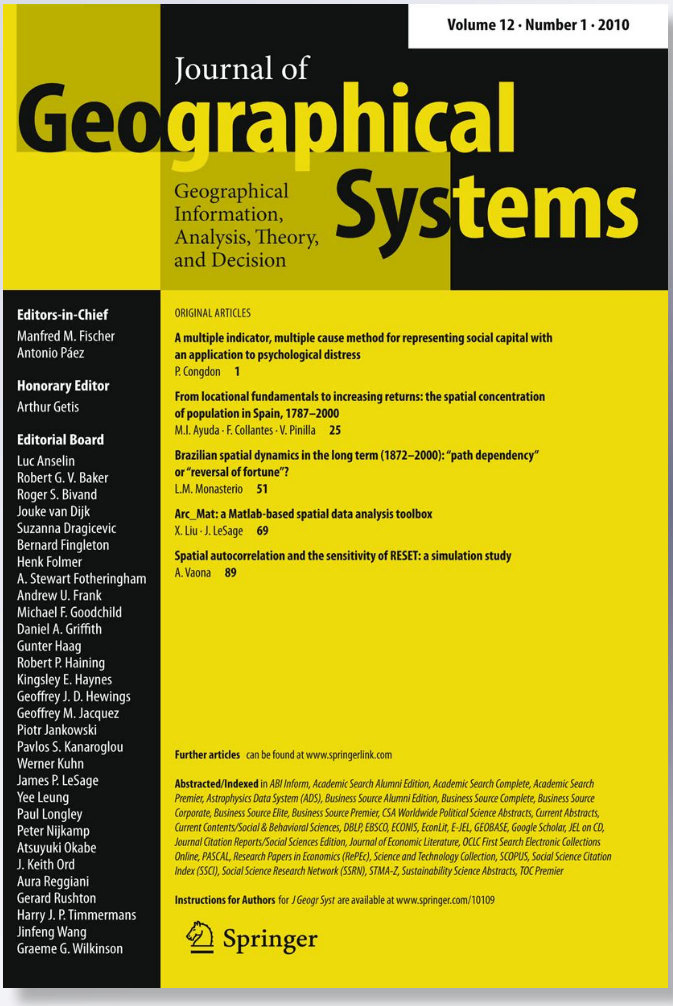
The effects of spatial autoregressive dependencies on inference in ordinary least squares: a geometric approach

Tony E. Smith & Ka Lok Lee

Journal of Geographical Systems
Geographical Information, Analysis,
Theory, and Decision

ISSN 1435-5930

J Geogr Syst
DOI 10.1007/s10109-011-0152-x



Springer

J Geogr Syst
DOI 10.1007/s10109-011-0152-x

ORIGINAL ARTICLE

The effects of spatial autoregressive dependencies on inference in ordinary least squares: a geometric approach

Tony E. Smith · Ka Lok Lee

Received: 10 February 2011 / Accepted: 15 July 2011

Abstract There is a common belief that the presence of residual spatial autocorrelation in ordinary least squares (OLS) regression leads to inflated significance levels in beta coefficients and, in particular, inflated levels relative to the more efficient spatial error model (SEM). However, our simulations show that this is not always the case. Hence, the purpose of this paper is to examine this question from a geometric viewpoint. The key idea is to characterize the OLS test statistic in terms of angle cosines and examine the geometric implications of this characterization. Our first result is to show that if the explanatory variables in the regression exhibit no spatial autocorrelation, then the distribution of test statistics for individual beta coefficients in OLS is *independent* of any spatial autocorrelation in the error term. Hence, inferences about betas exhibit all the optimality properties of the classic uncorrelated error case. However, a second more important series of results show that if spatial autocorrelation is present in both the dependent and explanatory variables, then the conventional wisdom is correct. In particular, even when an explanatory variable is statistically independent of the dependent variable, such joint spatial dependencies tend to produce "spurious correlation" that results in over-rejection of the null hypothesis. The underlying geometric nature of this problem is clarified by illustrative examples. The paper concludes with a brief discussion of some possible remedies for this problem.

Electronic supplementary material The online version of this article (doi:10.1007/s10109-011-0152-x) contains supplementary material, which is available to authorized users.

T. E. Smith (✉)
Department of Electrical and Systems Engineering,
University of Pennsylvania, Philadelphia, PA, USA
e-mail: tesmith@seas.upenn.edu

K. L. Lee
The Wharton School, University of Pennsylvania, Philadelphia, PA, USA
e-mail: kaloklee@wharton.upenn.edu

Published online: 09 August 2011

Keywords Spatial dependence · Spatial autocorrelation · Spatial error model · OLS regression · Geometric approach

JEL Classification C12 · C31

1 Introduction

There is a common belief that the presence of residual spatial autocorrelation in ordinary least squares (OLS) regression leads to inflated significance levels in beta coefficients and, in particular, inflated levels relative to the more efficient spatial error model (SEM). However, it is well known that OLS can continue to be very efficient in cases where spatial dependencies are very weak.[1] What is less well known is that the same can happen even when the spatial residual autocorrelation is very strong. In particular, this is true when spatial autocorrelation in the *explanatory* variables is very weak or nonexistent. Hence, one objective of this paper is to clarify this type of OLS efficiency from a geometric viewpoint.

A second more important objective is to show that if similar types of spatial autocorrelation are present in both the dependent and explanatory variables, then the conventional wisdom is correct. In particular, even when an explanatory variable is statistically independent of the dependent variable, such joint spatial dependencies tend to produce "spurious correlation" that results in over-rejection of the null hypothesis.

This problem of "spurious regression" between independent autocorrelated variates, y and x, is by no means new and appears to have been first studied by Bivand (1980) in the context of estimating correlations between such variables. Subsequently, Fingleton (1999) studied this problem more explicitly in terms of regression and, in particular, linked such spurious regression problems in space to the long-standing literature on spurious regressions between independent time series. Both of these papers focus on simulation analyses (as we do in Sect. 2 below; see also Legendre et al. 2002). However, a separate line of formal investigation, starting with the work of Clifford et al. (1989), has attempted to develop improved test statistics for dealing with this problem (including the work of Dutilleul 1993, 2008, Alpargu and Dutilleul 2003a, b, 2006, Mur and Trivez 2003 and Lauridsen and Kosfeld 2006). We shall return to this issue in the concluding section, but for the present, it should be emphasized that even these analytical results are often indirect and leave open the question of *why* these spurious correlation problems occur.

Hence, the major purpose of this paper is to propose a geometric approach that can serve to shed further light on this issue. This geometric approach is made possible by characterizing the standard OLS test statistic in terms of angle cosines. Although this characterization is well known (as for example in Davidson and

[1] In fact, the same is true for the much broader class of feasible generalized least square (FGLS) estimators. For as out by Green (2003, p. 211) and others, OLS is usually more efficient than FGLS when departures from classical assumptions of linear models are not too severe. For specific simulation results in the context of temporal autocorrelation, see for example Dutilleul and Alpargu (2001).

MacKinnon 1993), to the best of our knowledge, it has not been fully exploited in the present context. In particular, this approach allows the general case of multiple regression to be reduced to an equivalent simple regression for analyzing test statistics on individual beta coefficients. This simplification yields geometric insights that do not appear to be accessible by other means.

Using this approach, our analysis will show that the presence or absence of spurious correlation between independent spatial variables, $y = (y_1, \ldots, y_n)$ and $x = (x_1, \ldots, x_n)$, is determined by the *sphericity* properties of their marginal distributions on the underlying n-dimensional sample space. In particular, it is the degree of nonsphericity created by spatial autocorrelation among their individual components that leads to spurious correlation.

To develop these results, we begin in the next section with a series of simulated examples that illustrate the behavior of OLS tests as described above. Our geometric approach is then developed in Sect. 3 and is applied to show why OLS efficiency persists when explanatory variables are free of spatial autocorrelation. This is summarized by invariance theorems (Theorems 1 and 2) that make this property explicit. The more important case of spatially autocorrelated explanatory variables is then addressed in Sect. 4. Here, our strategy is to begin with a nonspatial setting in which the role of individual component correlations can be made more transparent. These general relationships are then applied to the spatial case. In particular, a limit theorem (Theorem 3) is developed which characterizes the exact type of linear dependencies arising among components of random vectors as spatial autocorrelation approaches its upper limit. This yields a natural range of dependencies from zero-correlated to perfectly correlated components that exactly parallel the nonspatial case. The paper concludes in Sect. 5 with a brief discussion of methods for resolving these spurious correlation problems.

2 Examples to illustrate the main ideas

To illustrate the general statistical phenomena described above, it is convenient to begin with the "classroom" example shown in Fig. 1 below. This one-dimensional example is sufficiently simple to provide a visual motivation for spatial regression models. Here, one can imagine that the x-axis represents distances (locations) along a city street emanating from the city center (CBD), with y denoting population density at each location. The true trend, $E(y|x)$, shown by the solid line in Fig. 1a, can be then viewed as expected density at distance x from the CBD throughout the city.[2] The actual densities shown at sampled points along this street exhibit some degree of deviation from the mean but are highly correlated with their neighbors and vary much more smoothly than random deviations.[3] Hence, an OLS regression of y on x using such points will tend to yield a line of fit with a smaller sum-of-squared deviation than the true trend, as shown by the dashed line in Fig. 1b. More

[2] As with most linear models, this trend line is at best only locally linear. But it may still provide a reasonable description of mean population density within the range shown.

[3] Of course not all correlated deviation patterns will be as smooth as those depicted. This stylized representation is only meant to illustrate a general tendency toward smoothness.

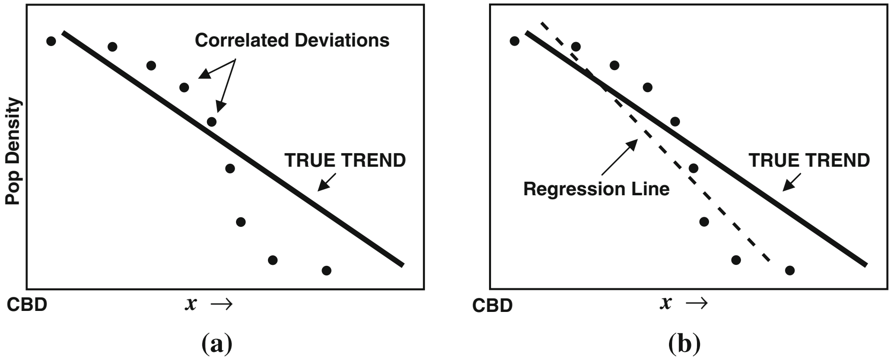


Fig. 1 Stylized example of spatial autocorrelation. **a** Autocorrelated urban density levels. **b** Effects on linear regression

generally, there is a clear tendency of least-squares estimation to *underestimate* true variance in the presence of such spatially correlated errors. But since smaller variance estimates tend to inflate the associated t statistics for betas, it follows that significance levels for these betas will also be inflated. This simple example thus motivates the need for regression models such as SEM that take explicit account of such spatial correlation effects.

2.1 OLS versus SEM for spatially independent explanatory variables

The simplicity of the example above turns out to be somewhat deceptive. In particular, the inference properties of OLS depend critically on the spatial dependency properties of *explanatory variables* as well as those of the dependent variable. This can be illustrated by the following more typical example. Here, we consider a two-dimensional simulated regression based on the 49 neighborhoods of Columbus, Ohio.[4] For most of the discussion and analysis to follow, we focus on *simple* regressions using the standard linear model:

$$y = \beta_0 1_n + \beta_1 x + \upsilon, \quad \upsilon \underset{iid}{\sim} N(0, \sigma^2) \tag{1}$$

where 1_n is the unit n-vector and (y, x, υ) are random n-vectors (with $n = 49$ in the present case). Our main interest focuses on significance tests for explanatory variables, which amounts in the present case to tests of the null hypothesis that $\beta_1 = 0$. In this context, we shall evaluate the efficacy of specific testing procedures in terms of the *actual size* of such tests (i.e., the fraction of null hypothesis rejections) versus their *nominal size*, which we here set to be 0.05.

To introduce spatial autocorrelation effects into Eq. 1, we simulate υ as a standard spatial autoregressive process, with *spatial weight matrix*, $W = (w_{ij}$: i,

[4] This set of spatial boundaries (as illustrated in Fig. 2 of Smith and Lee 2011a) is taken from Anselin (1988) and also constitutes one of the standard examples used in Geoda software.

