NBER WORKING PAPER SERIES

SPATIAL ERRORS IN COUNT DATA REGRESSIONS

Marinho Bertanha
Petra Moser

Working Paper 20374
http://www.nber.org/papers/w20374

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2014

We thank Han Hong, Guido Imbens, Aprajit Mahajan, and Scott Stern for helpful conversations and comments. Moser gratefully acknowledges financial support through NSF CAREER grant 1151180. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Spatial Errors in Count Data Regressions
Marinho Bertanha and Petra Moser
NBER Working Paper No. 20374
August 2014
JEL No. C23,C33,O3

ABSTRACT

Count data regressions are an important tool for empirical analyses ranging from analyses of patent counts to measures of health and unemployment. Along with negative binomial, Poisson panel regressions are a preferred method of analysis because the Poisson conditional fixed effects maximum likelihood estimator (PCFE) and its sandwich variance estimator are consistent even if the data are not Poisson-distributed, or if the data are correlated over time. Analyses of counts may be affected by correlation in the cross-section. For example, patent counts or publications may increase across related research fields in response to common shocks. This paper shows that the PCFE and its sandwich variance estimator are consistent in the presence of such dependence in the cross-section - as long as spatial dependence is time-invariant. In addition to the PCFE, this result also applies to the commonly used Logit model of panel data with fixed effects. We develop a test for time-invariant spatial dependence and provide code in STATA and MATLAB to implement the test.

Marinho Bertanha
Department of Economics
Stanford University
579 Serra Mall
Stanford, CA 94305-6072
bertanha@stanford.edu

Petra Moser
Department of Economics
Stanford University
579 Serra Mall
Stanford, CA 94305-6072
and NBER
pmoser@stanford.edu

Online supplements STATA and MATLAB commands to implement estimation procedures described in the paper are available at:
www.stanford.com/~bertanha/xtpsse_stata.zip www.stanford.com/~bertanha/xtpsse_matlab.zip

1 Introduction

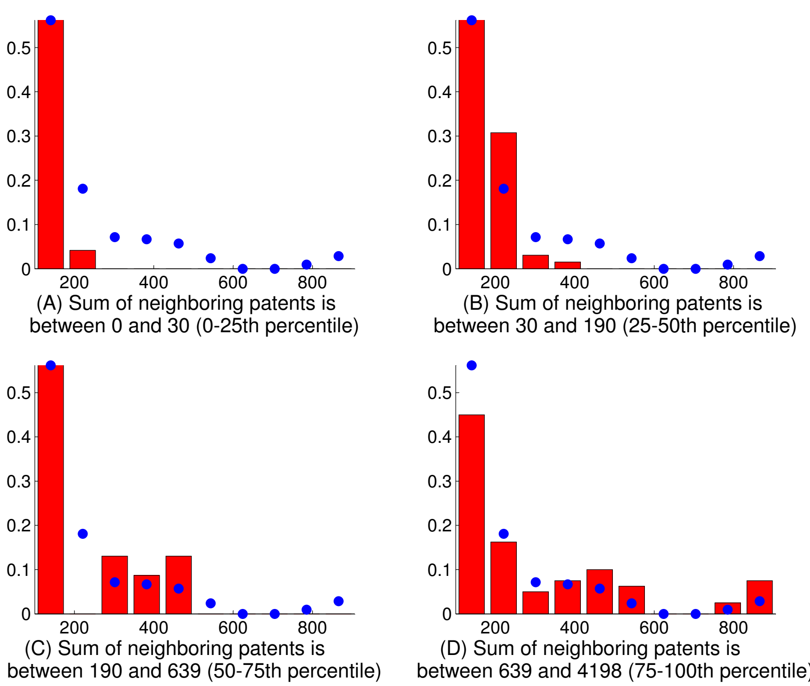
Empirical analyses in economics and other disciplines rely heavily on regressions of count data, such as patents, citations, publications, and counts of accidents or unemployed workers. Along with the negative binomial estimator, the Poisson Conditional Fixed Effects Maximum Likelihood Estimator (PCFE) has become the standard approach to analyze count data. A key advantage is that the PCFE is robust to misspecifications of the underlying distribution as long as the conditional mean is correctly specified (Wooldridge (1999), and Gourieroux, Monfort and Trognon (1984)). For example, the PCFE is robust even if the variance of the distribution is different from the mean (violating the assumption of equidispersion), or if the data includes an excessive number of zeros. Another key benefit is that the PCFE is robust to dependence over time. For example, counts of patents or unemployed workers may be correlated over time and the PCFE is robust to such correlation.

Count data can, however, also be affected by cross-sectional dependence. For example, Jaffe, Trajtenberg and Henderson (1993) show that knowledge spillovers are geographically localized. Patent data analyzed by Hall, Griliches and Hausman (1986) show evidence of spatial dependence. Figure 1 depicts patterns of spatial dependence for patents by firms with more than 100 patents per year. These firms may benefit from localized knowledge spillovers (as described by Jaffe, Trajtenberg and Henderson (1993)) or from shared access to human capital and other inputs to invention (Marshall (1890)).

Counts of patents, publications, and other measures for innovation may also be correlated in idea space - in addition to geographic space - if knowledge spillovers and scientific breakthroughs encourage innovation across related fields. This type of spatial dependence in idea space is likely to affect a growing literature (e.g. Azoulay, Zivin and Wang (2010), Furman and Stern (2011), Moser and Voena (2012), Borjas and Doran (2012), Williams (2013), Kogan, Papanikolaou, Seru and Stoffman (2012), and Moser, Voena and Waldinger (2014)).

This paper investigates whether whether spatial dependence (or correlation in the cross-section) threatens statistical inference from Poisson estimates of count data. To account for spatial dependence, we specify the likelihood scores of the PCFE as moment conditions in Conley (1999)'s spatial estimation framework. We find that the PCFE is consistent and asymptotically normal. The asymptotic variance of the PCFE is generally different from the variance that assumes spatial independence. To address this issue, we present a consistent spatial variance estimator for the case of spatial dependence. In comparison with clustering as a standard approach to address cross-sectional dependence, the spatial variance estimator is more general than clustering because it allows for dependence between any pair of individuals while clustering only allows for dependence within groups.[1]

Figure 1: Patents per Firm Conditional on Patents by Neighboring Firms



Notes: Firm A is defined to be a *neighbor* of firm B if its distance in terms of latitude and longitude is less or equal to the 10th quantile of distances in the sample; results are robust to using the 25th or 50th quantile as a cutoff to define neighbors. Dots plot the unconditional distribution of patents per year by firms with more than 100 patents per year. Bars plot the distribution of patents per year for firms whose neighbors produce few patents (sum of neighboring patents is between 0 and the 25th quantile, Panel A), medium level of patents (25-50th and 50-75th quantiles, Panels B and C) and many patents (75-100th quantile, Panel D).

[1] For a discussion of clustering see Cameron and Trivedi (2005), section 24.5.6, or STATA 13 User's Guide, section 20.21.2.

Allowing for spatial dependence changes the variance of the PCFE if covariances of the likelihood score function between observations in the cross-section are nonzero. We also show that covariances of the likelihood scores are zero - and the standard sandwich estimator of the PCFE is consistent - as long as spatial dependence is time-invariant.For example, firm patent applications may be spatially correlated because of unobservable shocks that affect the local supply of high-skilled workers such as cultural goods, high-quality restaurants or schools. If access to these amenities is constant during the sample period, then the structure of spatial dependence is time-invariant.

More formally, spatial dependence is time-invariant if unobservable shocks can be factored into a time-variant and a time-invariant component, and only the time-invariant component generates spatial dependence. Time-invariant spatial dependence does not affect the asymptotic variance of the PCFE because the source of spatial dependence are unobservable time-invariant components that are not separately identified from the unobservable fixed effects of the PCFE model.

The fixed effects of the PCFE model are modified to absorb time-invariant components of unobservable shocks that cause spatial dependence. With this modification, covariances of the likelihood score between different observations in the cross-section are zero. As a result, the asymptotic variance of the PCFE under time-invariant spatial dependence is equal to the variance under spatial independence, and the standard sandwich variance estimator is consistent. In addition to the PCFE, this result also applies to the commonly used Logit model of panel data with fixed effects.

To detect spatial dependence that is not time-invariant, a test-statistic is constructed based on the coefficients of a regression of the likelihood score of each individual on an average of the likelihood scores of neighboring individuals. Under the null hypothesis of time-invariance, the estimated parameters of this regression should be close to zero, because the covariance of the likelihood scores between different individuals is zero. The test-statistic is a Wald test for the restriction that these parameters are jointly equal to zero.

We illustrate the PCFE, its variance estimators, and the test statistic by applying them to Hall, Griliches and Hausman (1986)'s analysis of the link between firms' expenditures on research and patent counts, as a measure of research output. We also apply our approach to Furman and Stern (2011)'s analysis of the effects of a biological research center on cumulative research output, measured by count data on scientific citations. We provide new commands in STATA and MATLAB that compute the PCFE, its variance estimators, and the test statistic.[2]

The remainder of this paper is organized as follows. In Section 2, we summarize existing robustness results for the PCFE that rely on the assumption of spatial independence. Section 3 derives the asymptotic distribution of the PCFE under spatial dependence, and provides a consistent spatial estimator for the variance of the PCFE. We show that the sandwich variance estimator is consistent for the variance of the PCFE if spatial dependence is time-invariant. Section 4 constructs a test-statistic and derive its distribution under the null hypothesis of time-invariance. Section 5 illustrates the estimator with the examples of Hall, Griliches and Hausman (1986) as well as Furman and Stern (2011). An appendix presents proofs for the results in this paper and provides additional details about the code in MATLAB and STATA.

