.188	0.089	0.046
	Overconfidence	1.080	0.954	1.007	1.178	1.060	1.090	1.223	1.209	1.251
	Non-convergence	1	1	0	0	1	5	3	6	10

Table E.7: Summary statistics for γ parameter for RIS and PMLE estimators from 500 Monte Carlo iterations (100 iterations for RIS with N=64 and T=16).

		$\gamma = 0$			$\gamma = 0.25$			$\gamma = 0.5$		
	N x T	**64 x 4**	**64 x 16**	**256 x 16**	**64 x 4**	**64 x 16**	**256 x 16**	**64 x 4**	**64 x 16**	**256 x 16**
RIS	Mean Bias	0.075	0.032		0.085	0.082		0.162	0.197	
	RMSE	0.093	0.039		0.106	0.089		0.178	0.200	
	Overconfidence	5.909	1.997		11.607	9.399		22.784	16.475	
	No convergence	0	0		0	0		0	0	
PMLE	Mean Bias	0.086	0.039	0.019	0.075	0.035	0.018	0.060	0.034	0.017
	RMSE	0.106	0.048	0.024	0.095	0.044	0.022	0.076	0.043	0.021
	Overconfidence	1.025	1.008	1.004	1.036	0.975	0.965	1.048	1.148	1.118
	No convergence	0	0	0	0	0	0	0	0	0

Table E.8: Summary statistics for ρ parameter for RIS and PMLE estimators from 500 Monte Carlo iterations (100 iterations for RIS with N=64 and T=16).

		$\gamma = 0.25$ & $\rho = 0.25$			$\gamma = 0.25$ & $\rho = 0.5$			$\gamma = 0.5$ & $\rho = 0.25$		
	N x T	**64 x 4**	**64 x 16**	**256 x 16**	**64 x 4**	**64 x 16**	**256 x 16**	**64 x 4**	**64 x 16**	**256 x 16**
RIS	Mean Bias	0.219	0.168		0.317	0.294		0.275	0.201	
	RMSE	0.260	0.187		0.351	0.306		0.299	0.214	
	Overconfidence	22.108	19.233		24.361	22.673		30.293	29.351	
	No convergence	0	0		0	0		0	0	
PMLE	Mean Bias	0.154	0.080	0.042	0.160	0.075	0.037	0.128	0.079	0.039
	RMSE	0.202	0.099	0.053	0.231	0.099	0.047	0.176	0.101	0.049
	Overconfidence	1.203	1.166	1.167	1.474	1.394	1.283	1.354	1.393	1.338
	No convergence	2	0	0	3	3	2	1	1	0

Table E.9: Summary statistics for γ parameter for RIS and PMLE estimators from 500 Monte Carlo iterations (100 iterations for RIS with N=64 and T=16).

		$\gamma = 0.25$ & $\rho = 0.25$			$\gamma = 0.25$ & $\rho = 0.5$			$\gamma = 0.5$ & $\rho = 0.25$		
	N x T	**64 x 4**	**64 x 16**	**256 x 16**	**64 x 4**	**64 x 16**	**256 x 16**	**64 x 4**	**64 x 16**	**256 x 16**
RIS	Mean Bias	0.090	0.092		0.096	0.094		0.170	0.232	
	RMSE	0.112	0.099		0.122	0.105		0.189	0.236	
	Overconfidence	12.465	8.739		16.473	12.526		21.284	17.412	
	No convergence	0	0		0	0		0	0	
PMLE	Mean Bias	0.080	0.036	0.019	0.102	0.046	0.023	0.079	0.042	0.022
	RMSE	0.100	0.046	0.024	0.141	0.059	0.029	0.104	0.053	0.027
	Overconfidence	1.086	1.011	1.027	1.371	1.210	1.137	1.125	1.142	1.141
	No convergence	2	0	0	3	3	2	1	1	0

Figure E.2: Distribution of ρ estimates from Monte Carlo simulations for Bayes, GMM, MLE, RIS, and PMLE estimator.