$j = 1, \ldots, n$), and *spatial dependency parameter*, ρ_y. Thus, our simulated values of y are drawn from a model of the form:

$$y = \beta_0 1_n + \upsilon, \quad \upsilon = \rho_y W \upsilon + \varepsilon, \quad \varepsilon \sim N(0, \sigma_y^2 I_n) \tag{2}$$

where I_n is the n-square identity matrix, and the y-subscript in both ρ_y and σ_y^2 reflects the fact that y itself is spatially autocorrelated with spatial dependence parameter, ρ_y, and (intrinsic) variance, σ_y^2. Here, W is specified to be a "queen-contiguity" matrix,[5] so that direct spatial dependencies are essentially limited to immediate neighbors. For convenience, W is also scaled to have a maximum eigenvalue of one, so that the range of ρ_y is restricted to the unit interval.[6] Moreover, since we are primarily interested in positive spatial dependencies, the relevant range of ρ_y for our simulations is taken to be the interval, [0,1).

To complete the simulation model, we must also draw samples of the explanatory variable, x, in Eq. 1 which is by hypothesis taken to be independent of y. Moreover, since the marginal distribution of x is typically of less interest in this regression framework, we start with the simple assumption that the components of x are *iid* normal:

$$x \sim N(\mu_x 1_n, \sigma_x^2 I_n) \tag{3}$$

With these specifications, 100,000 simulations[7] of model (Eqs. 2, 3) were generated using values $\beta_0 = \mu_x = \sigma_x = \sigma_y = 1$ together with a selected set of ρ_y values in [0,1).[8] Each of the resulting y-vectors yielded data for *OLS* estimation of model Eq. 1. For comparison, the following *spatial error model* (SEM)

$$y = \beta_0 1_n + \beta_1 x + \upsilon, \quad \upsilon = \rho_y W \upsilon + \varepsilon, \quad \varepsilon \sim N(0, \sigma_y^2 I_n) \tag{4}$$

was also estimated for unknown parameters $(\beta_0, \beta_1, \rho_y, \sigma_y^2)$ using standard maximum likelihood (ML) procedures. To compare the inference properties of these two estimation procedures, we start with the *OLS test* for β_1, which in this case is simply the standard t test based on a t distribution with $n - 2$ degrees of freedom. However, to construct a comparable test of β_1 for SEM, we note first that in samples as small as the present one, the usual asymptotic z test for ML estimators (based on the standard normal distribution) is well known to be biased and in particular suffers from precisely the type of "over-rejection" problems that we wish to study. Hence, to minimize this particular source of over-rejections, our *SEM test* of β_1 is also based on the t distribution with $n - 2$ degrees of freedom, rather than the z test. This not only helps to minimize the small-sample bias of

[5] Queen contiguity implies that equal positive weights, $w_{ij} > 0$, are assigned to all distinct neighborhood pairs, ij, and that $w_{ij} = 0$ elsewhere.

[6] The restriction, $|\rho_y| < 1$, ensures that the matrix inverse, $(I_n - \rho_y W)^{-1}$, exists over the full range of ρ_y, so that the autoregressive process has a well-defined reduced form, $\upsilon = (I_n - \rho_y W)^{-1}\varepsilon$.

[7] This unusually large number of simulations was employed to minimize any possible sampling error in the results of Table 1a below.

[8] The specific values of ρ_y used were in increments of 0.1 from 0 to 0.9, together with the end value, 0.95. Given the singularity of $(I_n - \rho_y W)^{-1}$ at $\rho_y = 1.0$, values larger than 0.95 tend to exhibit computational instabilities.

SEM tests,[9] it also renders the test results more directly comparable to those of the *OLS test* for β_1.

But in spite of this modification favoring SEM tests, the simulation results in part (a) of Table 1 show that in terms of test size, OLS *is almost uniformly superior to* SEM for all values of ρ_y.[10] For example, when $\rho_y = 0.3$, the size of the test for β_1 under OLS is almost exactly 0.05, while that under SEM is slightly larger (0.057). This is even more surprising given that the SEM test is correctly specified in terms of the actual autocorrelation process simulated in Eq. 2,[11] while the OLS test is based on the misspecified error model in Eq. 1. However, one should hasten to add that SEM is not performing badly here (with sizes generally between 0.05 and 0.06) and is only noticeably worse that OLS under conditions of weak autocorrelation where OLS is expected to work well (see footnote 1). So the most striking feature of these results for our purposes is the uniformly strong performance of OLS, even at extreme levels of spatial autocorrelation in the dependent variable, y. This would appear to contradict all intuition gained from the one-dimensional example above and indeed forms one major focus of the present paper.

The key to this apparent contradiction can be found in our seemingly innocent assumptions about the distribution of the explanatory variable, x. In particular, we have assumed that the components of x are independently distributed and hence exhibit *no spatial autocorrelation*. Under these conditions, it is shown in Sect. 3 below that OLS exhibits all the optimality properties of the classical linear model with respect to tests about β_1. In particular, spatial autocorrelation in y has no effect whatsoever. Hence, problems of inflated significance in OLS tests of betas only arise when the associated explanatory variables are also spatially autocorrelated. In particular, this is precisely the reason for the inflated significance observed in the one-dimensional example above. Indeed, the explanatory variable in this case, namely "distance to CBD", is spatially autocorrelated in the most obvious way: locations close together in space must necessarily exhibit similar distances to the CBD.

2.2 OLS versus SEM for spatially autocorrelated explanatory variables

While the strong performance of OLS under conditions of spatially independent explanatory variables is very striking, it should be emphasized that in terms of practical applications, this result is generally not very helpful. Indeed, since most spatial data exhibit some degree of spatial dependence, it is natural to expect that explanatory variables, x, are as likely to be spatially dependent as is the y variable of interest. So perhaps the main practical consequence of this result is to suggest that one can expect OLS to perform reasonably well whenever spatial autocorrelation in explanatory variables is very weak. But when this is not the case, spatial regression models such as SEM do indeed perform better than OLS.

[9] Of course all variance estimates are still based on the standard asymptotic covariance matrix for ML estimation, so that some small-sample bias remains.

[10] There are six separate illustrations in Table 1, labeled (a) through (f). We shall refer to each by its label, such as Table 1a for the present case. Graphical representations of each illustration are given in the longer version of this paper, Smith and Lee (2011a), available online.

[11] In particular, Eq. 2 is precisely Eq. 4 under the null hypothesis, $\beta_1 = 0$.

Table 1 Test sizes for simulated examples with nominal size = 0.05

	(a) Columbus ($n = 49$) $\rho_x = 0$		(b) Columbus ($n = 49$) $\rho_x = 0.5$		(c) Columbus ($n = 49$) $\rho_x = 0.8$	
	OLS	SEM	OLS	SEM	OLS	SEM
ρ_y						
0	0.049	0.060	0.051	0.071	0.049	0.080
0.1	0.049	0.059	0.055	0.069	0.067	0.079
0.2	0.049	0.058	0.059	0.068	0.066	0.078
0.3	0.050	0.057	0.064	0.067	0.069	0.076
0.4	0.050	0.056	0.071	0.066	0.094	0.074
0.5	0.049	0.055	0.079	0.064	0.113	0.073
0.6	0.050	0.054	0.083	0.063	0.142	0.070
0.7	0.049	0.053	0.096	0.061	0.169	0.063
0.8	0.050	0.051	0.110	0.059	0.218	0.060
0.9	0.049	0.050	0.136	0.056	0.291	0.059
0.95	0.049	0.049	0.154	0.053	0.343	0.052

	(d) Columbus ($n = 49$) $\rho_y = 0.5$		(e) Columbus ($n = 49$) $\rho_y = 0.8$		(f) Philadelphia ($n = 367$) $\rho_y = 0.5$	
	OLS	SEM	OLS	SEM	OLS	SEM
ρ_x						
0	0.051	0.058	0.051	0.051	0.051	0.050
0.1	0.053	0.059	0.056	0.051	0.058	0.051
0.2	0.055	0.061	0.064	0.053	0.062	0.052
0.3	0.065	0.062	0.072	0.054	0.069	0.053
0.4	0.072	0.063	0.093	0.055	0.081	0.054
0.5	0.079	0.065	0.124	0.056	0.088	0.055
0.6	0.084	0.066	0.142	0.058	0.094	0.056
0.7	0.092	0.068	0.181	0.061	0.101	0.057
0.8	0.123	0.070	0.245	0.067	0.122	0.058
0.9	0.134	0.078	0.332	0.079	0.163	0.059
0.95	0.153	0.092	0.363	0.121	0.194	0.061

To illustrate these properties more systematically, we begin by extending the simulation framework above to allow for spatial autocorrelation in x as well as in y. In particular, we now assume that both y and x are governed by independent autocorrelation processes as in Eq. 2 above.[12] More precisely, it is now assumed that

$$y = \mu_y 1_n + \upsilon_y, \quad \upsilon_y = \rho_y W \upsilon_y + \varepsilon_y \tag{5}$$

$$x = \mu_x 1_n + \upsilon_x, \quad \upsilon_x = \rho_x W \upsilon_x + \varepsilon_x \tag{6}$$

[12] Note that μ_y in Eq. 5 below now plays the role of β_0 in Eq. 2.

$$\begin{pmatrix} \varepsilon_y \\ \varepsilon_x \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_y^2 I_n & \\ & \sigma_x^2 I_n \end{pmatrix} \right]. \tag{7}$$

Here, independence of y and x is implicit from the joint normal distribution of $(\varepsilon_y, \varepsilon_x)$. Since both the y and x distributions are simple instances of spatial error models, we shall refer to (Eqs. 5–7) as a *joint spatial error* (JSE) model.

For our later purposes, it is convenient to rewrite the JSE model in *reduced form* by eliminating υ_y and υ_x to obtain:

$$y = \mu_y 1_n + (I_n - \rho_y W)^{-1} \varepsilon_y = \mu_y 1_n + B_{\rho_y}^{-1} \varepsilon_y \tag{8}$$

$$x = \mu_x 1_n + (I_n - \rho_x W)^{-1} \varepsilon_x = \mu_x 1_n + B_{\rho_x}^{-1} \varepsilon_x \tag{9}$$

where $B_{\rho_y} = I_n - \rho_y W$ and $B_{\rho_x} = I_n - \rho_x W$. In these terms, an equivalent formulation of our JSE model is given by (Eqs. 7–9).

This model is essentially the same as that employed by Bivand (1980) and Fingleton (1999) except that here, we allow both y and x to have a nonzero mean. Note in particular that, like these authors, we assume a *common* weight matrix, W, for each process. Hence, a fundamental assumption of this JSE model is that spatial autocorrelation in both y and x is of a similar type and differs only in degree (as reflected by the spatial dependency parameters, ρ_y and ρ_x). While the presence of spurious correlation does not depend on this assumption, it is most easily illustrated in this setting.[13] Our only additional assumption for the present is that W is normalized to have unit eigenvalue,[14] so that the relevant range of both spatial dependence parameters, ρ_y and ρ_x, for all nonnegative spatial autocorrelation effects is again taken to be the interval [0,1).[15]

With this more general setup, we now extend the Columbus simulation example above to consider positive values of spatial dependency values, ρ_x, for the x process (where the same queen-contiguity matrix, W, is applied to x as well as to y). To illustrate these test-size results, we now assume a fixed spatial dependence value, $\rho_x = 0.5$, for the x process, which is taken to represent a substantial degree of spatial autocorrelation in x. Table 1b shows the results of 10,000 simulations of this JSE model for parameter values, $\mu_y = \mu_x = \sigma_x = \sigma_y = 1$, and a selected range of ρ_y values. Notice first that OLS again continues to outperform SEM for low values of ρ_y, where spatial dependency in the y process is relatively weak. However, as ρ_y increases, the situation changes dramatically. For while SEM behaves only slightly worse than in the independence case with $\rho_x = 0$, the over-rejection problem for OLS now becomes quite severe. So for substantial autocorrelation in the x process, we see that "conventional wisdom" about the failure of OLS inferences is restored.

[13] The effects of different weight matrices for y and x are illustrated in the longer version of this paper (Smith and Lee 2011a, Section 4.4.3).

[14] An additional restriction on W will be considered in Sect. 4.3.1 below.

[15] We also note that the inverses $B_{\rho_y}^{-1}$ and $B_{\rho_x}^{-1}$ are guaranteed to exist when $\rho_y, \rho_x \in [0, 1)$. For an analysis of this model in the case of "unit roots" where either $\rho_y = 1$ or $\rho_x = 1$, see for example Lauridsen and Kosfeld (2006).

Our ultimate goal is to explain why this should be so. But for the present, we explore the properties of this simulated example.