2 PCFE under Spatial Independence

This section introduces the PCFE model of Hausman, Hall and Griliches (1984) (HHG) and summarizes existing robustness results that rely on the assumption of spatial independence. We revisit these results in the next section where the assumption of spatial independence is relaxed. For $i = 1, \dots N$ (cross-sectional units), and $t = 1, \dots T$ (time periods), denote the random variables: $y_i = [y_{i1} \dots y_{iT}]'$ a $T \times 1$ vector of count dependent variables; $x_i = [x_{i1} \dots x_{iT}]'$ a $T \times K$ matrix of explanatory variables; $\phi = [\phi_1 \dots \phi_N]'$ a $N \times 1$ vector of

[2] Available at *www.stanford.com/˜bertanha/xtpsse_stata.zip*; and *www.stanford.com/˜bertanha/xtpsse_matlab.zip*

strictly positive individual fixed effects that are unobserved.[3] To compute asymptotic distributions, we let $N \to \infty$ while T remains fixed. The likelihood function is derived assuming observations are independent across i and t, and $y_{it}|x_i, \phi_i \sim iid$ Poisson $(\phi_i \exp(x_{it}'\beta_0))$. Unlike the Poisson random effects model, here individual fixed effects and explanatory variables can be dependent or independent.

Using the argument of Andersen (1972), HHG conditions the Poisson distribution above on the sum across t of y_{it}, $n_i = \sum_{t=1}^{T} y_{it}$. This leads to a likelihood function that does not depend on the unobserved ϕ. Following Wooldridge (1999), the pseudo log-likelihood function is $L_N(\beta) = \frac{1}{N}\sum_{i=1}^{N} \ell_i(\beta)$, where $\ell_i(\beta) = \sum_{t=1}^{T} y_{it} \log[p_t(x_i, \beta)]$, and $p_t(x_i, \beta) = \exp(x_{it}'\beta)/\sum_{k=1}^{T} \exp(x_{ik}'\beta)$. The maximum likelihood estimator is the zero of the score function[4] $S_N(\beta) = \frac{1}{N}\sum_{i=1}^{N} S_i(\beta)$ where

$$
\begin{aligned}
S_i(\beta) &= \nabla_\beta p(x_i, \beta)' W(x_i, \beta) u_i(\beta) \\
p(x_i, \beta) &= [p_1(x_i, \beta), \ldots, p_T(x_i, \beta)]' \\
W(x_i, \beta) &= [diag\{p_t(x_i, \beta)\}_t]^{-1} \\
u_i(\beta) &= y_i - p(x_i, \beta) n_i
\end{aligned}
$$

Using standard results for the asymptotics of maximum likelihood estimators,

[3] Despite being random, the ϕ_i's are called fixed effects because the Poisson distribution assumption is made conditionally on the ϕ_i's.

[4] HHG shows global concavity of this likelihood function for a compact parameter space. Note that we cannot compute $\hat{\beta}$ if there exist explanatory variables that are constant over time, or any linear combination of the explanatory variables that is constant over time.

$$\sqrt{N}\left(\hat{\beta}-\beta_0\right)\xrightarrow{d}N\left(0;A_0^{-1}B_0A_0^{-1}\right)$$

$$\text{where} \tag{1}$$

$$A_0=E\left[\nabla_\beta S_i\left(\beta_0\right)\right]$$

$$B_0=VAR\left[S_i\left(\beta_0\right)\right]$$

If the model is correctly specified, then $A_0=-B_0$, and the variance in the sandwich form simplifies to B_0^{-1} or $-A_0^{-1}$. An important reason for the widespread use of PCFE to analyze count data lies in its robustness to misspecification of the Poisson distribution. Empirically, the main concern is that count data typically do not satisfy the Poisson's assumption of equidispersion, i.e. the variance of the distribution is equal to the mean. Another issue is that count data typically include a larger number of zeros than the best fitted Poisson model would predict.

There are two important robustness results in the literature for the Poisson model that assume spatial independence. In the case of cross-sectional data with no fixed effects, Gourieroux, Monfort and Trognon (1984) show that consistency and asymptotic normality only relies on the correct specification of the conditional mean. Wooldridge (1999) extends this robustness result for the case of panel data with fixed effects, where the conditional mean is assumed to be correctly specified as

$$E[y_{it}|x_i,\phi_i]=\phi_i\exp\left(x_{it}'\beta_0\right) \tag{2}$$

In other words, Wooldridge (1999) proves that the PCFE proposed by HHG is consistent and asymptotically $N\left(0;A_0^{-1}B_0A_0^{-1}\right)$ under misspecification of the Poisson distribution as long as the conditional mean is correctly specified. If the Poisson distribution is misspecified, A_0 may be different than $-B_0$, and we cannot simplify the sandwich form of the variance

$A_0^{-1}B_0A_0^{-1}$. This is the reason why current applied research estimates $A_0^{-1}B_0A_0^{-1}$ instead B_0^{-1} or $-A_0^{-1}$. Using $\hat{A} = \frac{1}{N}\sum_{i=1}^{N} \nabla_\beta S_i(\hat{\beta})$, and $\hat{B} = \frac{1}{N}\sum_{i=1}^{N} S_i(\hat{\beta})S_i(\hat{\beta})'$ as consistent estimators for A_0 and B_0, the so called 'sandwich variance estimator' is consistent for the asymptotic variance of PCFE under misspecification of the Poisson distribution:

$$\hat{A}^{-1}\hat{B}\hat{A}^{-1} \tag{3}$$

This paper relaxes the assumption of independence across individuals in the cross-section of the PCFE model. Many types of count-data regressions are affected by spatial dependence, for example, in geographic or idea space. For example, counts of patent applications by a firm may be correlated to patent applications by neighboring firms if knowledge spillovers are geographically localized (Jaffe, Trajtenberg and Henderson (1993)). Knowledge spillovers can also influence patent counts of similar research fields, generating spatial dependence in idea space. In the next section, we show that spatial dependence changes the distribution of $\hat{\beta}$ given in (1) above. Standard-errors that are computed based on the distribution in (1) may lead to invalid inferences in the presence of spatial dependence.

3 PCFE under Spatial Dependence

In this section we apply Conley (1999)'s results to derive the asymptotic distribution of the PCFE under spatial dependence. We show that this asymptotic distribution is the same as the asymptotic distribution that assumes spatial independence (1) as long as spatial dependence is time-invariant. This is a new robustness result for the PCFE. In the next section, we construct a test-statistic and derive its distribution under the null hypothesis of time-invariant spatial dependence.

Using the PCFE score function as a moment condition, consistency and asymptotic normality follow from Conley (1999)'s spatial GMM estimator. In the following paragraph, we introduce definitions that are central to our current analysis; see Conley (1999) for a more detailed discussion.[5]

From now on, we substitute the i index in any random variable by $s_i \in \mathbb{R}^2$, which means that each unit in the cross-section has an Euclidean coordinate locating it in the $\mathbb{R}^2$ space (this can also be generalized to $\mathbb{R}^k$). We assume that these coordinates are observed for all units in the sample. The subscript s represents a coordinate in $\mathbb{R}^2$, and p, a coordinate in $\mathbb{Z}^2$.

It is assumed that observations are at least $d_0 > 0$ apart from each other. To better index observations, the $\mathbb{R}^2$ is covered by a regular lattice. Construct a grid with $d^* \times d^*$ identical squares with a diagonal that is less than d_0. This implies that there is always at most one point in each square. Let D_1 and D_2 denote the number of lines and columns, respectively, of the largest rectangular grid with $d^* \times d^*$ squares that contains the observed sample. Label the grid lines using integers and label each square by its southwest corner coordinate, a vector $p \in \mathbb{Z}^2$. Not every square in this grid necessarily has an observation, that is, $N \leq D_1 D_2$. The probability of observing a realization in any square is given by λ, and this event is independent of everything else and across different squares. Let $Y_p(\beta) = S_s(\beta)$ if s is observed in square p, but $Y_p(\beta) = 0$ otherwise. Using Conley (1999)'s central limit theorem, if the conditional mean (eq. 2) is correctly specified, then:

$$\sqrt{N}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N\left(0, A_0^{-1} C_0 A_0^{-1}\right) \tag{4}$$

[5] Lemma 2 in the appendix shows that the PCFE score function satisfies Conley (1999)'s identification assumption; that is, β_0 is the unique solution to $E[S_i(\beta)] = 0$.

where

$$A_0 = \lambda^{-1} E[\nabla_\beta Y_p(\beta_0)]$$

$$V_0 = \lim_{D_1, D_2 \to \infty} (D_1 D_2)^{-1} \sum_{p,q} COV\left[Y_p\left(\beta_0\right), Y_q\left(\beta_0\right)\right] \text{ non-singular}$$

$$C_0 = \lambda^{-1} V_0$$

In this setting, $\hat{A} = N^{-1} \sum_i \nabla_\beta S_{s_i}(\hat{\beta})$ is a consistent estimator for the Hessian matrix A_0. The consistent estimator $\hat{C}$ for C_0 is an weighted average of sample spatial covariances which is generally different than $\hat{B}$. Like other covariance estimators, we need to cut off the number of maximum lags we use when computing covariances between all possible pairs of individuals. We denote the vertical and horizontal bandwidths as L_{D_1} and L_{D_2}. These are assumed to go to infinity more slowly than the sample size.

Define

$$\hat{C} = \frac{1}{N} \sum_{j=-L_{D_1}+1}^{L_{D_1}-1} \sum_{k=-L_{D_2}+1}^{L_{D_2}-1}$$

$$\sum_{\substack{m=1 \\ \text{st } 1 \leq m+j \leq D_1}}^{D_1} \sum_{\substack{n=1 \\ \text{st } 1 \leq n+k \leq D_2}}^{D_2} K_{D_1,D_2}(j,k) \left[Y_{m,n}\left(\hat{\beta}\right) Y_{m+j,n+k}\left(\hat{\beta}\right)' \right]$$

where we used the notation $p = (m, n) \in \mathbb{Z}^2$ and

$$K_{D_1,D_2}(j,k) = \begin{cases} \left(1 - |j| / L_{D_1}\right)\left(1 - |k| / L_{D_2}\right) & , \text{ for } |j| < L_{D_1}, \ |k| < L_{D_2} \\ 0 & , \text{ o.w.} \end{cases}$$

Under additional assumptions, Conley (1999) shows that $\hat{C} \xrightarrow{p} C_0$. Therefore, a consistent estimator for the asymptotic variance of the PCFE under spatial dependence is given by

$$\hat{A}^{-1}\hat{C}\hat{A}^{-1} \tag{5}$$

We call this estimator the 'spatial variance estimator'.