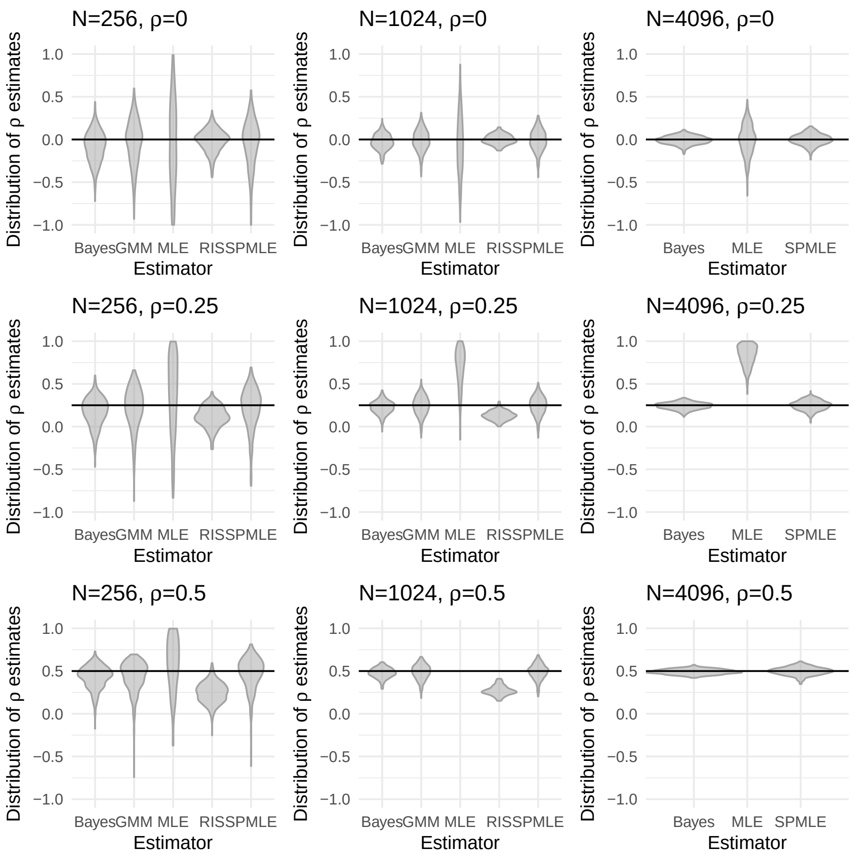


Figure E.3: Distribution of γ estimates from Monte Carlo simulations for RIS and PMLE estimator.