First one may ask how test sizes vary with ρ_x for a fixed value of ρ_y. Here, we again choose $\rho_y = 0.5$ to represent substantial spatial dependence in the y process, and for the same selected values of ρ_x and parameter values above, show the results for 10,000 simulations of the JSE model in Table 1d.[16] Here, we see that OLS exhibits essentially the same over-rejection behavior as in Table 1b above. Our later results will show that is to be expected. First, it is already clear that y and x are treated symmetrically in the JSE model. Moreover, it will be shown that in the case of simple regression, there is also symmetry in the way that spatial autocorrelation in y and x influences the relevant test statistics. Hence, what is most striking about Table 1d versus Table 1a and b is that now SEM is actually doing *worse* as spatial dependency in x (rather than y) increases.

We shall attempt to explain (or at least clarify) these patterns of test-size behavior in subsequent sections. For the present, we simply illustrate a number of different aspects of this behavior informally. First, as a parallel to Table 1b and d above, the same results are given for $\rho_x = 0.8$ and $\rho_y = 0.8$, respectively, in Table 1c and e. In qualitative terms, these results are very similar to those for $\rho_x = 0.5$ and $\rho_y = 0.5$. The single most dramatic difference is with respect to OLS, where at these higher levels of spatial autocorrelation, the over-rejection rates have virtually doubled at all scales.

Perhaps a more important question relates to the effect of sample size. As is well known, ML estimation procedures (for correctly specified models) are asymptotically efficient. So for sufficiently large samples, ML estimation of JSE models should yield reliable test sizes even for extreme levels of spatial autocorrelation in y and/or x. This is indeed the case, as is illustrated by the following example. Here we extend the simulation framework for the $n = 49$ neighborhoods in Columbus, Ohio, to the larger set of $n = 367$ census tracts in Philadelphia, Pennsylvania. Here again we use a normalized queen-contiguity weight matrix for Philadelphia, together with the same parameter settings for the JSE model above. The test-size results for 10,000 simulations for this larger example are shown in Table 1f, using the same range of ρ_x values and fixed value $\rho_y = 0.5$. A comparison with Table 1b for Columbus shows that for sample sizes this large, the asymptotic efficiency properties of ML estimation are now in force, and the inference anomalies for SEM have essentially vanished.[17] However, the story is quite different for OLS, where over-rejection rates are seen to remain essentially the same at these larger sample sizes. So even in terms of this single example, it should be clear that in the presence of spatially autocorrelated x variables, over-rejection rates for OLS are not simply a "small sample" problem.

Given this range of illustrative examples, we turn now to the deeper question of what is actually causing this behavior. In the next section, we begin by analyzing the

[16] A more systematic analysis would of course involve tables of test-size values allowing variation in both ρ_y and ρ_x. However, our purpose here is simply to illustrate the key properties of these testing procedures. Our main objective is to *explain* these properties in geometric terms.

[17] This still leaves open the over-rejection problem for SEM seen in smaller samples such as Table 1d above. We shall return to this issue in the concluding section of the paper.

robust inference properties of OLS in the presence of only spatial autocorrelation in the y variable. This will be followed in Sect. 4 by a consideration of full spatial dependence in both the y variable and explanatory variables.

3 OLS inference with spatially independent explanatory variables

To analyze the inference properties of OLS, it is most natural to begin with a full linear model involving many explanatory variables. In this setting, our first objective is to show that inference about an individual beta parameter can be developed by reduction to an appropriate simple linear model that produces the same estimate of this beta parameter. This forms the basis of our present geometric approach to inference and also helps to justify our primary focus on the case of a single explanatory variable. It should be emphasized at the outset that this reduction is by no means new and is based on the *Frisch-Waugh-Lovell (FWL) theorem*, as developed thoroughly in Chapter 2 of Davidson and MacKinnon (2004). Our present treatment draws more heavily on their earlier analysis of regression inference in Davidson and MacKinnon (1993, Section 3.5).

3.1 Cosine representation of F-statistics

Given the linear regression model,

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n), \tag{10}$$

with an $n \times k$ matrix, X, of explanatory variables, we focus on the F-statistic, F_1, for a single beta coefficient, β_1. To do so, we first decompose $X\beta$ as

$$X\beta = (x_1, X_2)\begin{pmatrix}\beta_1 \\ \beta_2\end{pmatrix} = x_1\beta_1 + X_2\beta_2 \tag{11}$$

where β_2 denotes the $(k-1)$-vector of all beta coefficients other than β_1. In these terms, it is a direct consequence of the FWL Theorem that if the *orthogonal projection* into the *orthogonal complement*, $X_2^\perp$, of X_2 is denoted by

$$M_2 = I_n - X_2(X_2'X_2)^{-1}X_2' \tag{12}$$

so that by definition,

$$M_2' = M_2, \quad M_2M_2 = M_2, \quad M_2X_2^\perp = X_2^\perp, \quad \text{and} \quad M_2X_2 = 0, \tag{13}$$

then by Eqs. 10 and 11 it follows that

$$M_2y = M_2x_1\beta_1 + M_2X_2\beta_2 + M_2\varepsilon = M_2x_1\beta_1 + M_2\varepsilon \tag{14}$$

Hence, by defining the modified data sets, $\tilde{y} = M_2y$ and $\tilde{x}_1 = M_2x_1$, we obtain a *simple linear model*,

$$\tilde{y} = \tilde{x}_1\beta_1 + \tilde{\varepsilon} \tag{15}$$

with normally distributed error term, $\tilde{\varepsilon} = M_2\varepsilon$.[18] This reduced form focusing on β_1 allows the F-statistic for inferences about β_1 to be given a simple geometric interpretation. In particular, if we recall that the cosine, $\cos(z, v)$, of the angle between two n-vectors, z and v, is given by:

$$\cos(z, v) = \frac{z'v}{\|z\|\|v\|} \tag{16}$$

then it is shown in Appendix 1 (ESM) that:[19]

Proposition 1 *For any given set of data* (y, x_1, X_2) *in regression model* (Eqs. 10, 11), *the F-statistic for testing the null hypothesis,* $\beta_1 = 0$, *is given by*[20]

$$F_1 = (n-k)\frac{\cos^2(M_2y, M_2x_1)}{1-\cos^2(M_2y, M_2x_1)}. \tag{17}$$

As a special case of Eq. 17, observe that for the *simple linear model*,

$$y = \beta_0 1_n + \beta_1 x + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n) \tag{18}$$

with $1_n = (1, \ldots, 1)'$, it follows from Eq. 11 that $X_2 = 1_n$, and thus that M_2 is given by the *deviation matrix*,

$$M = I_n - 1_n(1_n'1_n)^{-1}1_n' = I_n - \left(\frac{1}{n}\right)1_n1_n'. \tag{19}$$

Hence, for this case, the F-statistic in Eq. 17 is of the form:

$$F_1 = (n-2)\frac{\cos^2(My, Mx)}{1-\cos^2(My, Mx)} \tag{20}$$

While our main focus will be on the simple regression case in Eq. 18, it should be clear from Eq. 17 that these results are directly extendable to parameter inference for individual explanatory variables in the multiple regression case. We shall return to this question in Sect. 3.3 below. Note also that angles are only defined between *nonzero* vectors. Hence, throughout this analysis, we shall implicitly ignore the (measure-zero) set of exceptional realizations with either $My = 0$ or $Mx = 0$.

3.2 Invariance theorem for OLS

In this section, we attempt to clarify the robustness of OLS inference with respect to spatial autocorrelation in y when x is spatially independent (as was seen in Table 1a). Given the cosine representation of F_1 above, our approach exploits the geometric properties of this representation. To illuminate these properties

[18] Here, it should be noted that the covariance matrix, $\sigma^2 M_2$, of $\tilde{\varepsilon}$ has rank $n - 1$, so that technically $\tilde{\varepsilon}$ has a *singular* normal distribution. This can easily be remedied by replacing M_2 with an equivalent reduced form matrix of full column rank, as developed in Appendix 2 of the supplementary material.

[19] All appendices are included in the supplementary material for this paper and can also be found in the longer version of this paper, Smith and Lee (2011a), available on line.

[20] Davidson and MacKinnon (1993) also point out that the associated t statistic for this null hypothesis (which is simply the (signed) square route of F_1) corresponds to the *cotangent* of the angle.

graphically, it is convenient to focus on the simple linear model in Eq. 18 and hence on the specific representation of F_1 in Eq. 20. For this reason, we here state the result specifically in terms of Eq. 20. Assuming that the true model of data (y, x) data is a JSE model as in Eqs. 5 through 7, our result is stated most easily by making the underlying spatial dependency parameters (ρ_y, ρ_x) explicit in F_1 as $F_1(\rho_y, \rho_x)$. Next, if we write $Z \underset{D}{=} V$ whenever random variables Z and V are *identically distributed*, then our first key result is to show that:

Theorem 1 (OLS invariance) *If (y, x) are generated by a JSE model, then for all* $\rho_y \in [0, 1)$,

$$F_1(\rho_y, 0) \underset{D}{=} F_1(0, 0). \tag{21}$$

In other words, if there is no spatial dependency in x, i.e., $\rho_x = 0$, then the distribution of the F-statistic in Eq. 20 is *independent* of the degree of spatial autocorrelation in y. Most importantly, this implies that all the usual optimality properties for tests of $\beta_1 = 0$ in the classical regression case $[(\rho_y, \rho_x) = (0, 0)]$ continue to hold regardless of the value of ρ_y.

A more general statement of Theorem 1 is given in Sect. 3.3 below. The advantage of the present more limited formulation is to allow a simpler geometric explanation of the underlying reasons for this invariance property. To illuminate the relevant geometry here, we begin by observing that if there is no spatial dependency in x, then $\rho_x = 0$ implies that $\upsilon_x = \varepsilon_x$ and hence that Eq. 6 reduces to

$$x = \mu_x 1_n + \varepsilon_x \tag{22}$$

where $\varepsilon_x \sim N(0, \sigma^2 I_n)$ by Eq. 7. Given this reduced (linear model) form of Eq. 6, observe next from Eq. 20 that F_1 does not directly involve the angle between y and x, but rather the angle, θ, between their *projected images*, My and Mx, in the orthogonal complement, $1_n^\perp$, of the unit vector, 1_n, as shown in Fig. 2 below. (In order to allow a meaningful graphical representation, we here focus on samples of size $n = 3$.)[21] Since $M1_n = 0$ by definition, this in turn implies from Eq. 14 that

$$Mx = \mu_x M 1_n + M\varepsilon_x = M\varepsilon_x. \tag{23}$$

Of particular importance from a geometric viewpoint is the fact that M is an *orthogonal projection* and hence *maps spheres into spheres of smaller dimension.* This implies that the spherical contours of the normal density of ε_x (in $\mathbb{R}^n$) are mapped by M to smaller dimensional spherical contours in the $(n - 1)$-dimensional image space of M (i.e., the orthogonal complement, $1_n^\perp$, of the unit vector, 1_n). Hence, the contours of the (singular) normal density, φ, of Mx concentrated on $1_n^\perp$ are necessarily spherical.[22] For the case of $n = 3$, these contours must be *concentric circles* on $1_3^\perp$, as shown in Fig. 3 below. Note also that while the variable, x, may have a nonzero mean, μ_x, in Eq. 6, its projected image, Mx, always has *zero mean* since by Eq. 23,

[21] Fortunately, such samples are just large enough to yield nontrivial estimates of slopes such as β_1.

[22] As shown in Appendix 2 in ESM, this singular density can be replaced by a proper density, $\phi(U'x)$, where U are eigenvectors for the non-null eigenvalues of M.

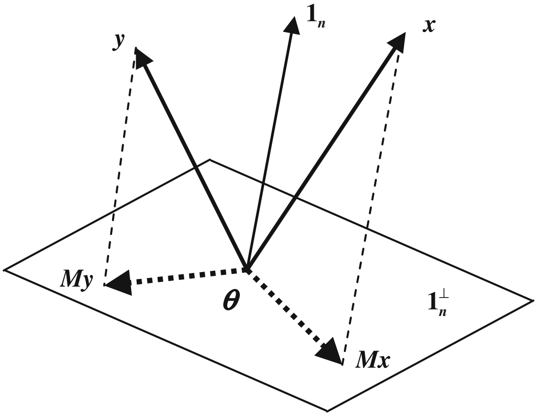


Fig. 2 Projected angle for $n = 3$

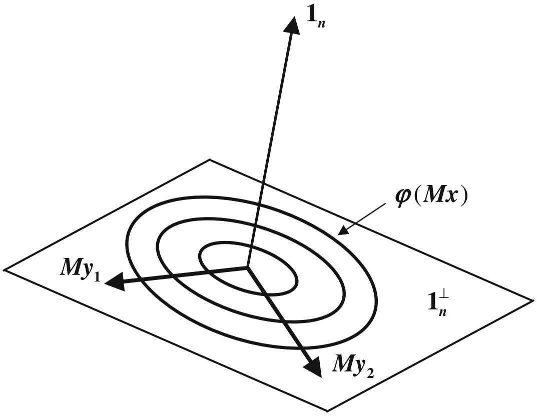


Fig. 3 Spherical invariance for $n = 3$

$$E(Mx) = ME(\varepsilon_x) = 0. \tag{24}$$

Hence, the spherical contours of the density, $\varphi(Mx)$, are necessarily centered about the origin, as also shown for the $n = 3$ case in Fig. 3.