Our approach allows for spatial dependence, whereas HHG and Wooldridge (1999) assume spatial independence. The only restrictions that Conley (1999) imposes on the nature of spatial dependence are stationarity and mixing conditions. Although accounting for spatial dependence preserves the consistency and asymptotic normality of the PCFE, estimates of the variance may be affected by spatial dependence, such that inference based on the sandwich variance estimator may be misleading.

Comparing the distribution of $\hat{\beta}$ under spatial dependence (4) to the distribution of $\hat{\beta}$ under spatial independence (1), the difference arises in the expression for the variance of the sum of the individual score functions (C_0 vs B_0). Under spatial dependence, this variance contains covariances between scores of different individuals. If the covariance of the score function of any two different individuals is zero, then $C_0 = \lambda^{-1}E[Y_p(\beta_0)Y_p(\beta_0)'] = B_0$, and $\hat{B} = N^{-1}\sum_i \nabla_\beta S_{s_i}(\hat{\beta})\nabla_\beta S_{s_i}(\hat{\beta})'$ is a consistent estimator for $B_0 = C_0$. The asymptotic results for the PCFE given in (1) along with its sandwich variance estimator continue to be valid under spatial dependence if $C_0 = B_0$. If the covariance of the score function is not zero for some pairs of individuals, then C_0 can be different than B_0, and the sandwich variance estimator can be inconsistent for the variance of the PCFE.

We define the unobserved shock vector $\varepsilon_s = [\varepsilon_{s,1}, \cdots, \varepsilon_{s,T}]'$ such that $\varepsilon_{s,t} \equiv y_{s,t}/\left(\phi_s \exp(x_{s,t}'\beta_0)\right)$ for $s \in \mathbb{R}^2$ and $t = 1, \ldots, T$. Given the observed data $\{y_s, x_s\}_s$ and the conditional mean assumption (eq. 2), we can identify β_0 and the joint distribution of $\{y_s, x_s, \phi_s\varepsilon_s\}_s$ but we cannot identify the distribution of ϕ_s and ε_s separately. There can be different $\widetilde{\varepsilon}_{s,t}$ and $\widetilde{\phi}_s$ such that $\widetilde{\phi}_s\widetilde{\varepsilon}_{s,t} = \phi_s\varepsilon_s$ and the conditional mean equation holds for β_0 and $\widetilde{\phi}_s$. It is important to emphasize that the spatial dependence of unobserved shocks conditional on explanatory variables x_s and fixed effects ϕ_s is what makes C_0 and B_0 different. However, if there is spatial dependence on $\varepsilon_{s,t}$ conditional on (x_s, ϕ_s), but no spatial dependence on $\widetilde{\varepsilon}_{s,t}$ conditional on $(x_s, \widetilde{\phi}_s)$, then C_0 and B_0 will be the same. This is the case when the spatial dependence on $\{\varepsilon_s\}_s$ is time-invariant.

Definition 1. *The spatial dependence on $\{\varepsilon_s\}_s$ is said to be 'time-invariant' if there exists $c > 0$ and $\{\{u_{s,t}\}_t, \eta_s\}_s$ strictly positive scalar random variables such that*

$$\varepsilon_{s,t} = \eta_s u_{s,t}$$

with $u_s \perp (u_{s'}, x_{s'}, \phi_{s'}\eta_{s'})|x_s, \phi_s\eta_s$ for every $s \neq s'$, where $u_s = [u_{s,1}, \cdots, u_{s,T}]'$, and $E[u_{s,t}|x_s, \phi_s\eta_s] = c$ for every s, t .

Intuitively, time-invariant spatial dependence means that $\varepsilon_{s,t}$ can be decomposed in a time-invariant component η_s, and a time-variant component $u_{s,t}$, and that the spatial dependence in $\varepsilon_{s,t}$ is generated by the time-invariant component η_s. Note that time-invariance only restricts the spatial dependence on the unobserved variable $\varepsilon_{s,t}$ to be constant over time. It does not say that variables shouldn't be correlated at all. For example, we can have spatial and time dependence in $y_{s,t}$, $x_{s,t}$, $\varepsilon_{s,t}$, and spatial dependence in ϕ_s and still have time-invariant spatial dependence on $\varepsilon_{s,t}$.

If the spatial dependence on $\{\varepsilon_s\}_s$ is time-invariant, then the covariance of the score function between any two different individuals is zero. Therefore, the asymptotic results for the PCFE given in (1) along with its sandwich variance estimator are robust to spatial dependence under time-invariance. Theorem 1 states the result; we present the complete proof in the appendix.

Theorem 1. *Under correct specification of the conditional mean function (eq. 2), if the spatial dependence on $\{\varepsilon_s\}_s$ is 'time-invariant', then*

$$COV\left[S_s(\beta_0), S_{s'}(\beta_0)\right] = 0, \textit{ for every } s \neq s'$$

Consequently, $\sqrt{N}\left(\widehat{\beta}-\beta_0\right) \xrightarrow{d} N\left(0,\ A_0^{-1}B_0A_0^{-1}\right)$*, and the sandwich estimator given in equation 3 is consistent for the variance of the PCFE.*

Time-invariance on the spatial dependence of $\varepsilon_{s,t}$ does not affect the asymptotic variance of the PCFE because the source of spatial dependence is η_s. This time-invariant component of $\varepsilon_{s,t}$ is not separately identified from the fixed effects ϕ_s that we control for in the PCFE model. In other words, we can write an indistinguishable model with fixed effects $\tilde{\phi}_s$ that absorb the time-invariant components η_s leaving no spatial dependence on the unobserved shocks. For example, firm patent applications may be spatially correlated because of unobservable factors that affect the local supply of high-skilled workers such as cultural goods, high-quality restaurants or schools. If access to these amenities is constant during the sample period, then the structure of spatial dependence is time-invariant.

4 Testing for Time-invariant Spatial Dependence

In this section we construct a test-statistic and derive its distribution under the null hypothesis of time-invariant spatial dependence on the unobservable shocks $\{\varepsilon_s\}_s$. The idea is to detect if the asymptotic variance of $\hat{\beta}$ changes from $A_0^{-1}B_0A_0^{-1}$ to $A_0^{-1}C_0A_0^{-1}$ when there is spatial dependence. Throughout this section, we maintain Conley (1999)'s assumptions of spatial dependence (as in section 3). If the null hypothesis of time-invariance is false, then $A_0^{-1}B_0A_0^{-1}$ is generally different than $A_0^{-1}C_0A_0^{-1}$, and we should use the spatial variance formula (eq. 5) to estimate the variance of $\hat{\beta}$ consistently. If the null hypothesis is true, then $A_0^{-1}B_0A_0^{-1} = A_0^{-1}C_0A_0^{-1}$, and the simpler sandwich variance estimator is consistent for the variance of $\hat{\beta}$. The spatial variance estimator is also consistent if the null hypothesis is true, and the difference between sandwich and spatial estimates is small in large samples.

To compute the test statistic, we regress each of the K elements of the vector $S_s(\hat{\beta})$ on averages of the spatial lags of the elements of the vector $S_s(\hat{\beta})$. Under the null hypothesis of time-invariance, the covariance of the score function between different cross-sectional units is zero (Theorem 1). This leads to estimated regression coefficients that are close to zero in large samples. We use the Wald test statistic to test the restriction that all parameters are zero.

For a given observation located in $p = (m, n)$, we define the non-random set of neighbors up to the $l = (l_1, l_2)$ spatial lag to be $N(p, l) = \{p' = (m', n') \in \mathbb{Z}^2 : p' \neq p, |m' - m| \leq l_1, |n' - n| \leq l_2\}$. The number of elements in $N(p, l)$ does not depend on p, and we call it N_l. Remember that $Y_p(\beta) = S_s(\beta)$ if $s \in \mathbb{R}^2$ is in the grid cell $p \in \mathbb{Z}^2$, and zero otherwise. The average of the l-th spatial lags of the vector Y_p is defined as $\bar{Y}_{p,l} = N_l^{-1} \sum_{q \in N(p,l)} Y_q$. For a given choice of spatial lag $l = (l_1, l_2)$, and for each $k = 1, \ldots, K$, we regress $Y_p^{(k)}(\hat{\beta})$ on k explanatory variables $\bar{Y}_{p,l}^{(1)}(\hat{\beta}), \cdots, \bar{Y}_{p,l}^{(k)}(\hat{\beta})$ to obtain the $k \times 1$ vector of estimates $\hat{\theta}_k$. We have a total of $K^* = K(K+1)/2$ estimated parameters that are stacked in $\widehat{\Theta} = [\widehat{\theta}_1' \ \ldots \ \widehat{\theta}_K']'$. We define the test statistic $\widehat{T}$ as:

$$\hat{T} = N \ \hat{\Theta}' \ \hat{W}^{-1} \ \hat{\Theta}$$

where

$$\widehat{\Theta} = \left(\sum_p Z_p(\widehat{\beta}) Z_p(\widehat{\beta})' \right)^{-1} \sum_p Z_p(\widehat{\beta}) Y_p(\widehat{\beta})$$
$$Z_p(\beta)_{K^* \times K} = diag \left\{ \bar{Y}_{p,l}^{(1:k)}(\beta) \right\}_{k=1}^{K}$$

$$\bar{Y}_{p,l}^{(1:k)}(\beta)_{k \times 1} = \left[\bar{Y}_{p,l}^{(1)}(\beta), \cdots, \bar{Y}_{p,l}^{(k)}(\beta) \right]'$$
$$\widehat{W}_{K^* \times K^*} = \widehat{\Gamma}^{-1} \widehat{\Omega} \widehat{\Gamma}^{-1}$$
$$\widehat{\Gamma}_{K^* \times K^*} = -\frac{1}{N} \sum_p Z_p(\widehat{\beta}) Z_p(\widehat{\beta})'$$

$$\widehat{\Omega}_{K^*\times K^*} = \frac{1}{N} \sum_{j=-L_{D_1}+1}^{L_{D_1}-1} \sum_{k=-L_{D_2}+1}^{L_{D_2}-1}$$

$$\sum_{\substack{m=1 \\ \text{st } 1\leq m+j\leq D_1}}^{D_1} \sum_{\substack{n=1 \\ \text{st } 1\leq n+k\leq D_2}}^{D_2} K_{D_1,D_2}\left(j,k\right) \left[Z_{m,n}(\widehat{\beta}) Y_{m,n}(\widehat{\beta}) Y_{m+j,n+k}(\widehat{\beta})' Z_{m+j,n+k}\left(\widehat{\beta}\right)' \right]$$

