Testing Spatial Correlation in Probit Models*

Joris Pinkse‡

The University of British Columbia

August 1997

Abstract

I propose tests for exclusion of spatially lagged–dependent variables and absence of spatial–error correlation in Probit models. I give a Lagrange–Multiplier justification and discuss when the asymptotic distribution can be used to draw inferences and what to do when it cannot.

.

*Key words: Spatial statistics, generalized residuals, binary choice, probit model, discrete choice.

JEL Classifications: C12,C52, C21

†Joris Pinkse, The University of British Columbia, Department of Economics, 997–1873 East Mall, Vancouver B.C., Canada V6T 1Z1, tel.: (604) 822 4608, fax: (604) 822 5915, e–mail: joris@econ.ubc.ca.

‡This research was supported by a grant from the Social Sciences and Humanities Research Council of Canada. I thank Craig Brett and participants at the North–American Regional Science Association meetings at Cincinatti for useful suggestions.

1 Motivation

Spatial–error correlation plays the same role in cross–sectional studies involving spatial data that serial correlation plays in studies involving time–series data. Moreover, there exists a rich set of econometric techniques that can be used to test for spatial dependence and to obtain efficient estimates of model parameters in the presence of such dependence (see Anselin 1988). With few exceptions, however, attention has been limited to continuous dependent variables.

Nevertheless, many interesting problems involve discrete choices. Examples include firms' and households' decisions to relocate, and firms' choices of technologies when network externalities are present. Spatial dependence arises in these examples because an agent's decision can be influenced by the choices made by her neighbors. Moreover, the notion of "neighbor" or "closeness" can be geographic, as with location problems, or more general, as with networks. Unfortunately, most spatial tests and estimation techniques yield inconsistent results when applied to discrete–choice problems.

A few researchers have proposed techniques that can be used to estimate spatial discrete–choice models (Case 1992, McMillen 1992, and Pinkse and Slade 1995). These techniques, however, are somewhat more complex than in a time–series context. Indeed, the ordinary probit yields consistent estimates in the presence of serial but not spatial correlation,[1] which makes two–step estimation problematic for the latter. It is therefore more than usually desirable to have a test that can be used to detect spatial dependence before embarking on estimation.

In this paper I propose three tests for spatial correlation in a probit model. The tests are based on a principle originally due to Moran (1950) and further explored by Cliff and Ord (1973), Burridge (1980), and Anselin, Bera, Florax, and Yoon (1995), among others.

Cliff and Ord were particularly interested in a linear–regression model with a spatial–autoregressive structure on normal errors where the observations lie on a perfect grid. They derived conditions under which the Moran–test statistic has a limiting–normal distribution under the null hypothesis: the absence of spatial dependence.[2] These conditions are satisfied by a finite–order stationary spatial–

[1] This is because most spatial–correlation patterns that are used in practice cause heteroskedasticity in the errors.

[2] The mere absence of spatial correlation would not suffice.

autoregressive process on a regular grid, but they also allow for more general spatial–correlation patterns. In addition, Cliff and Ord extended their result to the case in which the errors are nonnormal. Their proof, however, assumes that the error distribution has an infinite number of moments.

Burridge gives a Maximum–Likelihood justification for the Moran test and shows that the square of the Moran–test statistic is essentially a Lagrange–Multiplier–test statistic.

Anselin, Bera, Florax, and Yoon (1995) consider a model in which both spatially lagged–dependent variables and spatially correlated errors can appear. They derive tests for either of these forms of spatial dependence. Moreover, their tests have asymptotically correct size under a local alternative to the maintained hypothesis that the other type of dependence does not occur.

As far as I know, none of the above tests have been extended to discrete choice models.

In the next section, I derive three joint tests for the exclusion of spatially lagged–dependent variables and absence of spatial correlation in the errors of a probit model and three sets of tests for each of the marginal hypotheses. Like Burridge, I take the Lagrange–Multiplier route.

2 The Cliff and Ord Test

Consider the spatial–regression model

$$\begin{aligned} Y &= \delta W_1 Y + X\beta + U, && (1) \\ U &= \psi W_2 U + \varepsilon, && (2) \end{aligned}$$

where $\varepsilon \sim N(0, \sigma^2 I_n)$ and W_1 and W_2 are weighting matrices with diagonal elements equal to zero.

Cliff and Ord assumed that the rows of the matrix W_2 were normalized to sum to one. Dividing by the row sums, however, changes the model unless all row sums are equal. Our test does not impose this restriction.

Moran's test statistic for $\psi = 0$ against the alternative $\psi \neq 0$, with $\delta = 0$, is

$$\hat{\tau}_M = \frac{\hat{U}^T W_2 \hat{U}}{\{\mathrm{tr}(W_2^2 + W_2^T W_2)\}^{\frac{1}{2}} \hat{\sigma}^2}, \qquad (3)$$

where $\hat{U} = Y - X\hat{\beta}$, $\hat{\beta}$ is the OLS estimate of β, $\hat{\sigma}^2 = \hat{U}^T \hat{U}/n$ and tr denotes the trace of a matrix.

The statistic $\hat{\tau}_M$ does not have a limiting standard–normal distribution for all choices of W_2. For instance, when W_2 is any orthogonal–projection matrix with k eigenvalues equal to one, $U^T W_2 U$ has an exact χ^2–distribution with k degrees of freedom. There is no renorming of $U^T W_2 U$ that will have a standard–normal distribution, and the presence of nuisance parameters is unlikely to alleviate this problem. Cliff and Ord's condition [cf. Cliff and Ord (1973), Theorem 1 on page 36] for a limiting–normal distribution is equivalent to

$$n \frac{\sum_{ij} w_{2ij}^2}{(\sum_{ij} w_{2ij})^2} = O_e(1), \tag{4}$$

where w_{2ij} is the (i,j)–element of W_2 and $O_e(\cdot)$ means the "exact order of", i.e. the left–hand side is $O(1)$ but not $o(1)$.