Now for any given values of y, say $y = y_1$ and $y = y_2$, with projections, My_1 and My_2, shown in Fig. 3, consider the *conditional* distributions of the angles, $\theta(My_1, Mx)$ and $\theta(My_2, Mx)$. Since these distributions depend only on the density, $\varphi(Mx)$, it then follows from the *rotational symmetry* of this density that these two distributions must be the same, i.e., that

$$\theta(My_1, Mx) \underset{D}{=} \theta(My_2, Mx). \tag{25}$$

To see this, note simply that if the subspace, $1_n^{\perp}$, is rotated about the origin by a transformation, R_{12}, which maps My_1 to $My_2 = R_{12}\,(My_1)$, then every angular event $\theta(My_1, Mx) \in \Theta_1$ is mapped into a corresponding angular event,

$$\theta(My_2, Mx) \in \Theta_2 = \{\theta(My_2, Mx) : \theta(My_1, Mx) \in \Theta_1\} \tag{26}$$

with *identical probability mass*.

To complete the invariance argument, note that the identity in Eq. 25 implies in particular that $\cos(My_1, Mx) \underset{D}{=} \cos(My_2, Mx)$, and hence from Eq. 22 that the corresponding conditional distributions of F_1 (ρ_y, 0) are identical, i.e., that

$$F_1(\rho_y, 0|y = y_1) \underset{D}{=} F_1(\rho_y, 0|y = y_2). \tag{27}$$

But since this in turn implies by definition that,

$$F_1(\rho_y, 0) = E_y\left[F_1(\rho_y, 0|y)\right] \equiv F_1(\rho_y, 0|y = y_1) \tag{28}$$

It follows (from the arbitrary choice of y_1) that the distribution of F_1 (ρ_y, 0) must be entirely *independent* of distribution of y. Finally, since ρ_y is simply a parameter of the y-distribution, we see in particular that the distribution of F_1 (ρ_y, 0) is independent of the value of ρ_y, and hence that Theorem 1 must hold. A more general statement of this result will be given in Sect. 3.3 below (and is given a more rigorous proof in Appendix 3 in ESM). But this geometric argument serves to illuminate the main idea.

Before proceeding, it should be noted that this geometric argument also suggests an obvious generalization of Theorem 1. In particular, the original density of x need not be spherical for the above result to hold. From Fig. 3, it is clear that as long as the *projection*, Mx, of x onto the orthogonal complement, $1_n^\perp$, has a spherical density *on* $1_n^\perp$, the above argument goes through in tact. This generalization has been established rigorously by Dutilleul (2008), who showed that for the normal case, this property can be characterized in terms of "circular" covariance matrices.[23]

Note finally from the *symmetry* of cos(My, Mx) in y and x that the roles of these two random vectors can be reversed.[24] Hence, an immediate corollary of this result is that for $\rho_x \in [0, 1)$, it must also be true that

$$F_1(0, \rho_x) \underset{D}{=} F_1(0, 0). \tag{29}$$

In other words, if y satisfies all conditions of the classical linear model, then the same argument shows that inference about β_1 is not influenced by the distribution (and in particular the spatial dependency) of x. When stated in this manner, the result might appear to be less surprising. Indeed, it might be argued that since the linear model in Eq. 18 implicitly focuses on the *conditional distribution* of y given x, the marginal distribution of x should have no effect whatsoever. However, this intuition, not correct, and in particular, breaks down when even the slightest spatial dependency is present in the dependent variable, y.

3.3 The multivariate case

It should be clear from Eq. 15 above that a more general statement of Theorem 1 is possible within a multivariate setting. To do so, we first extend the JSE model by including a set of additional explanatory variables, as described by the matrix, X_2,

[23] Dutilleul also credits earlier work on this topic by Huynh and Feldt (1970), and others.

[24] The consequences of this symmetry property have of course been noted by many authors, dating at least as far back as Fingleton (1999).

above. Here, we now assume in addition that (i) the first column of X_2 is 1_n (so that the regression contains an intercept), and that (ii) X_2 is of full column rank, $n - k$, so that the inverse $(X_2'X_2)^{-1}$ exists (and hence that M_2 in Eq. 3 is well defined as stated).[25] But our single most important new assumption (which we discuss further below) is that (iii) the explanatory variable of interest, namely x ($=x_1$), is *independent* of X_2, so that relation Eq. 6 continues to hold in tact. In this setting, the joint distribution of (y, x) given X_2 [satisfying assumptions (i), (ii), and (iii)] is now assumed to be a *conditional JSE model* of the following form:

$$y = X_2\beta_2 + \upsilon_y, \quad \upsilon_y = \rho_y W\upsilon_y + \varepsilon_y \tag{30}$$

$$x = \mu_x 1_n + \upsilon_x, \quad \upsilon_x = \rho_x W\upsilon_x + \varepsilon_x \tag{31}$$

$$\begin{pmatrix} \varepsilon_y \\ \varepsilon_x \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_y^2 I_n & \\ & \sigma_x^2 I_n \end{pmatrix}\right] \tag{32}$$

Here, for convenience, we have repeated (Eqs. 6, 7) as (Eqs. 31, 32). Hence, the only real difference from the JSE model is in Eq. 30 where all explanatory variables are included, and in particular, where μ_y in Eq. 5 now corresponds implicitly to the first component of β_2 in Eq. 30. With regard to the associated OLS model, it will here be convenient to include the conditioning variates, X_2, along with the spatial dependency parameters (ρ_y, ρ_x) and now write the F_1-statistic in Eq. 17 as $F_1(\rho_y, \rho_x, X_2)$, where again M_2 is given in terms of X_2 by Eq. 12. In these terms, we now have the more general version of this theorem:

Theorem 2 (OLS invariance) *For any data X_2 satisfying (i), (ii), and (iii), if the true distribution of (y,x) is given by the conditional JSE model above, then for all $\rho_y \in [0, 1)$,*[26]

$$F_1(\rho_y, 0, X_2) \underset{D}{=} F_1(0, 0, X_2) \tag{33}$$

(A complete proof of this result is given in Appendix 3 in ESM.) For the present, one can gain some insight by observing from the independence assumption (iii) that the only role of the additional explanatory variables in X_2 is to transform the orthogonal projection matrix from M in Eq. 19 to M_2 in Eq. 12. But since spheres are preserved under *all* orthogonal projections, it is not surprising that such an extension is possible.

In this multivariate setting, a more interesting question concerns the degree to which this result holds when there are *dependencies* between x and X_2. Here, the situation is far more complex. A key difference from the simple linear model case is that while x and y are necessarily independent under the null hypothesis, $\beta_1 = 0$ (together with normality), this hypothesis makes no assertions about the relation between x and X_2. So even if there were no spatial dependencies at all (i.e.,

[25] In fact, this second assumption is for convenience only and simply avoids the need to introduce generalized inverses.

[26] As shown in the Appendix, this result actually holds for all spatial dependency values, ρ_y, which yield a well-defined reduced form in Eq. 8 above, i.e., for which the matrix, $B_{\rho_y} = I_n - \rho_y W$, is nonsingular. But since our main interest is on nonnegative spatial dependencies, we choose here to focus on this case.

$\rho_y = 0 = \rho_{x_j}$, $j = 1, \ldots, k$), one would still encounter the same types of multicollinearity problems that distort inferences about β_1 in the classical OLS case. Hence, all that we are able to say at this point, based on simple continuity considerations (and borne out by simulation experimentations), is that if both spatial dependencies in x and statistical dependencies between x and X_2 are relatively mild, then OLS can be expected to perform reasonably well with respect to inferences about β_1.

4 OLS inference with sample-correlated explanatory variables

We turn now to the more important case of explanatory variables exhibiting spatial dependencies. Recall from the Columbus simulation results in Table 1b through 1e above that higher degrees of spatial dependency tended to produce higher levels of spurious significance in OLS tests of β_1, which we shall here refer to as *spurious correlation* between y and x (given X_2). Moreover, as illustrated by the larger Philadelphia example in Table 1f, this problem of spurious correlation does not disappear with larger sample sizes. Hence, our objective in the present section is to offer a geometric explanation of this phenomenon.

To develop this explanation, we begin by noting one key property of the F_1-statistic that holds the general case of Eq. 17. To do so, we start with the conditional JSE model [which is implicitly taken to include assumptions (i), (ii), and (iii)] and rewrite this model in *reduced form* (paralleling Eqs. 8 and 9 above) as:

$$y = X_2\beta_2 + (I_n - \rho_y W)^{-1}\varepsilon_y = X_2\beta_2 + B_{\rho_y}^{-1}\varepsilon_y \tag{34}$$

$$x = \mu_x 1_n + (I_n - \rho_x W)^{-1}\varepsilon_x = \mu_x 1_n + B_{\rho_x}^{-1}\varepsilon_x. \tag{35}$$

Next, to analyze the projected random vectors, $M_2 y$ and $M_2 x$, which define the F_1-statistic, recall first from the argument in Eq. 14 that

$$M_2 y = M_2 X_2 \beta_2 + M_2 B_{\rho_y}^{-1}\varepsilon_y = M_2 B_{\rho_y}^{-1}\varepsilon_y \tag{36}$$

Similarly, since 1_n is a column vector in X_2, it also follows from Eq. 13 that $M_2 1_n = 0$ and hence that

$$M_2 x = \mu_x M_2 1_n + M_2 B_{\rho_x}^{-1}\varepsilon_x = M_2 B_{\rho_x}^{-1}\varepsilon_x \tag{37}$$

Thus, we see that each of these random vectors has *mean zero* since

$$E(M_2 y) = M_2 B_{\rho_y}^{-1} E(\varepsilon_y) = 0, \text{ and} \tag{38}$$

$$E(M_2 x) = M_2 B_{\rho_x}^{-1} E(\varepsilon_x) = 0. \tag{39}$$

In short, if the true model of (y,x) given X_2 is a conditional JSE model, then the key F_1-statistic in Eq. 17 depends on angle between two *independent zero-mean* random vectors, $M_2 y$ and $M_2 x$.

As will become clear below, the present notion of "spurious correlation" in terms of angles between zero-mean random vectors is in fact quite general and in particular has nothing to do with "spatial correlation" between vector components. Indeed, similar problems arise from almost any form of statistical dependencies between components.[27] Moreover, the underlying nature of spurious correlation can in fact be made more transparent by considering this question in a somewhat broader setting. Hence, our strategy here will be to develop the main ideas in terms of general correlated samples and then return to the case of spatially correlated samples.

To simplify the present discussion, we start by considering an arbitrary pair of *independent zero-mean* random vectors, $z = (z_1, \ldots, z_n)'$ and $w = (w_1, \ldots, w_n)'$ distributed on $\mathbb{R}^n$.[28] If this pair of random vectors (z,w) is treated as a sample of size n from a joint statistical population, then from a geometrical perspective, their *sample correlation*, $r(z, w)$, is precisely the cosine in Eq. 16, i.e.,

$$r(z,w) = \frac{\sum_{i=1}^{n} z_i w_i}{\sqrt{\sum_{i=1}^{n} z_i^2}\sqrt{\sum_{i=1}^{n} w_i^2}} = \frac{z'w}{\|z\|\|w\|} = \cos(z,w). \tag{40}$$

In these terms, our geometric explanation of spurious correlation will focus on the structural differences between sample correlations, $r(z, w)$, under conditions of (i) *independent* random sampling and (ii) *dependent* (correlated) random sampling.

4.1 The perfect-correlation case

The classic properties of sample correlations under independent random sampling are well known. So the key differences that arise under dependent random sampling are seen most easily by examining the most extreme case. In particular, suppose that the individual components (samples), z_i, of random vector z are all *perfectly correlated*. As is well known, this implies (for zero-mean random vectors) that all components are linearly dependent and in particular can be written as linear functions of the first component, z_1, as follows:

$$z_i = a_i z_1, \quad i = 2, \ldots, n \tag{41}$$

Hence, letting the vector, $a = (1, a_2, \ldots, a_n)$, denote the *dependency structure* of these components, it follows that z can be written simply as,

$$z = z_1 a \tag{42a}$$

with *fixed* dependency structure, a, and *random scale*, z_1 (which may be negative).[29] In a similar manner, if the components of $w = (w_1, \ldots, w_n)$ are also *perfectly correlated*, then there must be a fixed dependency structure, $b = (1, b_2, \ldots, b_n)$ such that

$$w = w_1 b \tag{42b}$$

[27] In fact, spurious correlation can even arise for completely *independent* x-samples and y-samples. In particular, if such samples are *heteroscedastic*, then such differences in variation can produce non-spherical distributions that have the same effects as those for correlated samples. An explicit example of this type is developed in the longer version of this paper (Smith and Lee 2011a, Section 4.3).