Under the null hypothesis of time-invariant spatial dependence, this test-statistic is asymptotically Chi-square distributed (Theorem 2, for a proof see Appendix, 7.4).[6]

Theorem 2. *Under the null hypothesis of time-invariant spatial dependence,*

$$\hat{T} \xrightarrow{d} \chi^2_{K^*}$$

If the estimate of this test-statistic exceeds a critical value, the null hypothesis of time-invariant spatial dependence is rejected, which indicates that the sandwich variance estimator is inconsistent. Failing to reject the null hypothesis indicates both the sandwich and spatial variance estimators are consistent. In that case, there is either no spatial dependence or spatial dependence is time-invariant.

5 Applications

In this section, we illustrate the estimates and test-statistic with the examples of Hall, Griliches and Hausman (1986)'s analysis of patent counts, and Furman and Stern (2011)'s analysis of journal citations. In these examples, the differences between the sandwich and the spatial variance estimates are small; values of the test statistic fail to reject time-invariance spatial dependence which is consistent with small differences between spatial and sandwich variance estimates.

[6]It is unfeasible to compute $\widehat{\Theta}$ if $\widehat{\Gamma}$ or $\widehat{W}$ are not invertible. This may be the case when K^* is large compared to the sample size N. Choosing to regress a number smaller than K of elements of the score vector Y does not affect the limiting distribution of $\hat{T}$ except for the smaller number of degrees of freedom.

5.1 Application to Hall, Griliches and Hausman (1986)

Hall, Griliches and Hausman (1986) use patent counts to investigate whether patenting responds to contemporaneous investment in R&D or whether it responds with a lag. Their analysis of patent counts and R&D expenditures for 642 U.S. firms between 1972 and 1979 indicates that patenting is most responsive to contemporaneous expenditures in R&D. A plot of the spatial component of their data reveals evidence of spatial dependence for firms with more than 100 patent applications per year (Figure 1). In this section, we replicate the analysis of Hall, Griliches and Hausman (1986) to examine whether Poisson estimates are compromised by spatial dependence. Specifically, we replicate Table 6 (on p. 279 of their paper) as a PCFE with firm fixed effects:

$$E[p_{i,t}|R_{i,t}, \cdots, R_{i,t-3}, \phi_i] = \phi_i \exp\left(\sum_{\tau=0}^{3} \beta_\tau \log R_{i,t-\tau} + \alpha_t\right) \tag{6}$$

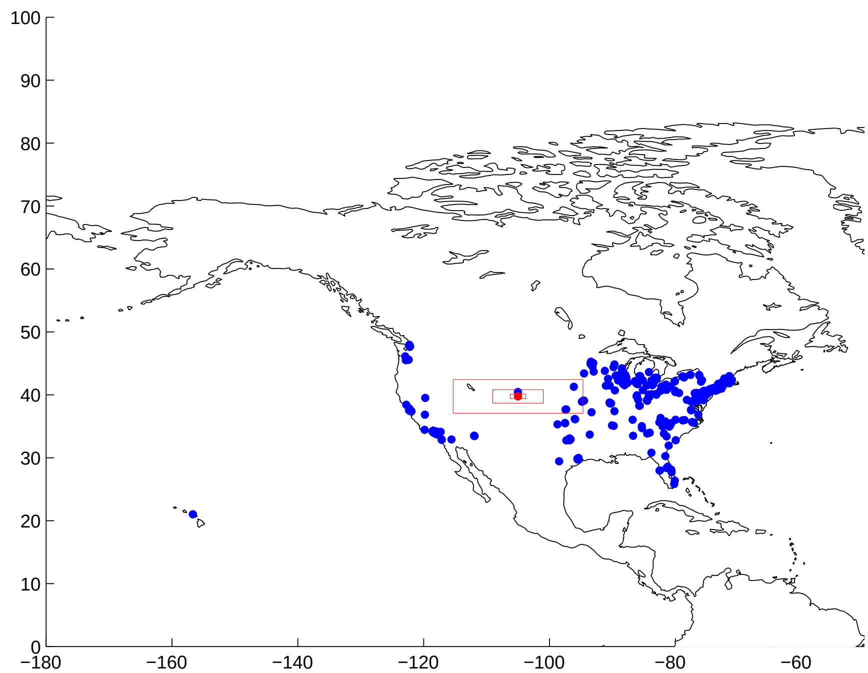
where

- $p_{i,t}$ is the number of patent applications by firm i in year t;
- $R_{i,t}$ measures R&D expenditures in millions of 1972 US dollars;
- α_t are year fixed effects;
- ϕ_i are firm fixed effects.

In these data, patent applications may be correlated across space, because the unobservable factors that influence patenting, such as the supply of high-skilled labor, specialized inputs, or knowledge spillovers, are correlated across space. Theorem 1 implies that the sandwich variance estimator is consistent in the presence of such spatial dependence, as long as the structure of spatial dependence is time-invariant. For example, the geographic location of high-skilled workers may vary with the supply of amenities that attract them, such as cultural goods, high-quality restaurants or schools. If access to these amenities is constant during the sample period, then the structure of spatial dependence is time-invariant.

To construct a measure for the geographic proximity between patent applications, we use information from the CUSIP (Committee on Uniform Security Identification Procedures) codes of firms who apply for patents to obtain information on their street addresses from COMPUSTAT.[7] Address data are available from COMPUSTAT for 460 of 642 firms. We then convert addresses to latitude and longitude coordinates, and calculate distances between coordinates.

Figure 2: Location of Patentees in Hall, Griliches and Hausman (1986)



Notes: Squares represent three bandwidth choices (the 10th, 25th, and 50th quantile of the distribution of the distance computed along the latitude and longitude coordinates). These bandwidths define nearness to the firms in the center of the three squares. For example, firms inside the largest square (for the 50th quantile of the distance distribution) are located in Colorado(center), Kansas, and Nebraska.

[7] Available at *www.compustat.com*, accessed on August, 2012. Figure 2 excludes one firm that is located in Singapore.

Table 1: Sandwich and spatial standard errors based on Hall, Griliches and Hausman (1986), Table 6-1

Log R&D investment in time t			
$\hat{\beta}_0$		0.2468	
sandwich s.e.		0.0786***	
spatial s.e.	0.0946***	0.0887***	0.0899***
bandwidth	10^{th}	25^{th}	50^{th}
Log R&D investment in time $t-1$			
$\hat{\beta}_1$		-0.0930	
sandwich s.e.		0.0797	
spatial s.e.	0.0792	0.0757	0.0774
bandwidth	10^{th}	25^{th}	50^{th}
Log R&D investment in time $t-2$			
$\hat{\beta}_2$		0.0687	
sandwich s.e.		0.0630	
spatial s.e.	0.0623	0.0601	0.0548
bandwidth	10^{th}	25^{th}	50^{th}
Log R&D investment in time $t-3$			
$\hat{\beta}_3$		-0.0224	
sandwich s.e.		0.0704	
spatial s.e.	0.0648	0.0622	0.0676
bandwidth	10^{th}	25^{th}	50^{th}
$\hat{T}$	17.85	10.7	7.34
Firms		413	
Years		5	
Observations		2065	

Notes: Poisson Conditional Fixed-Effect Maximum Likelihood estimates of equation 6 with sandwich standard errors (equation 3) and spatial standard errors (equation 5). The estimation excludes three years of observations to allow for three lags in $R_{i,t}$, and 47 firms without patents during the sample period. There are four year dummies for 1976-1979. The test-statistic $\hat{T}$ has an asymptotic Chi-square distribution with 36 degrees of freedom under the null hypothesis of time-invariant spatial dependence. We compute the spatial s.e. and the test-statistic for different bandwidth choices chosen to be the different quantiles of the distribution of distances. Significance of 1%, 5%, 10% is indicated with ***, **, * respectively.

Table 1 reports sandwich and spatial estimates for the variance of the PCFE. The statistically insignificant value of 10.7 for the test statistic (for the 25th quantile bandwidth) indicates that variance estimates are robust to allowing for spatial dependence. Consistent with time-invariant spatial dependence, the difference between the sandwich estimate and the spatial estimate of the standard errors is small. This result suggests that the sandwich estimator of the standard errors is consistent. As a kernel estimator, the estimated value of the spatial standard error depends on the researcher's choice of a bandwidth. Table 1 reports estimates for spatial standard errors using different choices of the bandwidth, defined by the 10th, 25th, and 50th quantile of nearness along coordinates of latitude and longitude; results are robust to these alternative choices of the bandwidth.