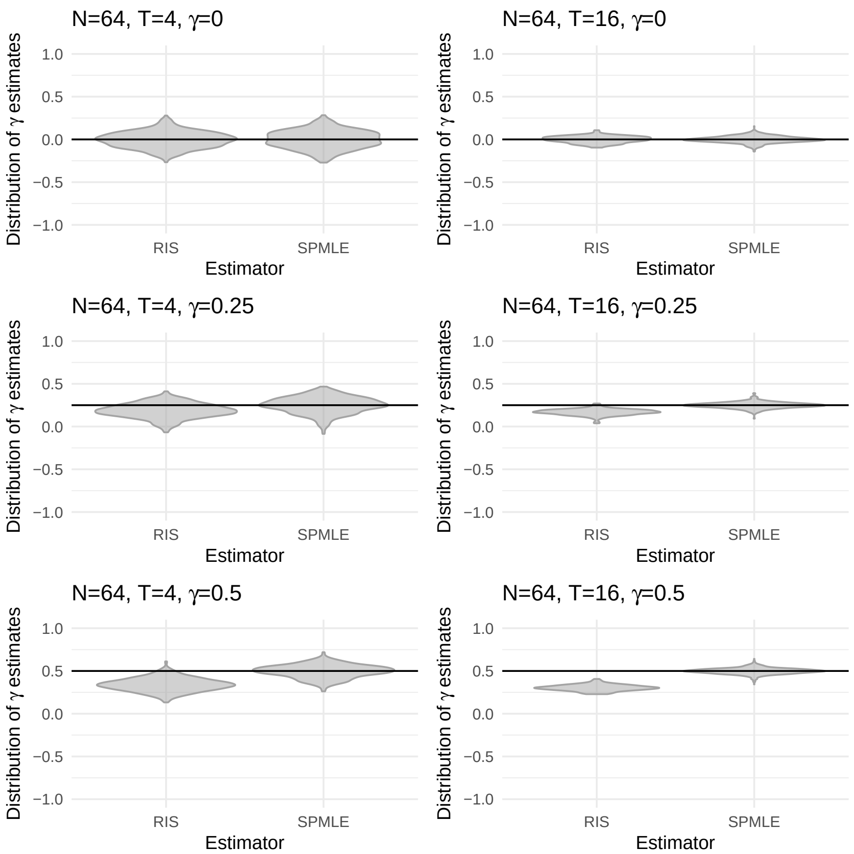
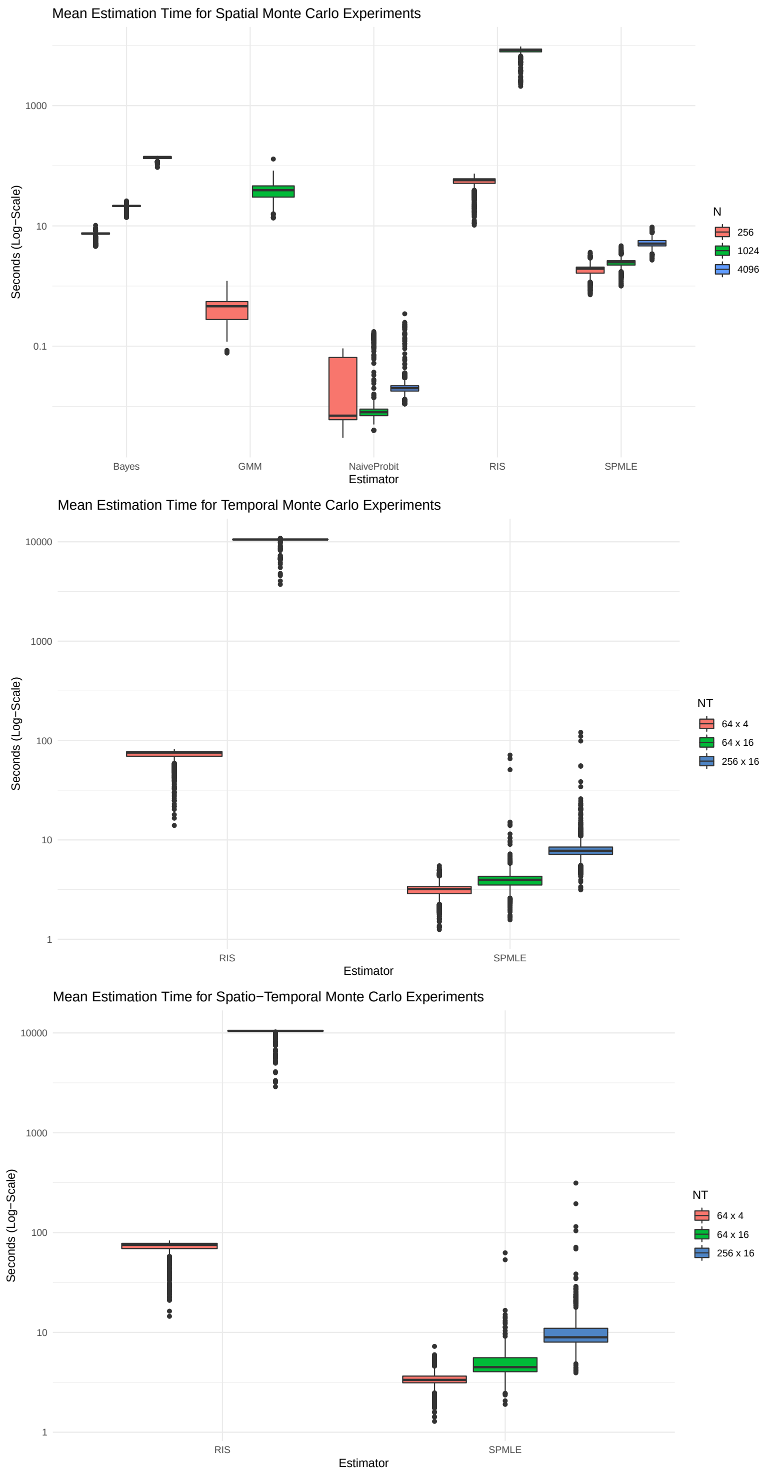