An example of an economic model in which (4) is not satisfied is the symmetric price–competition model of Chamberlin (1933), which implies that $w_{2ij} = 1$ for $i \neq j$, and $w_{2ii} = 0$. Accordingly, the left–hand side in (4) is $O(n^{-1}) = o(1) \neq O_e(1)$.

Even if condition (4) is satisfied asymptotically, the left–hand–side quantity in (4) may be relatively large or small, which can cause actual critical values to differ considerably from nominal critical values in small samples.[3] It should be noted though that for weight sequences that decline slowly as the distance between locations increases, the coefficient estimates are generally consistent but not asymptotically normal.[4]

3 Our Test

The probit analog to the spatial regression model is

$$Y^* = \delta W_1 Y^* + X\beta + U, \tag{5}$$

$$U = \psi W_2 U + \varepsilon, \tag{6}$$

$$Y_i = I(Y_i^* \geq 0), \ i = 1, \ldots, n, \tag{7}$$

where $I(\cdot)$ is the indicator function and Y^* is a vector of latent variables. I use the usual normalization, so that $\varepsilon \sim N(0, I_n)$.

[3] I deliberately leave the notions of "large" and "small" vague here.

[4] See Pinkse and Slade (1996).

Residuals in a probit model cannot be determined because Y^* is latent. Instead, I use the concept of generalized errors due to Cox and Snell (1968). The i–th generalized error is defined by

$$E[U_i|Y_i] = \frac{\phi(x_i^T\beta)}{\Phi(x_i^T\beta)\{1-\Phi(x_i^T\beta)\}}\{Y_i - \Phi(x_i^T\beta)\}, \tag{8}$$

where ϕ and Φ are the density and distribution functions of the standard–normal distribution. The left–hand side of (8) is often taken as a definition of a generalized error. The equality holds for the standard–probit model. The right–hand side in (8) is denoted $\tilde{U}_i(\beta)$. The generalized residual is then $\hat{U}_i = \tilde{U}_i(\hat{\beta})$, where $\hat{\beta}$ is the ordinary–probit estimate.

I now derive test statistics for $\psi, \delta = 0$ against the alternatives that at least one is nonzero, for $\psi = 0$ against $\psi \neq 0$ under the maintained hypothesis that $\delta = 0$ and for $\delta = 0$ against $\delta \neq 0$ under the maintained hypothesis that $\psi = 0$. Tests that do not assume the maintained hypothesis can be constructed but the derivation of their limiting distribution is far more complex.

Like Burridge (1980), I follow a Maximum–Likelihood route. Let $\theta = [\psi, \delta, \beta^T]^T$. The likelihood function is (using lower case letters for the realizations of random variables)

$$L(\theta; y, x) = \int_{R(y)} \phi_{\mu,V}(y^*)dy^*, \tag{9}$$

where $R(y) = R_1(y_1) \times \cdots \times R_n(y_n)$ with $R_i(y_i) = (-\infty, 0)$ if $y_i = 0$ and $R_i(y_i) = [0, \infty)$ if $y_i = 1$, $\phi_{\mu,V}$ is the n–variate normal density with mean μ and variance V. In the linear–spatial–probit model used here

$$\mu = \mu(\psi, \delta, \beta) = (I_n - \delta W_1)^{-1}X\beta, \tag{10}$$

$$V = V(\psi, \delta, \beta) = (I_n - \delta W_1)^{-1}(I_n - \psi W_2)^{-1}(I_n - \psi W_2^T)^{-1}(I_n - \delta W_1^T)^{-1}. \tag{11}$$

Let $\theta_0 = [0, 0, \beta^T]^T$, the parameter vector under the null hypothesis and by $\hat{\theta}_0$ an estimator thereof.

The score is (see Appendix A for details)

$$S = \frac{\partial \log L}{\partial \theta}_{\theta_0} = \begin{bmatrix} \tilde{U}^T W_2 \tilde{U} \\ (\tilde{Y}^*)^T W_1^T \tilde{U} \\ X^T\tilde{U}, \end{bmatrix}, \tag{12}$$

where $\tilde{Y}^* = X\beta + \tilde{U}$.

The information matrix is then

$$J = E[SS^T] = \begin{bmatrix} T_{D;22} & T_{D;21} & 0 \\ T_{D;21} & T_{D;11} + ||D^{\frac{1}{2}} W_1^T X\beta||^2 & \beta^T X^T W_1 DX \\ 0 & X^T D W_1^T X\beta & X^T DX, \end{bmatrix} \tag{13}$$

where $T_{D;ij} = \mathrm{tr}(W_i D(W_j + W_j^T)D)$, $i = 1, 2$, and where D is a diagonal matrix with elements $d_i = E[\tilde{U}_i^2] = \phi_i^2/(\Phi_i(1-\Phi_i))$ with $\Phi_i = \Phi(x_i^T\beta)$ and $\phi_i = \phi(x_i^T\beta)$. The matrix D appears because the generalized errors are heteroskedastic even though the errors themselves are not.

The LM test statistics for the three hypotheses are

$$\hat{\tau}_\psi = \frac{\hat{S}_\psi^2}{\hat{J}_{\psi\psi}}, \quad \hat{\tau}_\delta = \frac{\hat{S}_\delta^2}{\hat{J}_{\delta\delta} - \hat{J}_{\delta\beta} J_{\beta\beta}^{-1} J_{\beta\delta}}, \quad \hat{\tau}_{\psi\delta} = \frac{(\hat{J}_{\delta\delta} - \hat{J}_{\delta\beta}\hat{J}_{\beta\beta}^{-1}\hat{J}_{\beta\delta})\hat{S}_\psi^2 - 2\hat{J}_{\psi\delta}\hat{S}_\psi\hat{S}_\delta + \hat{J}_{\psi\psi}\hat{S}_\delta^2}{\hat{J}_{\psi\psi}(\hat{J}_{\delta\delta} - \hat{J}_{\delta\beta}\hat{J}_{\beta\beta}^{-1}\hat{J}_{\beta\delta}) - \hat{J}_{\psi\delta}^2} \tag{14}$$

where

$$\hat{S} = \begin{bmatrix} \hat{S}_\psi \\ \hat{S}_\delta \\ \hat{S}_\beta \end{bmatrix}, \ \hat{J} = \begin{bmatrix} \hat{J}_{\psi\psi} & \hat{J}_{\psi\delta} & 0 \\ \hat{J}_{\psi\delta} & \hat{J}_{\delta\delta} & \hat{J}_{\delta\beta} \\ 0 & \hat{J}_{\beta\delta} & \hat{J}_{\beta\beta}, \end{bmatrix}. \tag{15}$$

and again hats indicate that the corresponding expression is estimated.