5.2 Application to Furman and Stern (2011)

Furman and Stern use citation counts to investigate whether the creation of a biological research center (BRC), which certified and catalogued information about biomaterials, helped to amplify the cumulative impact of scientific discoveries. In the United States, the "American Culture Collection" (ATCC) is the largest BRC; Furman and Stern exploit exogenous transfers of biomaterials to the ATCC to investigate whether materials in the ATCC became more heavily cited after they became certified and catalogued by the ATCC. Baseline specifications compare changes in citations to publications that use materials that were transferred to the ATCC with other materials that were not transferred to the ATCC (Furman and Stern (2011), p. 1949, specification 2):

$$\begin{aligned} &\mathit{FORWARD\ CITATIONS}_{i,t} \\ &= f\left(\varepsilon_{i,t}; \gamma_i + \beta_t + \delta_{t\text{-}pubyear} + \psi_{WINDOW}\ \mathit{BRC\text{-}ARTICLE} \times \mathit{WINDOW\ PERIOD}_{i,t}\right. \\ &\quad \left. + \psi\ \mathit{BRC\text{-}ARTICLE} \times \mathit{POST\text{-}DEPOSIT}_{i,t}\right) \end{aligned} \tag{7}$$

where :

- $\mathit{FORWARD\ CITATIONS}_{i,t}$ is the number of citations to article i in year t;
- $\mathit{BRC\text{-}ARTICLE} \times \mathit{WINDOW\ PERIOD}_{i,t}$ is a dummy variable that equals 1 if article

i is referenced by a BRC deposit and year t is equal to the year of the deposit or the year of the deposit plus or minus 1;[8]

- $BRC\text{-}ARTICLE \times POST\text{-}DEPOSIT_{i,t}$ is a dummy variable that equals 1 if article i is referenced by a BRC deposit and YEAR > DEPOSIT YEAR + 1 (i.e., deposit has already occurred and deposit WINDOW PERIOD already passed);
- γ_i denote article fixed effects;
- β_t denote year fixed effects;
- $\delta_{t\text{-}pubyear}$ are fixed effects for the age of articles.

Under the assumption that - without the transfers - changes in citations would have been identical for publications about materials that were transferred and other materials that were not transferred, the coefficient ψ measures the impact of the transfers on citations, as a measure of knowledge flows. Using a balanced panel of 99 articles and 23 periods between 1979 and 2001, with a total of 2,277 observations, we estimate equation 7 (Table 3, column 4 in Furman and Stern (2011), p. 1948).[9]

Citations to articles may, however, be affected by unobservable factors, such as scientific breakthroughs or the invention of new research tools, which increase research productivity across related fields. Intuitively, scientific breakthroughs are more likely to increase the speed of innovation in fields that are more closely related to the original field in which the breakthrough occurred than in other distant fields. Similarly, research tools, which have been developed for a specific field may encourage cumulative innovation in related fields but are less likely to benefit fields that are more distant in knowledge space. Results in section 3

[8] The authors allow for a one-year window after the deposit, because the exact length of the lag between the date of the deposit and the date when materials become accessible is unknown.

[9] Furman and Stern (2011) estimate specification (7) using the negative binomial model for an unbalanced panel of 216 articles and 32 periods with a total of 4,857 observations. The difference between spatial and sandwich standard errors is also small when we estimate the Poisson model using the unbalanced panel.

imply that the standard sandwich estimator is consistent - even in the presence of spatial dependence in idea space - as long as this dependence remains substantially unchanged during the sample period. For example, results in section 3 imply that the standard sandwich estimator is consistent - even if research output is correlated across related fields - as long as the way in which related fields are affected by common shocks (such as scientific breakthroughs or research tools) does not change during the sample period.

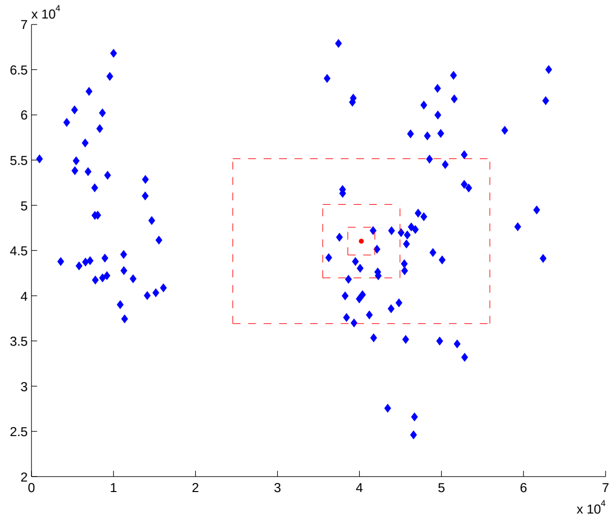
Table 2: Measures of Distance

Variable	**Definition**	**Measure of Distance**
pair_num	id number for treatment-control pairs	difference in id numbers
collectn	biological material's collection	zero if same collection
atcccode	ATCC code and collection of material	difference in codes
price	price of the referenced biological material	difference in price
journ_abb	journal where the article was published	zero if same value
author	article's author	zero if same value
authinst	author's institution	zero if same value
country	country of institution	zero if same value
pub_year	publication year of article	difference in years

Notes: For the qualitative variables *collectn*, *journ_abb*, *author*, *authinst*, *country*, each distance is zero if two different articles have the same value for the corresponding variable, or one if different values. For the numeric variables *pair_num*, *atcccode*, *price*, *pub_year*, each distance is equal to the absolute value of the difference of the corresponding variable for two different articles. This difference is normalized to be at most one. For a given pair of articles, if a variable has a missing value, this distance is dropped from the averaging. The variable *atcccode* is considered missing if two different articles are from different collections. 126 of 216 articles have missing values for at least one of the nine variables, and all of the articles have observed values for at least five variables.

To illustrate these results, we create a new measure of distance in idea space based on the analysis of journal citations in Furman and Stern (2011). Specifically, we create a proxy for distance in knowledge space based on the average distance for nine variables that Furman and Stern (2011) use to characterize articles (Table 2). This proxy allows us to generate coordinates, which we use to illustrate distance in knowledge space (Figure 3).

Figure 3: Dispersion of papers in idea space and bandwidth choices



Notes: Squares with dotted lines represent three bandwidth choices (the 10th, 25th, and 50th quantile of the distribution of the distance computed along the x and y coordinates). These bandwidths define nearness to the article in the center of the three squares. For example, two articles in the smallest square (for the 10th quantile of the distance distribution) share the same collection and similar publication year with the article that is at the center of that square.

Table 3 reports estimation results for the PCFE and its variance by the standard sandwich estimator (equation 3) and spatial estimator that accounts for spatial dependence (equation 5). Insignificant values for the test statistic in the range of 6.68 to 25.9 (for bandwidths between the 10th and 50th quantile) indicate that variance estimates are robust to allowing for spatial dependence. Consistent with time-invariant spatial dependence, the difference between the sandwich estimate and the spatial estimate of the standard errors is small. These results suggest that the sandwich estimator of the standard error is consistent.

Table 3: Sandwich and spatial standard errors based on Furman and Stern (2011), Table 3-4

BRC-ARTICLE $\times$ *WINDOW PERIOD*			
$\hat{\psi}_{WINDOW}$		0.4089	
sandwich s.e.		0.1657**	
spatial s.e.	0.1728**	0.1873**	0.2260*
bandwidth	10^{th}	25^{th}	50^{th}
BRC-ARTICLE $\times$ *POST-DEPOSIT*			
$\hat{\psi}$		0.5959	
sandwich s.e.		0.2802**	
spatial s.e.	0.3058*	0.3110*	0.3418*
bandwidth	10^{th}	25^{th}	50^{th}
$\hat{T}$	25.9	15.37	6.68
Articles		96	
Years		23	
Observations		2208	

Notes: Poisson Conditional Fixed-Effect Maximum Likelihood estimates of equation 7 with sandwich standard errors (equation 3) and spatial standard errors (equation 5). We exclude 3 articles that do not have any citations from the sample. There are year dummies for every year except 1979 and 1990, and 30 age dummies. The test-statistic $\hat{T}$ has an asymptotic Chi-square distribution with 45 degrees of freedom under the null hypothesis of time-invariant spatial dependence. We use the first nine elements of the score vector to calculate the test-statistic because it is not possible to compute the test-statistic using all elements of the score vector due to the large number of regressors in equation 7. We compute the spatial s.e. and the test-statistic for different bandwidth choices chosen to be the different quantiles of the distribution of distances. Significance of 1%, 5%, 10% is indicated with ***, **, * respectively.

6 Conclusion

Count data, such as patents, unemployed workers, or hospital visits, play an important role in empirical analyses across a broad range of research fields. Due to its robustness to misspecification of the Poisson distribution and to time dependence, the Poisson (PCFE) estimator has become a standard approach to analyze count data. Count data may, however, also be affected by cross-sectional dependence. For example, changes in patent counts for firms that are geographically close may be correlated if they are affected by geographically localized unobservable factors, such as knowledge spillovers or changes in the supply of inputs and skilled workers. Similarly, changes in research output may be correlated across related research fields, if they benefit from the same scientific advances.

This paper extends the robustness properties of the PCFE to the case of spatial dependence in the cross-section. We show that the asymptotic distribution of the PCFE – derived by Wooldridge (1999) under the assumption of cross-sectional independence – is robust to cross-sectional dependence, as long as cross-sectional dependence is time-invariant. The sandwich variance estimator is consistent if spatial dependence is time invariant. We construct a test statistic to detect time-variant spatial dependence, and provide a spatial variance estimator for the PCFE that is consistent under time-invariant and time-variant spatial dependence. We provide new commands for STATA and MATLAB that compute the sandwich and spatial variance estimates for the PCFE, as well as the test statistic. We illustrate the revised estimates through empirical examples based on Hall, Griliches and Hausman (1986) and Furman and Stern (2011). In these applications, the test statistic does not indicate time-variance spatial dependence suggesting that the sandwich variance estimator is consistent.

References

Andersen, Erling B (1972), 'The numerical solution of a set of conditional estimation equations', *Journal of the Royal Statistical Society. Series B (Methodological)* **34**(1), 42–54.

Azoulay, Pierre, Joshua Graff Zivin and Jialan Wang (2010), 'Superstar extinction', *The Quarterly Journal of Economics* **125**(2), 549–589.

Borjas, George J and Kirk B Doran (2012), 'The collapse of the soviet union and the productivity of american mathematicians', *The Quarterly Journal of Economics* **127**(3), 1143–1203.