[28] The transpose notation here indicates that these are by convention *column* vectors.

[29] Note also that since all components of z are completely determined by its first component, this n-vector is effectively a sample of size one.

Hence, if z and w are *both perfectly correlated,* then Eq. 42 yields a very simple form for the associated sample correlation, $r(z, w)$. In particular, since the *sign* of any nonzero number α is defined by $\mathrm{sgn}(\alpha) = \alpha/|\alpha|$, it follows from Eq. 40 that[30]

$$\begin{aligned}\cos(z,w) = \cos(z_1 a, w_1 b) &= \frac{(z_1 a)'(w_1 b)}{\|z_1 a\|\|w_1 b\|} = \frac{z_1 w_1 (a'b)}{(|z_1|\|a\|)(|w_1|\|b\|)} \\ &= \frac{z_1}{|z_1|}\frac{w_1}{|w_1|}\frac{a'b}{\|a\|\|b\|} = \mathrm{sgn}(z_1)\mathrm{sgn}(w_1)\cos(a,b)\end{aligned} \tag{43}$$

But since this in turn implies that $|\cos(z,w)| = |\cos(a,b)|$, we may conclude from Eq. 40 that

$$|r(z,w)| = |r(a,b)| \tag{44}$$

Thus, if $|r(z,w)|$ is now taken to represent the *degree of correlation* between z and w, then regardless of the joint distribution of random scalar variables z_1 and w_1, we see that this degree of correlation is completely determined by the *fixed* dependency structures a and b. So even if z_1 and w_1 are independent random variables, this degree of correlation can assume any value in [0,1], as determined by a and b. In particular, if the dependency structures of z and w are the same (i.e., if $a = b$), then we must have $|r(z,w)| \equiv 1$ and may conclude that there is a maximum degree of spurious correlation between z and w. (An explicit example of this case is given in Sect. 4.2 below.)

Note finally that if z_1 and w_1 are independent normally distributed random vectors, then the symmetry of this density about zero implies that $\Pr(z_1 > 0) = \Pr(z_1 < 0) = 1/2$. Hence, by the independence of z_1 and w_1, we see from Eqs. 40 and 43 that

$$\begin{aligned}\Pr[r(z,w) = r(a,b)] &= \Pr[\mathrm{sgn}(z_1)\mathrm{sgn}(w_1) = 1] \\ &= \Pr(z_1 > 0)\Pr(w_1 > 0) + \Pr(z_1 < 0)\Pr(w_1 < 0) = 1/2\end{aligned} \tag{45}$$

and similarly that

$$\begin{aligned}\Pr[r(z,w) = -r(a,b)] &= \Pr[\mathrm{sgn}(z_1)\mathrm{sgn}(w_1) = -1] \\ &= \Pr(z_1 > 0)\Pr(w_1 < 0) + \Pr(z_1 < 0)\Pr(w_1 > 0) = 1/2\end{aligned} \tag{46}$$

Thus, for the case of independent z and w, we may conclude that

$$E[r(w,z)] = \tfrac{1}{2}r(a,b) + \tfrac{1}{2}[-r(a,b)] = 0. \tag{47}$$

In other words, if z and w are each perfectly correlated but mutually independent, then their sample correlation, $r(z, w)$, is *on average equal to zero.*[31] Hence, the

[30] Recall again that we ignore the measure-zero cases in which $z_1 = 0$ and/or $w_1 = 0$.

[31] Note also that this result depends only on *symmetry* of both the z_1 distribution and w_1 distribution about zero and does not require normality.

distribution of $r(z, w)$ under perfect correlation agrees with classical independent sampling in this sense. However, its realized values are *never close to zero* and are thus in strong disagreement with independent random sampling—where it is well known that $r(z, w)$ must converge to zero almost surely as sample size becomes large. This is the essence of *spurious correlation.*

4.2 A simple class of sample-correlation models

The extreme example above constitutes the natural limiting case as individual components (samples) become more and more correlated. One useful feature of this limiting case is to show that spurious correlation has nothing in particular to do with *spatial* correlation. Moreover, in a manner similar to the opposite extreme of independent samples (i.e., spherical distributions), the mathematical simplicity of this case allows exact results to be obtained for all sample sizes, n. However, the intermediate cases of partial dependence are analytically more complex. Hence, to gain geometric insights, we here develop a family of simple models that (i) allow the relevant sample correlations to be parameterized explicitly and (ii) allow the model properties to be displayed graphically for the $n = 3$ case.

To do so, we begin with the simple class of JSE models in Eqs. 5–7 and relax the spatial autocorrelation specification to allow a more direct parameterization of the relevant sample correlations. An explicit full-dimensional reduced form of this model is then developed to analyze the consequences of such correlations in the $n = 3$ case. Here, we again start with simple regression model,

$$y = \beta_0 1_n + \beta_1 x + \upsilon_y, \quad \upsilon_y \sim N(0, \sigma^2 I_n) \tag{48}$$

in Eq. 1 above, but now assume that the true model is given by the following relaxed version of JSE models:

$$y = \beta_0 1_n + \upsilon_y \tag{49}$$

$$x = \mu_x 1_n + \upsilon_x \tag{50}$$

$$\begin{pmatrix} \upsilon_y \\ \upsilon_x \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_y & \\ & \Sigma_x \end{pmatrix}\right] \tag{51}$$

Essentially, this relaxed version is an instance of a general linear model in which the independent normal residuals, υ_y and υ_x, are allowed to have arbitrary covariance structures.

Within this broader framework, we next make a number of simplifying assumptions that will yield an explicit class of models encompassing the full range of correlation possibilities. First, to allow graphical representations of key results, we again adopt the small-sample framework ($n = 3$) in Sect. 3.2. In this context, it can be shown (see Appendix 2 in ESM) that if the (3×3) projection matrix, M, is replaced by the (2×3) transformation:[32]

[32] In the more general development in the Appendix 2 in ESM, T is an instance of the matrix, U_2', for $n = 3$.

$$T = \begin{bmatrix} T_1' \\ T_2' \end{bmatrix} = \begin{bmatrix} (0 & -1/\sqrt{2} & 1/\sqrt{2}) \\ (-2/\sqrt{6} & 1/\sqrt{6} & 1/\sqrt{6}) \end{bmatrix} \tag{52}$$

which is easily verified (by direct multiplication) to satisfy the two identities,

$$TT' = I_2, \quad T'T = M \tag{53}$$

then, we obtain a well-defined 2-dimensional model by simply multiplying Eqs. 48–51 to obtain the following *reduced hypothesized model* from Eq. 48,

$$Ty = \beta_1 Tx + T\upsilon_y, \quad T\upsilon_y \sim N(0, I_2), \tag{54}$$

together with the following *reduced true model* from Eqs. 49 through 51,

$$Ty = T\upsilon_y, \tag{55}$$

$$Tx = T\upsilon_x, \tag{56}$$

$$\begin{pmatrix} T\upsilon_y \\ T\upsilon_x \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} T\Sigma_y T' & \\ & T\Sigma_x T' \end{pmatrix}\right]. \tag{57}$$

The key point to notice here is that the transformed residuals, $T\upsilon_y$ and $T\upsilon_x$, are now proper 2-dimensional random vectors with *nonsingular* (2×2) covariance matrices, $T\Sigma_y T'$ and $T\Sigma_x T'$, respectively. This particular transformation not only yields a full-dimensional model but also satisfies the identity (see Appendix 2 in ESM),

$$\cos(My, Mx) \equiv \cos(Ty, Tx) \tag{58}$$

and hence is seen from Eq. 20 to yield precisely the same F_1-statistic as M. Thus, all analyses can be carried out in terms of these new reduced models. To visualize these models more clearly, recall that the singular normal densities, $\varphi(My)$ and $\varphi(Mx)$, are both concentrated on the subspace, $1_n^{\perp}$, shown for $n = 3$ [using $\varphi(Mx)$] in Fig. 3 above. Hence, for this $n = 3$ case, the transformation T essentially collapses $\mathbb{R}^3$onto the 2-dimensional subspace, $1_3^{\perp}$, and (in terms of Eq. 52) maps each vector, $z \in \mathbb{R}^3$, into the 2-dimensional pair $(T_1' z, T_2' z)$. For example, the contours of $\varphi(Mx)$ in Fig. 3 now correspond precisely to the density contours, $\phi(Tx) = \phi(T_1' x, T_2' x)$, of the proper bivariate normal distribution, $N(0, T\Sigma_x T')$, given by Eqs. 56 and 57.

With these preliminaries, the family of models to be developed requires only a specification of the true covariance matrices, Σ_x and Σ_y. Here, our objective is to obtain a simple one-parameter family of covariance structures that range between "no correlation" and "perfect correlation" (as in Sect. 3.1 above). If the family of all nonnegative nonsingular (2×2) *correlation matrices* is now denoted by:

$$R_\tau = \begin{pmatrix} 1 & \tau \\ \tau & 1 \end{pmatrix}, \quad \tau \in [0, 1) \tag{59}$$

then for each *correlation parameter*, τ,[33] the relevant *covariance matrix*, Σ_τ, for our purposes can be defined in terms of transformation, T, by:[34]

[33] The symbol τ is employed here to avoid confusion with spatial dependency parameters, ρ.

[34] Note that Σ_r can also be expressed directly as a linear matrix function of τ. In particular, it may be verified that in terms of the row representation of T in Eq. 52, $\Sigma_\tau = T'T + \tau(T_1 T_2' + T_2 T_1')$.

$$\Sigma_\tau = T' R_\tau T. \tag{60}$$

In particular, we assume that the covariance matrices for Ty and Tx are precisely of this form, i.e., that

$$\Sigma_y = \Sigma_{\tau_y} \quad \text{and} \quad \Sigma_x = \Sigma_{\tau_x} \quad \text{for some } \tau_y, \tau_x \in [0, 1) \tag{61}$$

For convenience, we now designate (Eqs. 55–57, 61) as the (τ_y, τ_x)-*model*.[35] Note first that for each such model, the (3×3) covariance matrices, Σ_{τ_y} and Σ_{τ_x}, are clearly of rank 2 and hence (in a manner similar to *M*) are singular. But by the first equality in Eq. 53, we see that the reduced matrix

$$T\Sigma_\tau T' = (TT')R_\tau(TT') = R_\tau \tag{62}$$

is *nonsingular* for each $\tau \in [0, 1)$. Hence, by Eqs. 55–56 together with Eq. 61, it follows that for each (τ_y, τ_x)-model, the joint distribution of Ty and Tx can be rewritten more simply as:

$$\begin{pmatrix} Ty \\ Tx \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} R_{\tau_y} & \\ & R_{\tau_x} \end{pmatrix}\right]. \tag{63}$$

Given this representation, recall next from Eq. 20 that for both the y and x vectors, the relevant correlations for the F_1-statistic are not those between their individual components, but rather those between the components of their images, Ty and Tx. Hence, it should now be evident from Eq. 62 that this is precisely what is captured by the pair (τ_y, τ_x). In particular, we see from Eq. 63 (together with Eq. 52) that for the random vector, $Ty = (T_1' y, T_2' y)'$, the correlation between components $T_1' y$ and $T_2' y$ is precisely τ_y. Similarly for $Tx = (T_1' x, T_2' x)'$, it follows that τ_x is the correlation between $T_1' x$ and $T_2' x$. Thus, the main advantage of (τ_y, τ_x)-models is that they allow a direct parameterization of the relevant correlations influencing F_1.

To examine the consequences of such correlations, we begin by making the additional simplifying assumption that the two correlations (τ_y, τ_x) are *identical*, i.e., that ($\tau_y = \tau_x = \tau$) for some $\tau \in [0, 1)$. This special case, which may be referred to as simply the τ-*model*, will serve to facilitate our geometric explanation of spurious correlation. Here, it suffices to compare the extreme cases of *independent samples* ($\tau = 0$) and *perfectly correlated samples* ($\tau = 1$) with a representative intermediate case, $\tau = 0.8$. These three cases are developed successively in the following three subsections, followed by an illustration of a more general intermediate case with $\tau_y \neq \tau_x$.