A slight alternative to the above approach would be to replace the matrix $\hat{D}$ with the matrix $\hat{D}^*$ with diagonal elements $\hat{U}_i^2$.

A third approach one could consider is to start off with removing the heteroskedasticity. This is the approach taken by Pinkse and Slade (1996) in their simulation test for spatial error correlation. Note that

$$\frac{\tilde{U}_i}{\sqrt{E\tilde{U}_i^2}} = Y_i\sqrt{\frac{1-\Phi_i}{\Phi_i}} - (1-Y_i)\sqrt{\frac{\Phi_i}{1-\Phi_i}}, \tag{16}$$

which is estimated by

$$\hat{U}_i^+ = Y_i\sqrt{\frac{1-\hat{\Phi}_i}{\hat{\Phi}_i}} - (1-Y_i)\sqrt{\frac{\hat{\Phi}_i}{1-\hat{\Phi}_i}}. \tag{17}$$

The third set of test statistics $\hat{\tau}_\psi^+, \hat{\tau}_\delta^+, \hat{\tau}_{\psi\delta}^+$ is then identical to the second with $\hat{U}$ replaced with $\hat{U}^+$, $\hat{D}$ with I_n and all other quantities based on $\hat{U}$ or $\hat{D}$ with their corresponding substitutes.

4 Asymptotic Distribution

If Condition (4) holds it is easily shown that under fairly mild conditions the limiting distribution of all three triplets of test statistics is χ^2 under independence. The number of degrees of freedom is one except for the test statistics for the joint hypothesis, in which case it is two.

Theorem 1 *If (4) holds and in addition,*

1. $\{Y_i\}$ *is a sequence of independent variables,*

2. $(\sum_{ij} |w_{2ij}|)^2/T_{22} = O(n)$,

3. $\sup_i ||x_i|| = o(n^{\frac{1}{12}})$,

then $\hat{\tau}_\psi, \hat{\tau}_\delta, \hat{\tau}_\psi^+, \hat{\tau}_\delta^+ \xrightarrow{\mathcal{L}} \chi_1^2$ *and* $\hat{\tau}_{\psi\delta}, \hat{\tau}_{\psi\delta}^+ \xrightarrow{\mathcal{L}} \chi_2^2$.

In Appendix B I prove that the limiting distribution of $\hat{U}^T W_2 \hat{U}/\sqrt{\hat{T}_{22}}$ is $N(0,1)$ under the null hypothesis, where the other results follow similarly.

Condition 1 merely states the null hypothesis. The additional restriction on the weights and variances d_i, Condition 2, is equivalent to (4) if the weights are nonnegative and the variances are bounded away from zero. The variances are allowed to be arbitrarily close to zero as long as it is not in a systematic way in exactly those places in which the weights are nonzero. Finally, Condition 3 restricts how the regressor variables are generated. If the x_i's were random this condition could be replaced with $\sup_i ||x_i|| = o_p(n^{\frac{1}{12}})$, which is rarely violated for i.i.d. data.

5 Simulations

To establish the performance of the tests I have carried out a Monte Carlo simulation study of both the size and power of the tests. In Tables 1 through 4 are the rejection rates at the 5% level for the tests for residual correlation (RSD), spatially lagged dependent variables (STRCT), and joint (JOINT) tests, as well as the standard (STD) and alternative (ALT) specifications, i.e. without and with prior heteroskedasticity correction.

The four weighting matrix specifications chosen are:

1. The observations lie on a square grid where $w_{ij} = 1$ for observations that are one grid point removed and $w_{ij} = 0$, otherwise.

2. The observations have randomly selected (uniformly) lattitudes and longitudes and $w_{ij} = 1/(1 + |i - j|^3)$, for $i \neq j$.

3. Ditto, with $|i - j|^3$ replaced with $|i - j|^2$.

4. The observations lie on a square grid where $w_{ij} = 1$ for $i \neq j$.

I will denote these weighting matrices by W_A through W_D respectively.

W_A is included as the standard case of a spatial first order autoregressive model. w_{ij} declines in specification W_B with the Euclidean distance between i and j but only gradually instead of an abrupt cut–off. Specification W_C is as W_B, except that the decline takes place at a slower rate. W_D is included as an example of a case that is expected to mess up the test's performance.

Aside from boundary effects, the alternative of W_A is stationary, the others are not.[5] Specification W_B satisfies (4) and 2, whereas W_C and W_D do not.

In all cases the number of replications was taken to be 10,000. In the size experiments the number of observations was either 100 or 256 and the number of regressors 3 or 6. The regressors included a constant term and have a joint normal distribution for each observation but were independent across observations and of the error term.

The experiments were carried out on a Pentium Pro 200 MHz computer running the NeXTSteP operating system. Random numbers were generated using Earl T. Cohen's standard Unix nonlinear additive feedback random number generator which claims a period of approximately $2^{35} \approx 3.5 \times 10^{10}$. The probit estimates were computed using artificial regressions as described in Davidson and MacKinnon (1993), Section 15.4. The artificial regressions were carried out by means of the singular value decomposition (Householder transformations) [cf. Golub and van Loan (1996)] to reduce rounding error. All routines were programmed by me in the programming language C.

The results in Tables 1, 2 and 3 suggest that under these specifications the limiting distribution is reasonably well approximated. Note that the approximation found in Table 1 is likely to be

[5]Nonstationarity here does not mean unit roots.

worse than for a first order time series autoregressive specification in the same way that the normal approximation for a test for higher order time series autocorrelation is often poorer than one for first order time series autocorrelation. What is disappointing is that there is not much improvement when the sample size is increased to 256.