Cameron, A. Colin and Pravin K. Trivedi (2005), *Microeconometrics: Methods and Applications*, Cambridge University Press.

Conley, Timothy G. (1999), 'Gmm estimation with cross sectional dependence', *Journal of Econometrics* **92**(1), 1–45.

Furman, Jeffrey L. and Scott Stern (2011), 'Climbing atop the shoulders of giants: The impact of institutions on cumulative research', *The American Economic Review* **101**(5), 1933–1963.

Gourieroux, Christian, Alain Monfort and Alain Trognon (1984), 'Pseudo maximum likelihood methods: Theory', *Econometrica* **52**(3), 681–700.

Hall, Bronwyn H., Zvi Griliches and Jerry Hausman (1986), 'Patents and r&d: Is there a lag?', *International Economic Review* **27**(2), 265–283.

Hausman, Jerry, Bronwyn H. Hall and Zvi Griliches (1981), Econometric models for count data with an application to the patents-r&d relationship. NBER Technical Working Paper No. 17.

Hausman, Jerry, Bronwyn H. Hall and Zvi Griliches (1984), 'Econometric models for count data with an application to the patents-r&d relationship', *Econometrica* **52**(4), 909–938.

Jaffe, Adam B., Manuel Trajtenberg and Rebecca Henderson (1993), 'Geographic localization of knowledge spillovers as evidenced by patent citations', *The Quarterly Journal of Economics* **108**(3), 577–598.

Kogan, Leonid, Dimitris Papanikolaou, Amit Seru and Noah Stoffman (2012), Technological innovation, resource allocation, and growth. NBER working paper 17769.

Marshall, Alfred (1890), *Principles of Economics*, MacMillan, London.

Moser, Petra and Alessandra Voena (2012), 'Compulsory licensing: Evidence from the trading with the enemy act', *The American Economic Review* **102**(1), 396–427.

Moser, Petra, Alessandra Voena and Fabian Waldinger (2014), German-Jewish Emigres and U.S. Invention. Forthcoming in the American Economic Review.

Williams, Heidi L. (2013), 'Intellectual property rights and innovation: Evidence from the human genome', *Journal of Political Economy* **121**(1), 1–27.

Wooldridge, Jeffrey M. (1999), 'Distribution-free estimation of some nonlinear panel data models', *Journal of Econometrics* **90**(1), 77–97.

7 Appendix

7.1 Code for Estimation

The command 'xtpsse' computes the PCFE, the sandwich and spatial variance estimator, and the test statistic for the null hypothesis of time-invariant spatial dependence. Each cross-sectional unit in the data must have coordinate information in $\mathbb{R}^2$ space. The files are available online at:

- *www.stanford.com/˜bertanha/xtpsse_stata.zip*;
- *www.stanford.com/˜bertanha/xtpsse_matlab.zip*.

Instructions on how to use the command are available in 'xtpsse_help.txt' inside each zip file.

7.2 Proof of Robustness to Time-invariant Spatial Dependence (Theorem 1)

We first prove the equivalence between time-invariance and the existence of a different model with fixed effects $\tilde{\phi}_s$ and no spatial dependence on its unobserved shocks.

Lemma 1. *The following statements are equivalent:*

(i) there exists $\{\tilde{\phi}_s\}_s$ strictly positive such that $y_s \perp (y_{s'}, x_{s'}, \tilde{\phi}_{s'})|x_s, \tilde{\phi}_s$ for every $s \neq s'$ and $E[y_{s,t}|x_s, \tilde{\phi}_s] = \tilde{\phi}_s \exp(x'_{s,t}\beta_0)$ for every s, t

(ii) the spatial dependence structure of ε_s is 'time-invariant'.

Proof. (i) $\Rightarrow$ (ii) : Make $u_{s,t} = y_{s,t}/\tilde{\phi}_s \exp(x'_{s,t}\beta_0)$, $\eta_s = \tilde{\phi}_s/\phi_s$, and $c = 1$.

Note that $u_{s,t} = \phi_s \varepsilon_{s,t}/\tilde{\phi}_s$ which leads to $\varepsilon_{s,t} = \eta_s u_{s,t}$. By assumption, $y_s \perp (y_{s'}, x_{s'}, \tilde{\phi}_{s'})|x_s, \tilde{\phi}_s$ makes $u_s \perp (u_{s'}, x_{s'}, \eta_{s'}\phi_{s'})|x_s, \phi_s\eta_s$ for $\forall s \neq s'$. Also, the assumption that $E[y_{s,t}|x_s, \tilde{\phi}_s] = \tilde{\phi}_s \exp(x'_{s,t}\beta_0)$ makes $E[u_{s,t}|x_s, \phi_s\eta_s] = c = 1$.

(i) $\Leftarrow$ (ii) : Make $\tilde{\phi}_s = \phi_s \eta_s c$. We have $y_{s,t} = \phi_s \exp(x'_{s,t}\beta_0)\varepsilon_{s,t} = \phi_s \eta_s c \exp(x'_{s,t}\beta_0) u_{s,t}/c$. The assumption that $E[u_{s,t}|x_s, \phi_s\eta_s] = c$ makes $E[y_{s,t}|x_s, \tilde{\phi}_s] = \tilde{\phi}_s \exp(x'_{s,t}\beta_0)$. Also, the assumption that $u_s \perp (u_{s'}, x_{s'}, \phi_{s'}\eta_{s'})|x_s, \phi_s\eta_s$ makes $y_s \perp (y_{s'}, x_{s'}, \tilde{\phi}_{s'})|x_s, \tilde{\phi}_s$ for $\forall s \neq s'$. $\square$

Now we show for every $s \neq s'$ that,

$$y_s \perp (y_{s'}, x_{s'}, \tilde{\phi}_{s'})|x_s, \tilde{\phi}_s \Rightarrow y_s \perp y_{s'}|x_s, \tilde{\phi}_s, x_{s'}, \tilde{\phi}_{s'} \tag{8}$$

Let the vector $W_s = (x_s, \widetilde{\phi}_s)$, and denote $\mathcal{W}$ and $\mathcal{Y}$ denote measurable events for W_s and y_s respectively.

$$\begin{aligned}
P\left(y_s \in \mathcal{Y}, y_{s'} \in \mathcal{Y}'|W_s \in \mathcal{W}, W_{s'} \in \mathcal{W}'\right) &= \frac{P\left(y_s \in \mathcal{Y}, y_{s'} \in \mathcal{Y}', W_{s'} \in \mathcal{W}'|W_s \in \mathcal{W}\right)}{P\left(W_{s'} \in \mathcal{W}'|W_s \in \mathcal{W}\right)} \\
&= \frac{P\left(y_s \in \mathcal{Y}|W_s \in \mathcal{W}\right) P\left(y_{s'} \in \mathcal{Y}', W_{s'} \in \mathcal{W}'|W_s \in \mathcal{W}\right)}{P\left(W_{s'} \in \mathcal{W}'|W_s \in \mathcal{W}\right)} \\
&= \frac{P\left(y_s \in \mathcal{Y}|W_s \in \mathcal{W}, W_{s'} \in \mathcal{W}'\right) P\left(y_{s'} \in \mathcal{Y}', W_{s'} \in \mathcal{W}'|W_s \in \mathcal{W}\right)}{P\left(W_{s'} \in \mathcal{W}'|W_s \in \mathcal{W}\right)} \\
&= P\left(y_s \in \mathcal{Y}|W_s \in \mathcal{W}, W_{s'} \in \mathcal{W}'\right) P\left(y_{s'} \in \mathcal{Y}'|W_s \in \mathcal{W}, W_{s'} \in \mathcal{W}',\right)
\end{aligned}$$

Define $z_{s,t,q} = x_{s,t} - x_{s,q}$. Following Hausman, Hall and Griliches (1981), the score function can be written as:

$$\begin{aligned}
S_s(\beta) &= \sum_{t=1}^{T} y_{s,t} \frac{\sum_{q=1}^{T} \exp(-z'_{s,t,q}\beta) z_{s,t,q}}{\sum_{u=1}^{T} \exp(-z'_{s,t,u}\beta)} \\
&= S(y_s, x_s, \beta) \qquad (9)
\end{aligned}$$

Now, we use Lemma 1, result (8) and equation (9) to prove that $COV\left[S_s(\beta_0), S_{s'}(\beta_0)\right] = 0$ for every $s \neq s'$. If the conditional mean is correctly specified (eq. 2 holds), then $E[S_s(\beta_0)] = EE[S_s(\beta_0)|x_s, \tilde{\phi}_s] = 0$, and we can write the covariance between the score function of two different individuals as:

$$COV\left[S_s(\beta_0), S_{s'}(\beta_0)\right] = E\left[S_s(\beta_0)S_{s'}(\beta_0)'\right]$$
$$= E\ E\left[S_s(\beta_0), S_{s'}(\beta_0)\,\middle|\, x_s, \tilde{\phi}_s, x_{s'}, \tilde{\phi}_{s'}\right]$$

$$= E\ E\left[S(y_s, x_s, \beta_0)S(y_{s'}, x_{s'}, \beta_0)'|\, x_s, \tilde{\phi}_s, x_{s'}, \tilde{\phi}_{s'}\right]$$
$$= E\ \left\{E\left[S(y_s, x_s, \beta_0)|\, x_s, \tilde{\phi}_s, x_{s'}, \tilde{\phi}_{s'}\right] E\left[S(y_{s'}, x_{s'}, \beta_0)'|\, x_s, \tilde{\phi}_s, x_{s'}, \tilde{\phi}_{s'}\right]\ \right\}$$

$$= E\ \left\{E\left[S(y_s, x_s, \beta_0)|\, x_s, \tilde{\phi}_s\right] E\left[S(y_{s'}, x_{s'}, \beta_0)|\, x_{s'}, \tilde{\phi}_{s'}\right]\ \right\}$$
$$= E\ \left\{E\left[S_s(\beta_0)|\, x_s, \tilde{\phi}_s\right] E\left[S_{s'}(\beta_0)|\, x_{s'}, \tilde{\phi}_{s'}\right]\ \right\} = 0$$

Therefore, $COV\left[S_s(\beta_0), S_{s'}(\beta_0)\right] = 0$ for every $s \neq s'$, which implies $COV\left[Y_p(\beta_0), Y_q(\beta_0)\right] = 0$ for every $p \neq q$, and that $C_0 = B_0$ under spatial dependence. Using this in (4) gives $\sqrt{N}\left(\widehat{\beta} - \beta_0\right) \xrightarrow{d} N\left(0,\ A_0^{-1}B_0A_0^{-1}\right)$, and the sandwich estimator given in equation 3 is consistent for the variance of the PCFE.