Table E.10: Replication of Franzese et al.'s (2016) simulation results for spatio-temporal RIS (100 iterations). The experiments here use the DGP (incl. **W**) of Franzese et al. (2016). The implementation code is the authors' own—a direct translation of the original MATLAB code to our own R code.

	$\beta_0 = -1.5$	$\beta_1 = 3$	ρ	γ
Experiment #1: $\rho = 0.1$, $\gamma = 0.3$				
Mean Coefficient Estimate	-1.434	2.859	0.088	0.270
Bias	0.115	0.219	0.037	0.032
RMSE	0.145	0.263	0.045	0.038
Actual SD of estimates	0.130	0.224	0.043	0.022
Experiment #2: $\rho = 0.1$, $\gamma = 0.5$				
Mean Coefficient Estimate	-1.270	2.487	0.070	0.448
Bias	0.239	0.523	0.060	0.053
RMSE	0.277	0.601	0.107	0.061
Actual SD of estimates	0.156	0.316	0.104	0.033
Experiment #3: $\rho = 0.25$, $\gamma = 0.3$				
Mean Coefficient Estimate	-1.401	2.798	0.220	0.274
Bias	0.122	0.232	0.042	0.029
RMSE	0.152	0.283	0.051	0.035
Actual SD of estimates	0.116	0.199	0.041	0.023
Experiment #4: $\rho = 0.25$, $\gamma = 0.5$				
Mean Coefficient Estimate	-1.190	2.374	0.231	0.456
Bias	0.320	0.639	0.046	0.048
RMSE	0.373	0.747	0.064	0.062
Actual SD of estimates	0.209	0.409	0.061	0.044

Figure E.4: Mean Estimation Times for Spatial, Temporal and Spatio-Temporal Estimators



Reference

[1] Anselin, L. (2002). Under the hood: Issues in the specification and interpretation of spatial regression models. *Agricultural Economics*, 27(3):247–267.

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

[3] 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.

[4] Beron, K. J. and Vijverberg, W. P. M. (2004). Probit in a Spatial Context: A Monte Carlo Analysis. In Anselin, L., Florax, R. J. G. M., and Rey, S. J., editors, *Advances in Spatial Econometrics: Methodology, Tools and Applications*, pages 169–195. Springer, Berlin.

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

[6] 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.

[7] Darmofal, D. (2015). *Spatial analysis for the social sciences.* Cambridge University Press.

[8] Erisman, A. and Tinney, W. (1975). On computing certain elements of the inverse of a sparse matrix. *Communications of the ACM*, 18(3):177–179.

[9] 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.

[10] Kauppi, H. and Saikkonen, P. (2008). Predicting us recessions with dynamic binary response models. *The Review of Economics and Statistics*, 90(4):777–791.

[11] 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.

[12] 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.

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

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

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

[16] Takahashi, K., Fagan, J., and Chin, M.-S. (1973). Formation of a sparse bus impedence matrix and its application to short circuit study. *Power Industry Computer Applications Conference*, pages 4–6.

[17] Van der Vorst, H. A. (1992). Bi-cgstab: A fast and smoothly converging variant of bi-cg for the solution of nonsymmetric linear systems. *SIAM Journal on scientific and Statistical Computing*, 13(2):631–644.

[18] Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, pages 5–42.

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