Table 4 is less encouraging. Although W_D is an extreme example, there are likely to be more moderate examples in which the test performs poorly, for which the simulation procedure introduced in the next section can be useful.

The entries in tables 5 through 8 are the rejection rates at the 5% significance level under various alternatives. In all cases the weighting matrices were divided by $\sum_{ij} w_{ij}/n$ to make the comparison across alternatives easier and to ensure that all roots are between -1 and 1. In all cases the number of replications is 10000.

The magnitude of the numbers in the table have little significance in their own right. It is easy to generate a sample for which the values of $x_i^T\beta$ are so large in absolute value that for small to moderate values of the dependence parameters the likelihood of sign changes in $x_i^T\beta + U_i$ (and hence changes in Y_i) is very small, thereby decreasing rejection rates. Note also that the power of the test is likely to be less than the power of a comparable test in a linear model for similar reasons.

As before, the first table corresponds to W_A, the second to W_B and so on. Table 5 suggests that rejection levels increase with the value of ψ or δ and decrease with the number of regressors.

The test in its standard form appears to perform better than the alternative specification in which heteroskedasticity correction is carried out first. Indeed, the test with the prior heteroskedasticity correction rejects in less than 5% of the cases for small ψ or δ. Further experiments should reveal whether the intuition that this phenomenon disappears with a larger sample size in the same way that it does when ψ or δ is taken to be 0.5, is correct.

Tables 6 and 7 paint pretty much the same picture. Not surprisingly, it appears as if the test for a specific alternative performs better against that alternative than the joint test, which rejects more often than the test against the other alternative, at least for the standard form of the test.

The results in Table 8 are quite different, just like Table 4 was very different from Tables 1 through 3. It underlines the fact that there are alternatives against which the test should not be used.

6 When the Asymptotic Theory Breaks Down

When (4) fails, or when one wants to improve on the asymptotic approximation of the small sample distribution, it is possible to adopt a numerical approach. Indeed, for any of the test statistics one can (see also Pinkse and Slade 1996)

1. Estimate the probit model to obtain $\hat{\beta}$ and determine the test statistic of choice, say $\hat{\tau}$,
2. For $r = 1, \ldots, R$, with R the number of replications, draw independent random numbers Y_{ri} from the Bernoulli distribution with probability of success $\Phi(x_i^T \hat{\beta})$,
3. Estimate a probit model for each replication r for the sample $\{Y_{ri}, x_i\}$, $i = 1, \ldots, n$,
4. Generate test statistics $\hat{\tau}_r$ for each replication,
5. Determine for what fraction of the replications $\hat{\tau}_r$ exceeds $\hat{\tau}$; this is the p–value.

Since the test statistics are 'pivotal', bootstrap theory suggests that the bootstrap procedure should yield a better approximation to the finite sample distribution than if one used the asymptotic distribution.

7 Conclusion

In this paper I propose testing methods for dual–spatial correlation in a probit model. All tests are of the Lagrange–Multiplier type and differ in the way the implicit heteroskedasticity is dealt with.

I derive the asymptotic distribution of the test and carry out simulations to establish the performance of the test statistics against a range of alternatives. It appears as if the test performs reasonably well against a range of alternatives, but that there are alternatives against which the test has no power and/or for which the limiting distribution under the null hypothesis of the test is nonnormal. For those cases, I suggest a computational alternative.

References

Amemiya, T., 1985, Advanced Econometrics (Harvard University Press, Cambridge MA).

Anselin, L., 1988, Spatial Econometrics: Methods and Models, Kluwer Academic Publishers (Dordrecht, The Netherlands).

Anselin, L., Bera, A. K., Florax, R. and M.J. Yoon, 1995, Simple Diagnostic Tests for Spatial Dependence, forthcoming in Regional Science and Urban Economics.

Burridge, P., 1980, On the Cliff–Ord Test for Spatial Correlation, Journal of the Royal Statistical Society B 42, 107–108.

Case, A.C., 1992, Neighborhood Influence and Technological Change, Regional Science and Urban Economics 22, 491–508.

Chamberlin, E.H., 1933, The Theory of Monopolistic Competition, Harvard University Press, Cambridge MA.

Cliff, A. and K.J. Ord, 1973, Spatial Autocorrelation (Pion, London).

Cox, D.R. and Snell, E.J., 1968, A General Definition of Residuals, Journal of the Royal Statistical Society B 39, 248–275.

Golub, G.H., 1996, Matrix Computations, third edition (Johns Hopkins University Press, Baltimore Md)

McMillen, D.P., 1992, Probit with Spatial Autocorrelation, Journal of Regional Science 32, 335-348.

Moran, P.A.P., 1950, A Test for Serial Independence of Residuals, Biometrika 37, 178–181.

Pinkse, J. and M. Slade, 1996, Contracting in Space: An Application of Spatial Statistics to Discrete-Choice Models, forthcoming in the *Journal of Econometrics*.

A Derivation of the Test

I need the first two partial derivatives of the loglikelihood $\log L = \log \int \phi$, omitting arguments and integration area. Note that

$$\frac{\partial \log L}{\partial \theta} = \frac{\partial L}{\partial \theta}\frac{1}{L}, \tag{18}$$

$$\frac{\partial^2 \log L}{\partial \theta \partial \theta^T} = \frac{\partial^2 L}{\partial \theta \partial \theta^T}\frac{1}{L} - \frac{\partial L}{\partial \theta}\frac{\partial L}{\partial \theta^T}\frac{1}{L^2}. \tag{19}$$