It is useful for the proof of Theorem 2 to show that $E[\nabla_{\beta'}S_s(\beta_0)S_{s'}(\beta_0)] = 0$ for every $s \neq s'$ holds under time-invariance. The expression for $\nabla_{\beta'}S_s(\beta)$ is:

$$\nabla_{\beta'}S_s(\beta) = \sum_{t=1}^{T} y_{s,t}$$
$$\left[\begin{array}{c} \frac{1}{\left(\sum_{q=1}^{T}\exp(-z'_{s,t,q}\beta)\right)^2}\left(\sum_{q=1}^{T}\exp(-z'_{s,t,q}\beta)z_{s,t,q}\right)\left(\sum_{q=1}^{T}\exp(-z'_{s,t,q}\beta)z_{s,t,q}\right)' \\ -\frac{1}{\sum_{q=1}^{T}\exp(-z'_{s,t,q}\beta)}\sum_{q=1}^{T}\exp(-z'_{s,t,q}\beta)z_{s,t,q}z'_{s,t,q} \end{array}\right]$$
$$= H(y_s, x_s, \beta)$$

Using the same results that led to $E[S_s(\beta_0)S_{s'}(\beta_0)'] = 0$, we have that:

$$
\begin{aligned}
E[\nabla_{\beta'} S_s(\beta_0) S_{s'}(\beta_0)] &= E[H(y_s, x_s, \beta_0) S(y_{s'}, x_{s'}, \beta_0)] \\
&= EE[H(y_s, x_s, \beta_0) S(y_{s'}, x_{s'}, \beta_0) | x_s, \widetilde{\phi}_s, x_{s'}, \widetilde{\phi}_{s'}] \\
&= E\left\{ E[H(y_s, x_s, \beta_0) | x_s, \widetilde{\phi}_s, x_{s'}, \widetilde{\phi}_{s'}] \; E[S(y_{s'}, x_{s'}, \beta_0) | x_s, \widetilde{\phi}_s, x_{s'}, \widetilde{\phi}_{s'}] \right\} \\
&= E\left\{ E[H(y_s, x_s, \beta_0) | x_s, \widetilde{\phi}_s] \; E[S_{s'}(\beta_0) | x_{s'}, \widetilde{\phi}_{s'}] \right\} = 0
\end{aligned}
$$

7.3 Proof of Identification: $E[S_s(\beta)] = 0 \iff \beta = \beta_0$

We refer to Conley (1999) for the complete set of assumptions behind his results of consistency and normality of $\hat{\beta}$, and consistency of $\hat{C}$. In this section, we demonstrate that using the PCFE score function as a moment condition satisfies the identification assumption in Conley (1999). It is straightforward to verify that this moment condition also satisfies his other assumptions under standard regularity conditions.

Lemma 2. *$E\left[S_s\left(\beta\right)\right] = 0 \Leftrightarrow \beta = \beta_0$, where $\beta_0 \in int\left(\mathbf{B}\right)$ with $\mathbf{B} \subseteq \mathbb{R}^K$ compact.*

Proof. Correct specification of the conditional mean, $E\left[y_{st}|x_s, \phi_s\right] = \phi_s \exp\left(x_{st}'\beta_0\right)$, plus uniqueness of a global maximum leads to this result. Hausman, Hall and Griliches (1981) show global concavity of this log-likelihood function. We use the same arguments to show global concavity of the expected value of the log-likelihood function and that β_0 is the unique zero of $E\left[S_s\left(\beta\right)\right] = 0$. The conditional expectation of the log-likelihood is:

$$
\begin{aligned}
E[\ell_s(\beta)|\phi_s, x_s] &= \phi_s \sum_{t=1}^{T} \exp(x_{st}'\beta_0) \log\left(\frac{\exp(x_{st}'\beta)}{\sum_{q=1}^{T}\exp(x_{sq}'\beta)}\right) \\
&= -\phi_s \sum_{t=1}^{T} \exp(x_{st}'\beta_0) \log\left(\sum_{q=1}^{T} \exp(-(x_{st} - x_{sq})'\beta)\right) \\
&= -\phi_s \sum_{t=1}^{T} \exp(x_{st}'\beta_0) \log\left(\sum_{s=1}^{T} \exp(-z_{stq}'\beta)\right)
\end{aligned}
$$

where $z_{stq} = x_{st} - x_{sq}$. Looking at the first and second derivatives:

$$\frac{\partial}{\partial\beta}E[\ell_s(\beta)|\phi_s, x_s] = \phi_s \sum_{t=1}^{T} \frac{\exp(x'_{st}\beta_0)}{\sum_{q=1}^{T}\exp(-z'_{stq}\beta)} \sum_{q=1}^{T}\exp(-z'_{stq}\beta)z_{stq}$$

$$\frac{\partial^2}{\partial\beta\partial\beta'}E[\ell_s(\beta)|\phi_s, x_s] = \phi_s \sum_{t=1}^{T}\exp(x'_{st}\beta_0)$$

$$\left[\begin{array}{c} \frac{1}{\left(\sum_{q=1}^{T}\exp(-z'_{stq}\beta)\right)^2}\left(\sum_{q=1}^{T}\exp(-z'_{stq}\beta)z_{stq}\right)\left(\sum_{q=1}^{T}\exp(-z'_{stq}\beta)z_{stq}\right)' \\ -\frac{1}{\sum_{q=1}^{T}\exp(-z'_{stq}\beta)}\sum_{q=1}^{T}\exp(-z'_{stq}\beta)z_{stq}z'_{stq} \end{array}\right]$$

In order to see that this second derivative is negative definite, for any given s, t, and β, call $v_q = z_{stq}$ and $a_q = \left[\exp(-z'_{stq}\beta)\right] / \left[\sum_{j=1}^{T}\exp(-z'_{stj}\beta)\right]$. Note that $a_q \in (0,1)$ and $\sum_{q=1}^{T} a_q = 1$. Rewriting the term in square brackets above gives:

$$\left(\sum_{q=1}^{T} a_q v_q\right)\left(\sum_{q=1}^{T} a_q v_q\right)' - \sum_{q=1}^{T} a_q v_q v'_q$$
$$= -\sum_{q=1}^{T} a_q\left[v_q - (\Sigma a_q v_q)\right]\left[v_q - (\Sigma a_q v_q)\right]'$$

which is simply the negative weighted sample variance of v_q. This is negative definite because x_s's are not constant across time which makes the v_q vary across q. Therefore, since $\phi_s exp(x'_{st}\beta_0) > 0$, we conclude that the second derivative is negative definite, i.e. the function is strictly concave. This is true for any (x_s, ϕ_s), so it is true for the unconditional expectation of the log-likelihood.

Given that the domain $\mathbf{B}$ is compact, there is a unique maximum and critical point. If the conditional mean is correctly specified, β_0 makes the expected value of the score function zero:

$$E\left[S_s\left(\beta\right)|x_s,\phi_s\right]=\nabla_\beta p\left(x_s,\beta\right)'W\left(x_s;\beta\right)E\left[u_s\left(\beta\right)|x_s,\phi_s\right]$$

where

$$\begin{aligned}E\left[u_{st}\left(\beta\right)|x_s,\phi_s\right]&=E\left[y_{st}|x_s,\phi_s\right]-E\left[p_t\left(x_s,\beta\right)n_s|x_s,\phi_s\right]\\&=\phi_s\exp\left(x_{st}'\beta_0\right)-\phi_s\frac{\exp\left(x_{st}'\beta_0\right)}{\sum_{k=1}^T\exp\left(x_{sk}'\beta\right)}\sum_{k=1}^T\exp\left(x_{sk}'\beta_0\right)\\&=\phi_s\exp\left(x_{st}'\beta_0\right)\left[1-\frac{\sum_{k=1}^T\exp\left(x_{sk}'\beta_0\right)}{\sum_{k=1}^T\exp\left(x_{sk}'\beta\right)}\right]\end{aligned}$$

If $\beta=\beta_0$, then $E\left[u_s\left(\beta_0\right)|x_s,\phi_s\right]=0$, which implies that $E\left[S_s\left(\beta_0\right)\right]=0$ using iterated expectations.