4.2.1 The independent sampling case

First, it is important to show that the *independent sampling* case (for $n = 3$) does indeed correspond to $\tau = 0$ within the family of τ-models. This is not obvious, since

[35] Note that as a parallel to JSE models, the spatial dependency parameters (ρ_y, ρ_x) are here replaced by the correlation parameters (τ_y, τ_x). However, this simple parameterization is only possible in the present setting for the $n = 3$ case.

independence requires $\mathrm{cov}(x) = \sigma^2 I_3$. But for the τ-model with $\tau = 0$, we see from Eqs. 60 and 61 together with the identity $R_0 = I_2$ that

$$\mathrm{cov}(x) = TI_2T' = T\,T' = M \neq \sigma^2 I_3. \tag{64}$$

So even by setting $\sigma^2 = 1$, it is not evident that independent sampling is included in the present class of τ-models. However, by examining the *reduced form* of this independent sampling model (with $\sigma^2 = 1$), we see from Eq. 53 that

$$\mathrm{cov}(Tx) = TI_3T' = TT' = I_2 = R_0. \tag{65}$$

Similarly, since the same argument shows that $\mathrm{cov}(Ty) = R_0$ for independently sampled y components, it follows that Eq. 63 does indeed hold with $\tau_y = \tau_x = 0$.[36] Moreover, since it is only these image vectors, Ty and Tx, that are relevant for the F_1-statistic in Eq. 22, we see that for our present purposes, the independent sampling case is adequately represented within the framework of τ-models.

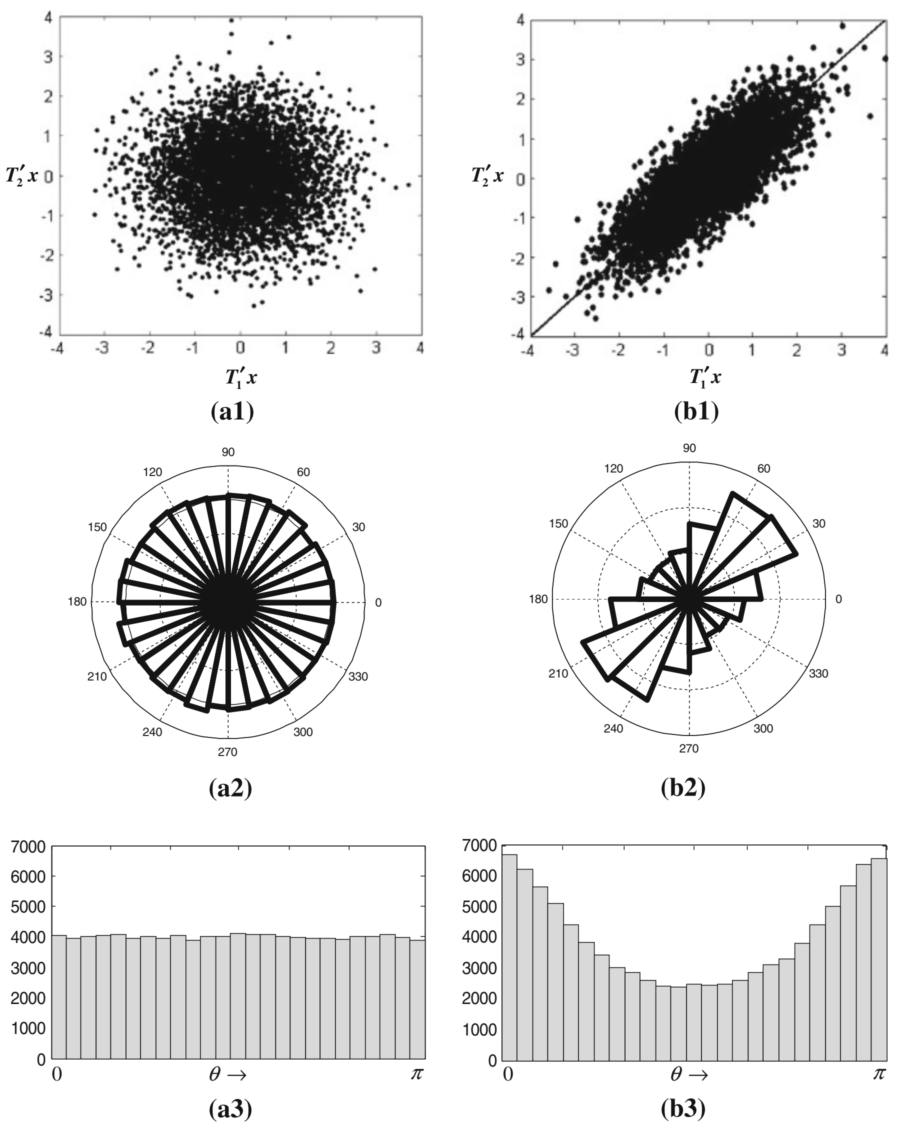
With this preliminary observation, the distributions of both Tx and Ty can be obtained by sampling from Eq. 63 with $\tau_y = \tau_x = 0$. Since these distributions are identical, it suffices for the moment to focus on Tx with (standard normal) bivariate distribution, $Tx = (T_1'x, T_2'x)' \sim N(0, I_2)$. A scatter plot of 5,000 simulated draws from this distribution is shown in panel (a1) of Fig. 4. (The panels (b1), (b2), and (b3) are included to facilitate a comparison of the $\tau = 0$ and $\tau = 0.8$ cases, and may be ignored for the present). As we have already seen from Sect. 3.1, this scatter plot simply reflects the underlying circular symmetry of $N(0, I_2)$. Most important for our purposes are the *directional frequencies* of these points (as vectors from the origin). These frequencies are depicted by the directional histogram in panel (a2), where the length of each directional (pie-shaped) sector is proportional to the frequency of vectors in that sector. Since these frequencies are seen to be virtually identical in each sector, it is clear that such directions are indeed completely random.[37] This process was repeated for an independent set of 5,000 simulated draws of $Ty = (T_1'y, T_2'y)'$ from the same distribution. The resulting scatter plot and directional frequencies for Ty are virtually identical to those for Tx and are not shown.

Given these two independent samples, if we compute the angles, $\theta\,[(Ty)_s, (Tx)_s]$, between each pair of simulated vectors $[(Tx)_s, (Ty)_s], s = 1, \ldots, 5{,}000$, then the sample histogram of these values (in radians) can be plotted, as shown in panel (a3) of Fig. 4. As expected, this histogram is again seen to be almost flat, indicating that these angles are completely random. Hence, the cosines of these angles must be also consistent with complete randomness,[38] implying that F_1 in Eq. 20 is indeed

[36] Note that this argument is in fact an instance of the more general geometric fact (mentioned in Sect. 3.2 above) that orthogonal projections of spheres are always spheres of lower dimension.

[37] While the simulation size, 5,000, yields a clear visual scatter plot in panel (a1), it is not sufficiently large to overcome the extreme variation in samples of size $n = 3$. Hence, all histograms in this section [such as in panels (a1) and (a3)] are based on much larger simulations of 100,000 draws. At this simulation size, the true shape of each sampling distribution [such as the uniform distribution in panel (a3)] is much more evident.

[38] A histogram of these cosine values is somewhat less informative since the cosine itself is a very nonlinear function.



(a1)

(b1)

(a2)

(b2)

(a3)

(b3)

Fig. 4 Comparison of $\tau = 0$ and $\tau = 0.8$. **a1** Tx scatter plot for $\tau = 0$. **a2** Tx angle histogram for $\tau = 0$. **a3** (Ty, Tx) angle histogram for $\tau = 0$. **b1** Tx scatter plot for $\tau = 0.8$. **b2** Tx angle histogram for $\tau = 0.8$. **b3** (Ty, Tx) angle histogram for $\tau = 0.8$

F-distributed in accordance with the null hypothesis, $\beta_1 = 0$, for the classical OLS model in Eq. 48. Moreover, although the sample size ($n = 3$) is extremely small, it is just large enough to yield a nondegenerate F(1,1)-distribution (with numerator and denominator degrees of freedom both equal to 1). Hence, the size of the OLS test in this case can be checked by computing the F_1-statistic for each angle and comparing it with the 0.05 rejection level.[39] The estimated size of this test, 0.0498 ($\approx$0.05), confirms that the sampling distribution is indeed consistent with F(1,1), and hence that there is no over-rejection problem for the independent sampling case.[40]

4.2.2 The perfectly correlated sampling case

To develop the perfect-correlation case in the present setting, note first that $\tau = 1$ is *not* included in the definition of τ-models. Indeed, the covariance matrices, $R_1 = 1_2 1_2'$, for both Ty and Tx under this model add additional singularities that must be treated separately. To do so, we focus on Tx. For this case, note from Eq. 56 that since $Tx = T\upsilon_x$ and since we are only interested in this reduced model, we can assume without loss of generality that $x = \upsilon_x$, and hence that $E(x) = 0$. But for this case, we have already seen from Sect. 4.1 that x is perfectly correlated if and only if

$$x = x_1 a \tag{66}$$

for some fixed *dependency structure*, $a = (1, a_2, a_3)'$. Moreover, by rescaling x if necessary, we may also assume that

$$\text{var}(x_1) = E(x_1^2) = 1. \tag{67}$$

Within this framework, we can directly construct a dependency structure that will yield a proper (one-dimensional) representation of the singular normal distribution, $N(0, R_1)$, as follows. Observe that if the dependency structure, a, is chosen to satisfy

$$Ta = 1_2 \tag{68}$$

and for any standard normal variate, $x_1 \sim N(0,1)$, we set $x = x_1 a$, then it will follow by construction that x has perfectly correlated components, and moreover that

$$\text{cov}(Tx) = \text{cov}[T(x_1 a)] = \text{cov}[x_1(Ta)] = Ta\,\text{var}(x_1)\,(Ta)' = 1_2(1)\,1_2' = R_1. \tag{69}$$

Hence, it suffices to find a solution, a, of Eq. 68 with $a_1 = 1$. This can be accomplished by letting $e_1 = (1, 0, 0)'$ (so that $a_1 = e_1' a$) and then obtaining a as the *unique* solution of the augmented linear equation system,

$$\begin{bmatrix} T_1' \\ T_2' \\ e_1' \end{bmatrix} a = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \Rightarrow a = \begin{bmatrix} T_1' \\ T_2' \\ e_1' \end{bmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 + \left(3/\sqrt{6}\right) - \left(1/\sqrt{2}\right) \\ 1 + \left(3/\sqrt{6}\right) + \left(1/\sqrt{2}\right) \end{pmatrix} \tag{70}$$

(where the explicit solution for a on the right-hand side can be verified by simply multiplying out the first expression). Finally, by letting $y = y_1 a$ with $y_1 \sim N(0, 1)$ independent of x_1, we will obtain a pair of well-defined independent random vectors, (y, x), with perfectly correlated components that are consistent with the limiting form of the τ-model with $\tau = 1$. Note that by construction, the realized values of

[39] For this small sample size, the corresponding rejection level is enormous: F(0.05;1,1) = 161.45.

[40] As in footnote 37 above, each estimated test size in this section (such as the present value of 0.0498) is based on a larger simulation of 100,000 draws to overcome sampling variation.

$$(Ty, Tx) = (y_1 Ta, x_1 Ta) = (y_1 1_2, x_1 1_2) \tag{71}$$

must simply be independent pairs of points on the 45° line, each with distances and directions from the origin determined by realization of the normal variates, y_1 and x_1. Hence, there is no need to simulate this case. In fact, it suffices to consider two possible realizations of these random variables as shown in Fig. 5 below.

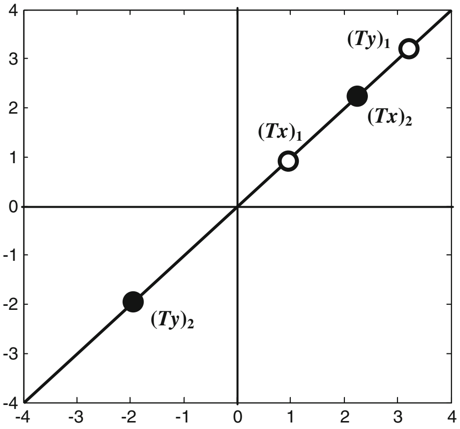
Here, the first realization $[(Ty)_1, (Tx)_1]$ (depicted by hollow points) shows a case in which $(Ty)_1$ and $(Tx)_1$ are both in the same direction from the origin, so that $r[(Ty)_1, (Tx)_1] = 1$. Similarly, the second realization $[(Ty)_2, (Tx)_2]$ (depicted by solid points) shows a case with $(Ty)_2$ and $(Tx)_2$ in opposite directions from the origin, so that $r[(Ty)_2, (Tx)_2] = -1$. Since these are the only two possibilities (except for a set of measure zero), it follows that $|r(Ty, Tx)| = 1$ must hold identically.