Thus using $\partial\phi/\partial\theta = \phi\partial\log\phi/\partial\theta$ and $\partial^2\phi/\partial\theta\partial\theta^T = \phi(\partial^2\log\phi/\partial\theta\partial\theta^T + \partial\log\phi/\partial\theta\partial\log\phi/\partial\theta^T)$, we get

$$\frac{\partial \log L}{\partial \theta} = \frac{\int \frac{\partial \phi}{\partial \theta}}{\int \phi} = \frac{\int \frac{\partial \log \phi}{\partial \theta}\phi}{\int \phi},$$

$$\frac{\partial^2 \log L}{\partial \theta \partial \theta^T} = \frac{\int \frac{\partial^2 \phi}{\partial \theta \partial \theta^T}}{\int \phi} - \frac{\int \frac{\partial \phi}{\partial \theta}\phi \int \frac{\partial \phi}{\partial \theta^T}\phi}{(\int \phi)^2} = \frac{\int \frac{\partial^2 \log \phi}{\partial \theta \partial \theta^T}\phi + \int \frac{\partial \log \phi}{\partial \theta}\frac{\partial \log \phi}{\partial \theta^T}\phi}{\int \phi} - \frac{\int \frac{\partial \log \phi}{\partial \theta}\phi \int \frac{\partial \log \phi}{\partial \theta^T}\phi}{(\int \phi)^2}.$$

Let $A_1 = I - \delta W_1$, $A_2 = I - \psi W_2$. Then $Y^* = A_1^{-1}X\beta + A_1^{-1}A_2^{-1}\varepsilon$ with $\varepsilon \sim N(0, I)$. Hence the density of Y^* is $\phi = c|A_1^{-1}A_2^{-1}(A_2^T)^{-1}(A_1^T)^{-1}|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(Y^* - A_1^{-1}X\beta)^T A_1^T A_2^T A_2 A_1 (Y^* - A_1^{-1}X\beta)\}$ with c a constant. Then

$$\log \phi = \log|A_1| + \log|A_2| - \frac{1}{2}(Y^*)^T A_1^T A_2^T A_2 A_1 Y^* + \beta^T X^T A_2^T A_2 A_1 Y^* - \frac{1}{2}\beta^T X^T A_2^T A_2 X\beta + \log c.$$

Thus,

$$\frac{\partial \log \phi}{\partial \beta} = X^T A_2^T A_2 (A_1 Y^* - X\beta), \tag{20}$$

$$\frac{\partial \log \phi}{\partial \delta} = -\mathrm{tr}(A_1^{-1}W_1) + (Y^*)^T W_1^T A_2^T A_2 (A_1 Y^* - X\beta), \tag{21}$$

$$\frac{\partial \log \phi}{\partial \psi} = -\mathrm{tr}(A_2^{-1}W_2) + (A_1 Y^* - X\beta)^T A_1^T A_2^T W_2 (A_1 Y^* - X\beta), \tag{22}$$

$$\frac{\partial^2 \log \phi}{\partial \beta \partial \beta^T} = -X^T A_2^T A_2 X, \tag{23}$$

$$\frac{\partial^2 \log \phi}{\partial \beta \partial \delta} = -X^T A_2^T A_2 W_1 Y^*, \tag{24}$$

$$\frac{\partial^2 \log \phi}{\partial \beta \partial \psi} = -X^T W_2^T A_2 (A_1 Y^* - X\beta) - X^T A_2^T W_2 (A_1 Y^* - X\beta), \tag{25}$$

$$\frac{\partial^2 \log \phi}{\partial \delta^2} = -\mathrm{tr}(A_1^{-1}W_1 A_1^{-1}W_1) - (Y^*)^T W_1^T A_2^T A_2 W_1 Y^*, \tag{26}$$

$$\frac{\partial^2 \log \phi}{\partial \delta \partial \psi} = -(Y^*)^T W_1^T W_2^T A_2 (A_1 Y^* - X\beta) - (Y^*)^T W_1^T A_2^T W_2 (A_1 Y^* - X\beta), \tag{27}$$

$$\frac{\partial^2 \log\phi}{\partial\psi^2} \quad = \quad -\mathrm{tr}(A_2^{-1}W_2A_2^{-1}W_2) - (A_1Y^* - X\beta)^TW_2^TW_2(A_1Y^* - X\beta), \tag{28}$$

such that (using the short–hand $U = Y^* - X\beta$)

$$\frac{\partial \log\phi}{\partial\theta}_{\theta=[0,0,\beta^T]^T} = \begin{bmatrix} U^TW_2U \\ (Y^*)^TW_1U \\ X^TU \end{bmatrix},$$

and

$$-\frac{\partial^2 \log\phi}{\partial\theta\partial\theta^T}_{\theta=[0,0,\beta^T]^T} = \begin{bmatrix} \mathrm{tr}(W_2^2) + U^TW_2^TW_2U & U^TW_2W_1Y^* + U^TW_2^TW_1Y^* & U^T(W_2^T + W_2)X \\ \cdot & \mathrm{tr}(W_1^2) + (Y^*)^TW_1^TW_1Y^* & (Y^*)^TW_1^TX \\ \cdot & \cdot & X^TX \end{bmatrix},$$

Now, note that under independence

$$E[U_iU_j|Y_i, Y_j] = E[U_i|Y_i]E[U_j|Y_j] = \tilde{U}_i\tilde{U}_j.$$

Let $d_i(u_i) = I^{Y_i}(u_i \geq -x_i^T\beta)I^{1-Y_i}(u_i < -x_i^T\beta)$ and $d(u) = \prod_{i=1}^n d_i(u_i)$. Then

$$E[U_i|Y_i] = \frac{\int u_id_i(u_i)\phi(u_i)}{\int d_i(u_i)\phi(u_i)},$$

and therefore (ϕ_n being the n–variate standard normal density),

$$\frac{\partial \log L}{\partial\beta}_{\theta=[0,0,\beta^T]^T} = \frac{X^T\int u\phi_n(u)d(u)}{\int d(u)\phi_n(u)} = \sum_{i=1}^n \frac{x_i\int u_i\phi(u_i)d_i(u_i)}{\int d_i(u_i)\phi(u_i)} = X^T\tilde{U},$$

such that with $\tilde{Y}^* = X\beta + \tilde{U}$,

$$\frac{\partial \log L}{\partial\theta}_{\theta=[0,0,\beta^T]^T} = \begin{bmatrix} \tilde{U}^TW_2\tilde{U} \\ (\tilde{Y}^*)^TW_1\tilde{U} \\ X^T\tilde{U} \end{bmatrix}.$$