□

7.4 Proof of the Asymptotic Distribution of the Test Statistic (Theorem 2)

First, consider the unfeasible estimator $\widetilde{\Theta}=\left(\sum_p Z_p(\beta_0)Z_p(\beta_0)'\right)^{-1}\sum_p Z_p(\beta_0)Y_p(\beta_0)$, where we use the following notation: $p=(m,n)$, $\sum_p=\sum_{m,n}$, $K^*=K(K+1)/2$, and:

$$Z_p(\beta)_{K^*\times K}=diag\left\{\bar{Y}_{p,l}^{(1:k)}(\beta)\right\}_{k=1}^K$$

$$\bar{Y}_{p,l}^{(1:k)}(\beta)_{k\times 1}=\left[\bar{Y}_{p,l}^{(1)}(\beta),\cdots,\bar{Y}_{p,l}^{(k)}(\beta)\right]'$$

For a fixed choice of spatial lags $l=(l_1,l_2)$, the set $N(p,l)$ is non-random and remains fixed as the sample size increases. Using the GMM framework of Conley (1999), we can find the asymptotic distribution of $\widetilde{\Theta}$ under the null hypothesis of time-invariance:

$$\widetilde{W}^{-1/2}\sqrt{N}\ \widetilde{\Theta} = \widetilde{W}^{-1/2}\widetilde{\Gamma}^{-1}\frac{1}{\sqrt{N}}\sum_{p} Z_p(\beta_0)Y_p(\beta_0) \xrightarrow{d} N(0; I_{K^*}) \tag{10}$$

where

$$\widetilde{W}_{K^*\times K^*} = \widetilde{\Gamma}^{-1}\widetilde{\Omega}\widetilde{\Gamma}^{-1}$$

$$\widetilde{\Omega}_{K^*\times K^*} = \frac{1}{N}\sum_{j=-L_{D_1}+1}^{L_{D_1}-1}\ \sum_{k=-L_{D_2}+1}^{L_{D_2}-1}$$

$$\sum_{\substack{m=1\\ \text{st } 1\leq m+j\leq D_1}}^{D_1}\ \sum_{\substack{n=1\\ \text{st } 1\leq n+k\leq D_2}}^{D_2} K_{D_1,D_2}(j,k)\left[Z_{m,n}(\beta_0)Y_{m,n}(\beta_0)Y_{m+j,n+k}(\beta_0)'Z_{m+j,n+k}(\beta_0)'\right]$$

$$\widetilde{\Gamma}_{K^*\times K^*} = -\frac{1}{N}\sum_{p} Z_p(\beta_0)Z_p(\beta_0)'$$

Our goal is to use (10) to show asymptotic normality of the feasible estimator $\widehat{\Theta}$:

$$\widehat{W}^{-1/2}\sqrt{N}\ \widehat{\Theta} = \widehat{W}^{-1/2}\widehat{\Gamma}^{-1}\frac{1}{\sqrt{N}}\sum_{p} Z_p(\hat{\beta})Y_p(\hat{\beta}) \xrightarrow{d} N(0; I_{K^*})$$

where $\widehat{\Gamma}$ and $\widehat{W}$ have been defined in the main text. It suffices to show that $\left\|\widehat{\Gamma}-\widetilde{\Gamma}\right\| \xrightarrow{p} 0$, $\left\|\widehat{\Omega}-\widetilde{\Omega}\right\| \xrightarrow{p} 0$, and $N^{-1/2}\left\|\sum_p Z_p(\hat{\beta})Y_p(\hat{\beta}) - Z_p(\beta_0)Y_p(\beta_0)\right\| \xrightarrow{p} 0$.

Step 1: Under the null hypothesis of time-invariance, the proof of Theorem 1 shows that

$E\left[\nabla_\beta Y_p^{(i)}(\beta_0) Y_q^{(j)}(\beta_0)\right] = 0$ for every $p \neq q$ and $i, j \in \{1, \ldots, K\}$. Hence,

$$
\begin{aligned}
& E\left[\nabla_\beta \left\{\bar{Y}_{p,l}^{(i)} Y_p^{(j)}(\beta_0)\right\}\right] \\
&= E\left[N_l^{-1} \sum_{q \in N(p,l)} \nabla_\beta Y_{q,l}^{(i)}(\beta_0) Y_p^{(j)}(\beta_0)\right] + E\left[\nabla_\beta Y_p^{(j)}(\beta_0) N_l^{-1} \sum_{q \in N(p,l)} Y_{q,l}^{(i)}(\beta_0)\right] \\
&= N_l^{-1} \sum_{q \in N(p,l)} E\left[\nabla_\beta Y_{q,l}^{(i)}(\beta_0) Y_p^{(j)}(\beta_0)\right] + N_l^{-1} \sum_{q \in N(p,l)} E\left[\nabla_\beta Y_p^{(j)}(\beta_0) Y_{q,l}^{(i)}(\beta_0)\right] \\
&= 0
\end{aligned}
$$

and $N^{-1} \sum_p \nabla_\beta \left\{\bar{Y}_{p,l}^{(i)} Y_p^{(j)}(\beta_0)\right\} = o_p(1)$. We can show that the same is true when we replace β_0 with $\beta^* \xrightarrow{p} \beta_0$. Assuming bounded moments of the data, and using the facts that $Y_p(\beta)$ has smooth derivatives, and that β^* belongs to a compact set with probability approaching one, we arrive at:

$$
\left\| \frac{1}{N} \sum_p \nabla_\beta \left\{\bar{Y}_{p,l}^{(i)} Y_p^{(j)}(\beta^*)\right\} \right\| = o_p(1) \tag{11}
$$

$$
\left\| \frac{1}{N} \sum_p \nabla_\beta \left\{\bar{Y}_{p,l}^{(i)} \bar{Y}_{p,l}^{(j)}(\beta^*)\right\} \right\| = o_p(1) \tag{12}
$$

Similarly, we have that for any $i_0, i_1, j_0, j_1 \in \{1, \ldots, K\}$:

$$
\begin{gathered}
\frac{1}{N} \sum_{j=-L_{D_1}+1}^{L_{D_1}-1} \sum_{k=-L_{D_2}+1}^{L_{D_2}-1} \sum_{\substack{m=1 \\ \text{st } 1 \leq m+j \leq D_1}}^{D_1} \sum_{\substack{n=1 \\ \text{st } 1 \leq n+k \leq D_2}}^{D_2} \\
K_{D_1,D_2}(j,k) \left\| \nabla_\beta \left\{ \bar{Y}_{(m,n),l}^{(i_0)} Y_{m,n}^{(i_1)} Y_{m+j,n+k}^{(j_1)} \bar{Y}_{(m+j,n+k),l}^{(j_0)}(\beta^*) \right\} \right\| = O_p(1)
\end{gathered} \tag{13}
$$

Step 2: $\left\|\widehat{\Gamma} - \widetilde{\Gamma}\right\| \xrightarrow{p} 0$

Let (i, j) denote an arbitrary element of the $K^* \times K^*$ matrix Γ. Let i_0 and j_0 index the elements in $\bar{Y}_{p,l}$ that are used to compute $\Gamma_{i,j}$. Using (12) we can show that:

$$
\begin{aligned}
\left|\widehat{\Gamma}_{i,j} - \widetilde{\Gamma}_{i,j}\right| &= \left|\frac{1}{N}\sum_{p} \bar{Y}_{p,l}^{(i_0)}(\hat{\beta})\bar{Y}_{p,l}^{(j_0)}(\hat{\beta}) - \bar{Y}_{p,l}^{(i_0)}(\tilde{\beta})\bar{Y}_{p,l}^{(j_0)}(\beta_0)\right| \\
&\leq \|\hat{\beta} - \beta_0\| \left\|\frac{1}{N}\sum_{p} \nabla_{\beta}\left\{\bar{Y}_{p,l}^{(i_0)}\bar{Y}_{p,l}^{(j_0)}(\beta^*)\right\}\right\| \\
&= o_p(1)o_p(1) = o_p(1).
\end{aligned}
$$

Step 3: $\left\|\widehat{\Omega} - \widetilde{\Omega}\right\| \xrightarrow{p} 0$

Let (i,j) denote an arbitrary element of the $K^* \times K^*$ matrix Ω. Let i_0 and i_1 index the elements in $\bar{Y}_{p,l}$, and j_0 and j_1 index the elements in Y_p that are used to compute $\Omega_{i,j}$. Using (13) we can show that:

$$
\begin{aligned}
&\left|\widehat{\Omega}_{i,j} - \widetilde{\Omega}_{i,j}\right| \\
&\leq \|\hat{\beta} - \beta_0\| \frac{1}{N} \sum_{j=-L_{D_1}+1}^{L_{D_1}-1} \sum_{k=-L_{D_2}+1}^{L_{D_2}-1} \\
&\sum_{\substack{m=1 \\ \text{st } 1 \leq m+j \leq D_1}}^{D_1} \sum_{\substack{n=1 \\ \text{st } 1 \leq n+k \leq D_2}}^{D_2} K_{D_1,D_2}(j,k) \left\|\nabla_{\beta}\left\{\bar{Y}_{(m,n),l}^{(i_0)} Y_{m,n}^{(i_1)} Y_{m+j,n+k}^{(j_1)} \bar{Y}_{(m+j,n+k),l}^{(j_0)}(\beta^*)\right\}\right\| \\
&= o_p(1)O_p(1) = o_p(1).
\end{aligned}
$$

Step 4: $N^{-1/2}\left\|\sum_p Z_p(\hat{\beta})Y_p(\hat{\beta}) - Z_p(\beta_0)Y_p(\beta_0)\right\| \xrightarrow{p} 0$

Let i denote an arbitrary element of the $K^* \times 1$ vector Z_pY_p. Let i_0 index the element in $\bar{Y}_{p,l}$, and i_1 index the element in Y_p that are used to compute $\{Z_pY_p\}_i$. Using (11), we can show that:

$$
\begin{aligned}
&N^{-1/2}\left|\sum_p \left\{Z_p(\hat\beta)Y_p(\hat\beta) - Z_p(\beta_0)Y_p(\beta_0)\right\}_i\right| \\
&= N^{-1/2}\left|\sum_p \bar{Y}_{p,l}^{(i_0)}(\hat\beta)Y_p^{(i_1)}(\hat\beta) - \bar{Y}_{p,l}^{(i_0)}(\beta_0)Y_p^{(i_1)}(\beta_0)\right| \\
&\leq \sqrt{N}\|\hat\beta - \beta_0\| \left\|\frac{1}{N}\sum_p \nabla_\beta \left\{\bar{Y}_{p,l}^{(i_0)}Y_p^{(i_1)}(\beta^*)\right\}\right\| \\
&= O_p(1)o_p(1) = o_p(1)
\end{aligned}
$$

Therefore, $\widehat{W}^{-1/2}\sqrt{N}\ \widehat\Theta \xrightarrow{d} N(0; I_{K^*})$, and $\widehat{T} = N\ \widehat\Theta'\widehat{W}^{-1}\widehat\Theta \xrightarrow{d} \chi^2_{K^*}$.