This is of course simply an instance of the more general result established in Sect. 4.1 above. But the advantages of this particular extension of τ-models to the $\tau = 1$ case are twofold. First, the visual representation of this case shows exactly *why* the sample correlation between any two *independent* random vectors with the same (perfect-correlation) dependency structure must necessarily be maximal. Moreover, it serves as a natural benchmark for interpreting the range of intermediate τ-models, between $\tau = 0$ and $\tau = 1$.

4.2.3 The intermediate-correlation sampling case

Finally, we turn to the most important case of intermediate correlations. It suffices to illustrate this case by a single example, $\tau = 0.8$, which is sufficiently close to $\tau = 1$ to be viewed as a "neighbor" of this perfect-correlation case. If we focus on the correlated random vector, $Tx = (T_1'x, T_2'x)' \sim N(0, R_\tau)$, then it is clear that for large correlations, τ, the random pair $(T_1'x, T_2'x)$ will tend to be close together, and hence that realizations of Tx will tend to cluster around the 45° line. This is confirmed by the elliptical-shaped scatter plot in panel (b1) of Fig. 4 showing 5,000 simulated draws of $Tx = (T_1'x, T_2'x)'$ from $N(0, R_{0.8})$. Even more important however

Fig. 5 Possible perfect-correlation pairs ($\tau = 1$)



is that (in contrast to the $\tau = 0$ case of panel (a2)), the *directional frequencies* of these realizations also concentrate around the 45° line, as is seen in the directional histogram of panel (b2). Hence, when viewed as vectors from the origin, realizations of Tx tend to be directionally close to either the unit vector, 1_2, or its negative, -1_2. Similarly, if we independently simulate 5,000 draws of $Ty = (T_1'y, T_2'y)'$, then these realizations will exhibit the same properties as those of Tx and hence are not shown.

As in the $\tau = 0$ case, our main interest focuses on the frequency distribution of angles between each pair $[(Tx)_s, (Ty)_s], s = 1, \ldots, 5{,}000$. Here, the results above suggest the basic form of this distribution. In particular, whenever both Ty and Tx are directionally close to the 45° line [such as in the biggest four angular sectors of panel (b2)], they must either be pointing in roughly the *same or opposite direction*, so that their angle, $\theta(Ty, Tx)$, must either be close to $\theta = 0$ or $\theta = \pi$. Hence, there is necessarily a tendency for realized angles to cluster around these two extremes, as is seen in the histogram of panel (b3).

The continuum of possible cases is now clear. For the independent sampling case ($\tau = 0$), the sampling distribution of angles, $\theta(Ty, Tx)$, is completely uniform. But as τ increases, and the directional histogram becomes more concentrated around the 45° line, the sampling distribution of $\theta(Ty, Tx)$ becomes more concentrated at the end points, $\theta = 0$ and $\theta = \pi$. Finally, in the extreme perfect-correlation case ($\tau = 1$) where Ty and Tx are exactly on the 45° line, the directional histogram must be completely concentrated on this line, and the sampling distribution of $\theta(Ty, Tx)$ collapses to a two-point mass distribution at these end points. This continuum of possibilities for $\theta(Ty, Tx)$ in turn implies that as τ increases, the realized values of squared cosines, $\cos^2(Ty, Tx)$, must concentrate near unity. Hence, even though Ty and Tx are statistically independent, we see from Eq. 40 that the spurious correlation reflected by $r^2(Ty, Tx)$ will necessarily increase toward unity. Even more important for statistical inference is that the associated F_1-values in Eq. 20 will increase without bound. In short, as τ increases, the *size* of OLS tests for $\beta_1 = 0$ in model Eq. 48 must increase toward unity, so that over-rejection is virtually guaranteed. For the case of $\tau = 0.8$, the size of the OLS test procedure at the end of Sect. 4.2.1 now increases to 0.0839. Of course, little can be said about the actual rate of such increases without extensive simulation. But the central purpose of these τ-models is not to quantify such increases, rather it is to illuminate their underlying cause.

4.2.4 Example of a (τ_1, τ_2)-model

While the simplifying assumption of $\tau_y = \tau_x = \tau$ is useful for illustrative purposes, it is important to emphasize that such spurious correlation problems go well beyond the case of identically distributed Ty and Tx. As one example, suppose that $\tau_y = 0.8$ as above, but that $\tau_x = 0.4$. In this case, the distribution of Ty will tend to be more circular than that of Tx. While this can in principle be depicted by overlapping scatter plots in a manner similar to panels (a1) and (b1) of Fig. 4, it is more difficult

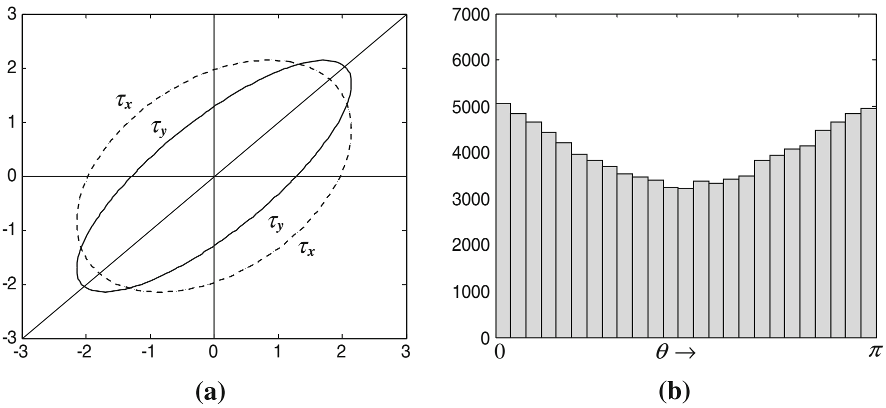


Fig. 6 Correlation properties for $(\tau_y, \tau_x) = (0.8, 0.4)$. **a** 90% contours for distributions. **b** (Ty, Tx) angle histogram

to see the relative shapes of these two patterns. One alternative approach is to plot representative probability contours for each distribution. This is done in panel (a) of Fig. 6, where the density contour containing 90% of all probability mass for distribution $N(0, R_{\tau_y})$ is shown as a solid ellipse, and that for $N(0, R_{\tau_x})$ is shown as a dashed ellipse.

The relatively more circular nature of the $N(0, R_{\tau_x})$ distribution is now evident. From the arguments above, the resulting shape of the sampling distribution for $\theta(Ty, Tx)$ in this case is easy to guess. In particular, it is natural to expect (by simple continuity) that the histogram for the (0.8, 0.4) case should be between that for (0.8, 0.8) and (0.8, 0). But by the (informal) proof of Theorem 1 above, we know that the distribution (0.8, 0) should be the same as that for (0, 0), since $\tau_x = 0$ in both cases implies that the underlying distribution of Tx remains perfectly circular. So the desired histogram should be roughly an averaging between panels (a3) and (b3) of Fig. 4. This is indeed the case, as can be seen from the actual simulated histogram in panel (b) of Fig. 6. As one final check, the size of the OLS test of $\beta_1 = 0$ in this (0.8, 0.4) case is 0.06211, which is roughly half way between the sizes, 0.0498 and 0.0839 for the (0.8, 0) and (0.8, 0.8) cases.

4.3 The spatial autocorrelation case

Given the broader results above, we now return to the case of spatial autocorrelation. Our objective is to show that the same range of cases produced by τ-models is also exhibited by JSE models with τ replaced by ρ. For ρ close to zero, the comparison is obvious. Like τ-models, all correlations between components of y and x (and hence of Ty and Tx) vanish as ρ approaches zero. Hence, the more interesting questions relate to the opposite extreme as ρ approaches one. In Sect. 4.3.1 below, it is shown that in the multiple regression setting of conditional JSE models, both Ty and Tx approach perfect-correlation cases in a manner similar to τ-models. This is followed in Sect. 4.3.2 with an illustration based on $n = 3$ paralleling τ-models.

4.3.1 A limit theorem for spatial error models

Here, we start with the hypothesized regression model in (Eqs. 10, 11) above and assume that the true model of y is given by

$$y = X_2\beta_2 + B_{\rho_y}^{-1}\varepsilon_y = X_2\beta_2 + (I_n - \rho_y W)^{-1}\varepsilon_y, \quad \varepsilon_y \sim N(0, \sigma_y^2 I_n) \tag{72}$$

which is simply the reduced form of Eq. 32 in the conditional JSE model. Here, our focus will be on y, since the model for x is simply the special case with $X_2 = 1_n$.

Up to this point, we have treated spatial weight matrices, W, in a rather informal manner. But the following result requires that we be more precise. Hence, we now define a *spatial weight matrix*, $W = (w_{ij}: i, j = 1, \ldots, n)$ to be any nonnegative matrix with zero diagonal ($w_{ij} = 0$: $i = 1, \ldots, n$).[41] The only additional condition we require is that no subregions be "isolated" in terms of spatial dependencies, i.e., that there be no subset if regions, $S \subset \{1, \ldots, n\}$ such that all spatial influences, w_{ij}, of regions $i \notin S$ on regions $j \in S$ are zero. Such weight matrices, W, are said to be *connected* in the sense that there must exist positive chains of spatial influence, $(w_{ik_1}, w_{k_1k_2}, \ldots, w_{k_mj})$, between each pair of regions i and j.[42] The key feature of *connected weight matrices*, W, for our present purposes is that the maximum eigenvalue, λ_1, of W is always positive and has unique positive eigenvector, v_1, of unit length, which we designate as the *maximal eigenvector* for W.[43] As above, we assume for convenience that W is scaled to have $\lambda_1 = 1$. In this setting, our main result is to show that as $\rho_y \to 1$, the realized values of y in Eq. 72 must eventually be approximately proportional to v_1 (see Appendix 4 in ESM):[44]

Theorem 3 (Limiting autocorrelation) *For any connected spatial weight matrix, W, with maximal eigenvector,* v_1, *and any random vector y satisfying the corresponding SEM* Eq. 72,

$$\lim_{\rho_y \to 1} \frac{y}{\|y\|} = \pm v_1 \quad \text{almost surely.} \tag{73}$$

The "almost surely" condition means that this limit will hold for all residual vectors, ε_y, in Eq. 72 except for a subset with probability zero. To understand the

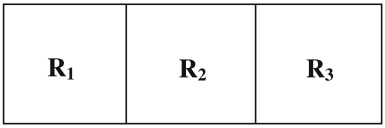
[41] Interestingly, the result to be developed does not require a zero diagonal. But the spatial error model itself requires zero diagonals to avoid self-referencing in the spatial autoregressive relation of Eq. 4 above.

[42] Such matrices are also said to be *irreducible* matrices. For further discussion of such weight matrices, see for example Appendix A in Martellosio (2010).

[43] The pairs (λ_1, v_1) are often designated as the *Perron* eigenvalue and eigenvector for W. See Lemma 1 in Appendix 4 in ESM for more detail.

[44] As mentioned in the introduction, this result is largely inspired by the work of Kramer and Donninger (1987) who developed a parallel result for the covariance matrix, $\mathrm{cov}(y) = \sigma^2(I_n - \rho_y W)^{-1}(I_n - \rho_y W')^{-1}$. A recent result closely related to Theorem 3 (in the context of testing for spatial autocorrelation) can be found the proof of Theorem 1 in Martellosio (2010).

Fig. 7 Three region case



notation, $\pm v_1$, recall that $v_1 > 0$ by definition. Hence, this notation implies that y components are either all positive or all negative.[45]

The key implication of this limiting result for our purposes is that (with the $\lambda_1 = 1$ normalization) the spatial dependency parameter, ρ_y, acts very much like the τ correlation coefficient in the sense that when $\rho_y \approx 1$, the components of y are almost perfectly correlated. More specifically, if we denote the common sign of all (nonzero) y components by sgn(y), and write Eq. 73 somewhat more loosely as

$$\rho_y \approx 1 \Rightarrow \frac{y}{\|y\|} \approx \operatorname{sgn}(y)\, v_1 \tag{74}$$

then (again ignoring possible zero values) it follows that for any component, $j = 1, \ldots, n$,

$$\begin{aligned} \rho_y \approx 1 \Rightarrow \left(\frac{y_1}{\|y\|} \approx \operatorname{sgn}(y) v_{11}\right) \quad \text{and} \quad \left(\frac{y_j}{\|y\|} \approx \operatorname{sgn}(y) v_{1j}\right) \\ \Rightarrow \frac{y_j}{y_1} \approx \frac{v_{1j}}{v_{11}} \equiv a_j \Rightarrow y_j \approx y_1 a_j, \quad j = 1, \ldots, n \Rightarrow y \approx y_1 a \end{aligned} \tag{75}$$

for this choice of $a = (1, a_2, \ldots, a_n)'$. Hence, we see that when $\rho_y \approx 1$, the random vector, y, is approximately *perfectly correlated* as in Eq. 42 above with dependency structure proportional to v_1.[46]

4.3.2 *Illustrations of spurious correlation effects*

As in Sect. 4.2 above, the consequences of Theorem 3 can be illustrated graphically for the simple case of $n = 3$. Here, we start with an $n = 3$ version of the proximity matrices used in the Columbus and Philadelphia examples above and, in particular, consider the "linear" three-region case in Fig. 7 with proximities as shown.