Further, note that

$$E\left[\frac{\partial^2 \log\phi}{\partial\theta\partial\theta^T}\right] = -\begin{bmatrix} T_{22} & T_{12} & 0 \\ T_{12} & T_{11} + ||W_1^TX\beta||^2 & \beta^TX^TW_1X \\ 0 & X^TW_1^TX\beta & X^TX \end{bmatrix},$$

where $T_{ij} = \mathrm{tr}(W_i(W_j + W_j^T)) = \sum_{ts} w_{its}(w_{jts} + w_{jst})$, for $i, j = 1, 2$. Also, note that

$$E\left[\frac{\partial \log \phi}{\partial \theta}\frac{\partial \log \phi}{\partial \theta^T}\right] = \begin{bmatrix} T_{22} & T_{12} & 0 \\ T_{12} & T_{11} + ||W_1^T X\beta||^2 & \beta^T X^T W_1 X \\ 0 & X^T W_1^T X\beta & X^T X \end{bmatrix} = -E\left[\frac{\partial^2 \log \phi}{\partial\theta\partial\theta^T}\right],$$

and that

$$E\left[\frac{\int \frac{\partial \log \phi}{\partial \theta}\phi}{\int \phi}\frac{\int \frac{\partial \log \phi}{\partial \theta^T}\phi}{\int \phi}\right] = \begin{bmatrix} T_{D;22} & T_{D;12} & 0 \\ T_{D;12} & T_{D;11} + ||D^{\frac{1}{2}} W_1^T X\beta||^2 & \beta^T X^T W_1 D X \\ 0 & X^T D W_1^T X\beta & X^T D X \end{bmatrix},$$

where $T_{D;ij} = \mathrm{tr}(W_i D(W_j + W_j^T)D)$, $i, j = 1, 2$, and D is a diagonal matrix with diagonal elements $d_i = E[\tilde{U}_i^2] = \phi_i^2/\{\Phi_i(1-\Phi_i)\}$.

B Proofs

Let $A(t) = \phi(t)/\Phi(t)$, $A_i = \phi_i/\Phi_i$, $\hat{A}_i = \hat{\phi}_i/\hat{\Phi}_i$ and $A_i^* = \phi(x_i^T\beta^*)/\Phi(x_i^T\beta^*)$ whenever β^* is defined. Further, let $A_\cdot^0$ denote the same quantities with Φ replaced with $1-\Phi$.

Lemma 1 $\sup_i |\hat{U}_i - \tilde{U}_i| = o_p(n^{-\frac{1}{4}})$.

Proof: Note that $\tilde{U}_i = \phi_i\{Y_i/\Phi_i - (1-Y_i)/(1-\Phi_i)\}$. I show that $Q = \sup_i |\hat{A}_i - A_i| = o_p(n^{-\frac{1}{4}})$ where the remainder can be established similarly. By the mean value theorem $A(t_1) - A(t_2) = -(tA(t) + A^2(t))(t_1 - t_2)$, for some t between t_1 and t_2. Hence it suffices to show that $Q^* = \sup_{i;\beta^*\in\Theta} |\{x_i^T\beta^* A(x_i^T\beta^*) - A^2(x_i^T\beta^*)\}x_i^T(\hat{\beta} - \beta)| = o_p(n^{-\frac{1}{4}})$. Now $\hat{\beta} - \beta = O_p(n^{-\frac{1}{2}})$. Further $\sup_t |A(t)/(|t|+1)| < 1$. Therefore Q^* is bounded by $O_p(n^{-\frac{1}{2}})\sup_{i;\beta^*\in\Theta}\{|x_i^T\beta^*|(|x_i^T\beta^*|+1) + |x_i^T\beta^*|^2\}||x_i||$. Since Θ is bounded and $\sup_i ||x_i|| = o(n^{\frac{1}{12}})$ by Condition 3, the result holds. **Q.E.D.**

Lemma 2 $\sup_i |\hat{U}_i - U_i + \{(x_i^T\beta A_i + A_i^2)Y_i - (x_i^T\beta A_i^0 - (A_i^0)^2)(1-Y_i)\}x_i^T(\hat{\beta} - \beta)| = o_p(n^{-\frac{1}{2}})$.

Proof: Like in Lemma 1 we only consider the observations for which $Y_i = 1$, where the result for $Y_i = 0$ can be established similarly. Note that by the mean value theorem $A(t_1) - A(t_2) = -\{t_2 A(t_2) + A^2(t_2)\}(t_1 - t_2) + \{(t^2-1)A(t) + 3tA^2(t) + 2A^3(t)\}(t_1-t_2)^2$ for some t between t_1 and t_2. Take $t_1 = x_i^T\hat{\beta}$ and $t_2 = x_i^T\beta$. Now, $\sup_{i,\beta^*\in\Theta} |\{((x_i^T\beta^*)^2 - 1)A_i^* + 3x_i^T\beta^*(A_i^*)^2 + 2(A_i^*)^3\}\{x_i^T(\hat{\beta} - \beta)\}^2$.

Note that $||\hat{\beta}-\beta||^2 = O_p(n^{-1})$ and that Θ is bounded. Since $\sup_t |A(t)/(|t|+1)| < 1$, the result follows from $\sup_i ||x_i|| = o(n^{\frac{1}{12}})$. Indeed, $\sup_i ||x_i|| = o(n^{\frac{1}{5}})$ would have sufficed here. **Q.E.D.**

Lemma 3 $T_{22}^{-\frac{1}{2}} \sum_{ij} w_{2ij}(\hat{U}_i - \tilde{U}_i)(\hat{U}_j - \tilde{U}_j) = o_p(1)$.

Proof: Note that the LHS is bounded by $T_{22}^{-\frac{1}{2}}(\sup_i |\hat{U}_i - \tilde{U}_i|)^2 \sum_{ij} |w_{2ij}| = o_p(n^{-\frac{1}{2}}) T_{22}^{-\frac{1}{2}} \sum_{ij} |w_{2ij}| = o_p(1)$ by Lemma 1 and Condition 2. **Q.E.D.**

Lemma 4 $T_{22}^{-\frac{1}{2}} \sum_{ij} w_{2ij} U_i(\hat{U}_j - \tilde{U}_j) = o_p(1)$.

Proof: Let $B_i = -(x_i^T \beta A_i + A_i^2)Y_i + (x_i^T \beta A_i^0 - (A_i^0)^2)(1-Y_i)$. Then $\hat{U}_j - \tilde{U}_j = B_j(\hat{\beta}-\beta) + R_j$ for implicitly defined R_j. I will show that (i) $Q_1 = T_{22}^{-\frac{1}{2}} \sum_{ij} w_{2ij} \tilde{U}_i R_j = o_p(1)$ and that (ii) $Q_2 = T_{22}^{-\frac{1}{2}} \sum_{ij} w_{2ij} \tilde{U}_i B_j = o_p(n^{\frac{1}{2}})$, which are sufficient since $\hat{\beta} - \beta = O_p(n^{-\frac{1}{2}})$. First (i). Note that $\sup_j |R_j| = o_p(n^{-\frac{1}{2}})$ by Lemma 2. So (i) is implied by $T_{22}^{-\frac{1}{2}} \sum_j |\sum_i w_{2ij} \tilde{U}_i| = O_p(n^{\frac{1}{2}})$. Note that $E|\sum_i w_{2ij} U_i| \leq \{\sum_i w_{2ij}^2 E[\tilde{U}_i^2]\}^{\frac{1}{2}}$ by the Schwarz inequality. Note that $E[\tilde{U}_i^2] = A_i A_i^0$ and is bounded uniformly in i, which follows from the fact that if $A_i > A_i^0$ then $A_i^0 \leq 2\phi_i$ and $\phi(t)t$ is uniformly bounded in absolute value. Hence for (i) to hold we need $T_{22}^{-\frac{1}{2}} \sum_j \{\sum_i w_{2ij}^2\}^{\frac{1}{2}} = O(n^{\frac{1}{2}})$, which follows with the Schwarz inequality and the condition that the w_{2ij}'s are summable across i or j. Now (ii). Note that for $i \neq j$, $\tilde{U}_i$ is independent of B_j and for $i = j$, $w_{2ij} = 0$. Thus, $E||Q_2||^2 = T_{22}^{-1} \sum_{ijkl} w_{2ij} w_{2kl} E[\tilde{U}_i B_j^T \tilde{U}_k B_l]$. If i,j,k,l are all different the expectation is zero. Indeed, when $i \neq k, l$ or $k \neq i, j$ the expectation is zero. I thus have to deal with two cases: $i = k$ and $i = l, j = k$. The second case is of lower order than the first so we only consider the first. I get $T_{22}^{-1} \sum_{ijl} w_{2ij} w_{2il} E[\tilde{U}_i^2 B_j^T B_l]$. Since $\sup_i E[\tilde{U}_i^4] + \sup_i E||B_i||^4$ is bounded uniformly in n (which is easily verified), Now, $T_{22}^{-1} \sum_{ijl} w_{2ij} w_{2il} = \sum_i (\sum_j w_{2ij})^2 / T_{22} = o(n)$ by Condition 2. **Q.E.D.**

Lemma 5 $T_{22}^{-\frac{1}{2}}(\hat{U}^T W_2 \hat{U} - \tilde{U}^T W_2 \tilde{U}) = o_p(1)$.

Proof: Since $\hat{U}^T W_2 \hat{U} - \tilde{U}^T W_2 \tilde{U} = (\hat{U} - \tilde{U})^T W_2 (\hat{U} - \tilde{U}) + \tilde{U}^T W_2 (\hat{U} - \tilde{U}) + (\hat{U} - \tilde{U})^T W_2 \tilde{U}$, the result follows immediately from Lemmas 3 and 4. **Q.E.D.**

Lemma 6 $\sup_i |\hat{d}_i - d_i| = o_p(1)$.

Proof: Note that $d_i = A_i A_i^0$. In Lemma 1 we established that $\sup_i |\hat{A}_i - A_i| = o_p(n^{\frac{1}{4}})$. **Q.E.D.**

Proof of Theorem 1

A mild generalization of the Cliff–Ord (1973) result has that $\tilde{U}^T W_2 \tilde{U} / T_{22} \xrightarrow{\mathcal{L}} N(0,1)$. Hence we need to show that $T_{22}^{-\frac{1}{2}}(\hat{U}^T W_2 \hat{U} - \tilde{U}^T W_2 \tilde{U}) = o_p(1)$ and that $T_{22}^{-1}(\hat{T}_{22} - T_{22}) = o_p(1)$. The former is established in Lemma 5, the latter follows directly from Lemma 6.

Q.E.D.

C Tables

C.1 Size

Rejection rates under the null hypothesis at a 5% significance level.

$k=$	$n=$	RSD STD	RSD ALT	STRCT STD	STRCT ALT	JOINT STD	JOINT ALT
3	100	4.86	5.04	4.63	5.36	3.95	5.87
	256	4.86	5.34	4.74	5.38	3.70	5.58
6	100	4.98	5.02	4.91	5.30	5.21	6.04
	256	4.77	5.11	4.80	5.33	4.37	5.39

Table 1: First Order Spatial Autocorrelation Weighting Matrix, Replications=10000

$k=$	$n=$	RSD STD	RSD ALT	STRCT STD	STRCT ALT	JOINT STD	JOINT ALT
3	100	4.32	4.64	4.32	5.00	3.26	5.60
	256	4.90	5.27	4.47	5.22	3.58	6.04
6	100	4.62	4.58	4.80	4.98	4.45	5.95
	256	4.84	5.22	4.75	5.38	4.14	5.66

Table 2: Inversely related to third power of distance, Replications=10000

$k=$	$n=$	RSD STD	RSD ALT	STRCT STD	STRCT ALT	JOINT STD	JOINT ALT
3	100	3.87	4.16	3.42	4.38	2.72	5.05
	256	4.41	4.88	3.59	4.95	2.78	5.77
6	100	3.89	3.94	3.91	4.66	3.75	5.65
	256	4.41	4.80	4.40	4.95	3.87	5.84

Table 3: Inversely related to second power of distance, Replications=10000

$k =$	$n =$	RSD STD	RSD ALT	STRCT STD	STRCT ALT	JOINT STD	JOINT ALT
3	100	0.00	0.00	0.00	1.93	0.00	100.00
	256	0.00	0.00	0.00	4.12	0.00	47.31
6	100	0.00	0.00	0.00	2.01	0.00	99.99
	256	0.00	0.00	0.00	3.50	0.00	46.76

Table 4: Fully Symmetric, Replications=10000

C.2 Power

Rejection rates at a 5% significance level.

				Residual		Structural		Both	
$k =$	$n =$	$\psi =$	$\delta =$	STD	ALT	STD	ALT	STD	ALT
3	100	0.1		4.8	2.9	5.4	4.1	5.1	3.8
6	100	0.1		4.4	0.8	5.2	1.8	4.7	1.5
3	100	0.5		31.3	21.6	16.5	16.4	25.7	17.1
6	100	0.5		17.1	5.8	9.1	5.3	15.3	5.1
3	100		0.1	4.8	3.0	7.1	5.2	6.3	4.5
6	100		0.1	4.4	0.9	9.1	3.7	7.5	2.6
3	100		0.5	36.7	27.1	79.0	70.1	70.1	60.8
6	100		0.5	24.0	9.6	86.0	70.5	78.2	59.1
3	100	0.1	0.1	6.8	4.0	10.0	7.1	7.9	5.3
6	100	0.1	0.1	4.8	1.0	10.0	4.1	7.4	2.8
3	100	0.5	0.5	87.3	81.5	93.9	93.0	92.3	88.4
6	100	0.5	0.5	70.4	52.9	91.5	86.5	89.1	77.7

Table 5: First Order Spatial Autocorrelation Weighting Matrix, Replications=10000

$k =$	$n =$	$\psi =$	$\delta =$	Residual		Structural		Both	
				STD	ALT	STD	ALT	STD	ALT
3	100	0.1	0.0	5.6	3.1	5.3	4.0	5.7	3.7
6	100	0.1	0.0	4.9	1.0	5.5	2.1	5.0	1.7
3	100	0.5	0.0	42.0	31.1	21.9	23.1	36.1	26.0
6	100	0.5	0.0	24.5	9.4	11.9	8.1	21.7	8.1
3	100		0.1	5.9	3.3	8.5	6.0	7.3	5.1
6	100		0.1	4.9	1.0	10.3	4.3	7.8	2.6
3	100		0.5	56.4	47.7	86.6	81.9	80.4	74.6
6	100		0.5	39.3	19.2	93.1	83.5	88.5	73.9
3	100	0.1	0.1	8.2	5.4	11.5	9.6	9.5	7.0
6	100	0.1	0.1	6.3	1.4	11.9	5.7	9.5	3.5
3	100	0.5	0.5	97.5	95.9	98.1	98.2	98.2	97.0
6	100	0.5	0.5	86.8	73.1	96.3	94.3	95.8	89.9

Table 6: Inversely related to third power of distance, Replications=10000

$k =$	$n =$	$\psi =$	$\delta =$	Residual		Structural		Both	
				STD	ALT	STD	ALT	STD	ALT
3	100	0.1	0.0	4.4	3.0	5.4	3.9	4.9	3.7
6	100	0.1	0.0	4.1	0.8	5.5	2.0	4.7	1.6
3	100	0.5	0.0	27.4	18.7	16.3	15.7	24.8	17.8
6	100	0.5	0.0	16.3	5.5	9.9	5.8	15.3	5.6
3	100		0.1	5.2	2.9	7.2	5.1	6.2	4.3
6	100		0.1	4.3	0.9	8.0	3.5	6.2	2.2
3	100		0.5	37.8	28.8	65.8	58.9	56.9	50.1
6	100		0.5	31.2	15.1	72.0	58.8	64.0	47.6
3	100	0.1	0.1	6.2	3.4	8.6	6.3	7.2	5.0
6	100	0.1	0.1	4.8	0.9	8.6	4.0	6.9	2.3
3	100	0.5	0.5	85.4	78.5	86.4	86.4	87.4	82.7
6	100	0.5	0.5	71.3	50.0	82.0	76.2	82.2	67.1

Table 7: Inversely related to second power of distance, Replications=10000

$k =$	$n =$	$\psi =$	$\delta =$	Residual		Structural		Both	
				STD	ALT	STD	ALT	STD	ALT
3	100	0.1	0.0	0.0	0.1	0.0	1.1	0.0	64.4
6	100	0.1	0.0	0.0	0.0	0.0	1.9	0.0	74.8
3	100	0.5	0.0	0.0	0.1	0.0	1.4	0.0	70.7
6	100	0.5	0.0	0.0	0.0	0.0	2.6	0.0	78.1
3	100		0.1	0.0	0.0	0.0	1.3	0.0	66.4
6	100		0.1	0.0	0.0	0.0	2.0	0.0	76.2
3	100		0.5	0.0	0.0	0.0	3.3	0.0	78.3
6	100		0.5	0.0	0.0	0.0	5.8	0.0	86.2
3	100	0.1	0.1	0.0	0.0	0.0	1.0	0.0	65.8
6	100	0.1	0.1	0.0	0.0	0.0	2.0	0.0	76.8
3	100	0.5	0.5	0.0	0.0	0.0	6.1	0.0	84.3
6	100	0.5	0.5	0.0	0.0	0.0	7.8	0.0	87.8

Table 8: Fully Symmetric, Replications=10000