The appropriate scaled proximity weight matrix, W_3, is then given by

$$W_3 = \lambda_1^{-1} \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \tag{76}$$

where the maximum eigenvalue of the unscaled matrix is $\lambda_1 = \sqrt{2}$. By symmetry, W in this case has a spectral decomposition given by

$$W_3 = V \Lambda V' \tag{77}$$

[45] Note also that y should technically be indexed by ρ_y to indicate that each specified value of ρ_y implicitly defines a different random vector, y.

[46] The limiting properties of test sizes for this perfect-correlation case are studied in Kramer (2003).

with eigenvalues, $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \lambda_3) = \mathrm{diag}(1, 0, -1)$, and associated eigenvectors

$$V = (v_1, v_2, v_3) = \left(\frac{1}{2}\right)\begin{pmatrix} 1 & -\sqrt{2} & 1 \\ \sqrt{2} & 0 & -\sqrt{2} \\ 1 & \sqrt{2} & 1 \end{pmatrix} \tag{78}$$

and satisfying $V^{-1} = V'$. Hence, W_3 is seen to satisfy our assumptions with unique maximal eigenvector given by $v_1 = (1/2,\ \sqrt{2}/2,\ 1/2)'$. If we now consider the simple JSE model for y in Eqs. 5–7, then since

$$\begin{aligned} B_{\rho_y} &= I_3 - \rho_y W_3 = VV' - \rho_y V\Lambda V' = V(I_3 - \rho_y \Lambda)V' \\ \Rightarrow B_{\rho_y}^{-1} &= V(I_3 - \rho_y\Lambda)^{-1}V' \\ &= (1-\rho_y\lambda_1)^{-1}v_1v_1' + (1-\rho_y\lambda_2)^{-1}v_2v_2' + (1-\rho_y\lambda_3)^{-1}v_3v_3' \\ &= (1-\rho_y)^{-1}v_1v_1' + v_2v_2' + (1+\rho_y)^{-1}v_3v_3' \end{aligned} \tag{79}$$

we see that

$$\begin{aligned} y &= \mu_y 1_n + [(1-\rho_y)^{-1}v_1v_1' + v_2v_2' + (1+\rho_y)^{-1}v_3v_3']\varepsilon_y \\ &= \mu_y 1_n + \left(\frac{v_1'\varepsilon_y}{1-\rho_y}\right)v_1 + (v_2'\varepsilon_y)v_2 + \left(\frac{v_3'\varepsilon_y}{1+\rho_y}\right)v_3 \end{aligned} \tag{80}$$

So in this case, Theorem 3 is transparent: as $\rho_y \to 1$ the absolute value of the coefficient on v_1 diverges to infinity almost surely,[47] while all others stay bounded. So it is clear that y is eventually close to this dominant component, which is proportional to v_1. In this $n = 3$ case, we can see this graphically by again using the transformation T in Eq. 52. Since $T1_3 = 0$, we then obtain,

$$Ty = \left(\frac{v_1'\varepsilon_y}{1-\rho_y}\right)Tv_1 + (v_2'\varepsilon_y)Tv_2 + \left(\frac{v_3'\varepsilon_y}{1+\rho_y}\right)Tv_3 \tag{81}$$

and see again that for $\rho_y \approx 1$, the 2-dimensional random vector, Ty, will tend to be approximately proportional to $\pm Tv_1$. This is shown in Fig. 8 for a simulated sample of 1,000 values of Ty with $\rho_y = 0.9$, where the heavy line denotes the span of the maximal eigenvector image, Tv_1.

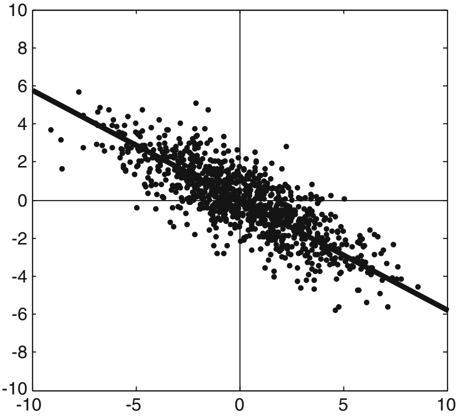
Moreover, by the symmetry of this simple JSE model, precisely the same behavior will be exhibited by Tx for spatial dependency values, $\rho_x \approx 1$. Hence, a comparison of Fig. 8 with panel (b1) of Fig. 4, for example, shows that exactly the same arguments must again lead to spurious correlation between y and x.[48]

However, there is one additional feature of this example that should be noted. While spurious correlation is evident for values of ρ_y and ρ_x very close to one (say

[47] Here, the exceptional subset of $\varepsilon_y$values are those with $v_1'\varepsilon_y = 0$, which has probability zero.

[48] It is also of interest to note that as a *spatial pattern*, the vector v_1 exhibits *maximal correlation* with its corresponding "spatial influence" pattern, Wv_1. In fact, this correlation is *perfect*, since by definition, $Wv_1 = \lambda_1 v_1$ (with $v_1 > 0$) implies $\mathrm{corr}(v_1,\ Wv_1) = 1$ (which is also closely related to the well-known extremal properties of Moran's I in terms of eigenvectors, as for example in DeJong et al. 1984 and Griffith 1996). So this type of spuriousness is again associated with an extreme form of spatial correlation.

Fig. 8 Tx for W_3 with $\rho_y = 0.9$



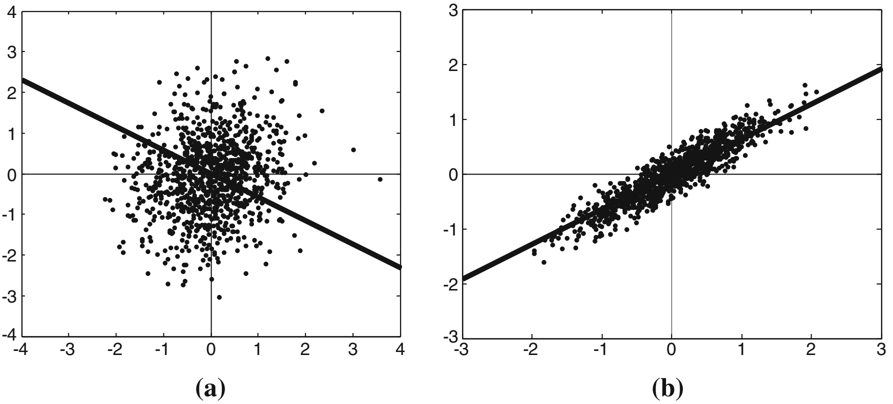


Fig. 9 Comparative effect of proximity to the unit vector. **a** Tx for W_3 with $\rho_x = 0.5$. **b** Tx for W_0 with $\rho_x = 0.5$

both above 0.9 as in Fig. 8), it disappears rapidly even for moderate levels of spatial dependence. This is seen for example in panel (a) of Fig. 9, where a simulated sample of 1,000 values of Tx with $\rho_x \approx 0.5$ produces an almost spherical pattern. So even if y is highly autocorrelated ($\rho_y \approx 1$), it follows from the arguments leading to Theorem 1 that OLS can be expected to do quite well in this case. Hence, it is of interest to ask why sphericity appears to hold even when x exhibits substantial spatial autocorrelation.

The key property exhibited by this particular case is that the maximal eigenvector, $v_1 = (1/2, \sqrt{2}/2, 1/2)' = (1/2)(1, \sqrt{2}, 1)'$, is almost proportional to the unit vector, 1_3, which in geometric terms, means that v_1 is close to span(1_3). Hence, the projection, Tv_1, must be close to the origin, which in turn dampens the directional tendencies of the random vector, Tx, even for large values of ρ_x. In fact, the situation here is seen even more clearly by examining the original hypothesized model Eq. 1 in terms of y. For if v_1 is proportionally close to 1_n, and y is proportionally close to v_1 (as in Theorem 3), then it can be expected that y is also be proportionally close to 1_n. But by definition, this implies that the term, $\beta_0 1_n$, in Eq. 1 must already account for most of y, and hence that there is little left for x_1 to "explain". Thus, when v_1 is proportionally close to 1_n, as in the present example, spurious correlation should be less of problem for simple OLS regression.[49]

In fact, this observation extends to the multivariate model in (Eqs. 10, 11) as well. For even if X_2 does not include the intercept, 1_n, exactly the same argument shows that when v_1 is close to span(X_2), the random vector y will tend to be well approximated by some linear combination of the columns of X_2, so that the regression term $X_2\beta_2$ now accounts for y. Thus, over-rejection of the hypothesis, $\beta_1 = 0$, will again be less of a problem for OLS.[50]

However, if v_1 is *not* proportionally close to X_2, then spurious correlation continues to be a problem for even moderate levels of spatial dependency. This can be illustrated in the present context by modifying the weight matrix, W_3, to produce a weight matrix with maximal eigenvector not proportionally close to 1_3. In particular, if we now let W_0 be given by

$$W_0 = \lambda_1^{-1} \begin{pmatrix} 0 & 1.0 & 0 \\ 0.9 & 0 & 0.1 \\ 0 & 0.2 & 0 \end{pmatrix} \tag{82}$$

(with unscaled maximum eigenvalue, $\lambda_1 = 0.95917$), then W_0 is seen to be a perturbation of W_3 with the same zeros, but with asymmetric positive weights. Here, the eigenvalues are exactly the same as those of W_3, but the maximum eigenvector is now given by $v_0 = (0.71429, 0.68512, 0.14286)'$. Since the last component is seen to be much smaller than the first two, it follows that v_0 is further from span(1_3) than is v_1. As a comparison with W_3, a simulated sample of 1,000 values of Tx using W_0 with $\rho_x = 0.5$ is shown in panel (b) of Fig. 9. Hence, for this weight matrix, it is clear (from the discussion of Fig. 8 above) that even moderate values of both ρ_y and ρ_x will tend to produce spurious correlation between y and x.

5 Concluding remarks

The main objective of this paper has been to illuminate the nature of spurious correlation in OLS from a geometric viewpoint. In particular, we have shown that such effects arise when both the dependent and explanatory variables, y and x,

[49] In this context, it is important to note that the "row normalization" convention often used with spatial weight matrices in fact *guarantees* that $v_1 = 1_n$. But as pointed out by Kelejian and Prucha (2010), the validity of this normalization procedure is subject to question. Here, it should also be noted that for matrices close to the "equal weights" matrix with constant off-diagonal components, one must again have v_1 close to *span*(1_n). Such weight matrices are known to exhibit a variety of special properties with respect to standard testing procedures, as studied for example by Kelejian and Prucha (2002) and Martellosio (2011).

[50] This observation is closely related to the more general result of Krämer and Donninger (1987) showing that OLS will be as efficient as SEM when $v_1 \in$ span(X) and W is symmetric (see also Tilke 1993, and Krämer and Baltagi 1996). However, this efficiency result holds much more generally, as recently shown by Martellosio (2011).

exhibit similar types of spatial autocorrelation. More generally, the same spurious correlation effects are shown to arise whenever the dispersion ellipsoids of y and x are both nonspherical and approximately aligned.

Hence, the main question left unaddressed by this paper is how to mitigate such spurious effects in statistical inference. Of course SEM itself is designed to account for spatial autocorrelation effects and, as illustrated in the Philadelphia example in Table 1f above, is quite effective when sample sizes are sufficiently large. However, the Columbus example in Table 1d shows that for smaller samples, even SEM exhibits spurious correlation effects when spatial autocorrelation in the explanatory variable is substantial. The basic reason of course is that SEM is formulated in a conditional setting where spatial autocorrelation in y is modeled, but *not* spatial autocorrelation in x. Hence, there remains the question of how to improve small-sample inference for SEM in the presence of spatial autocorrelated explanatory variables.

As mentioned in the introduction, there have been a number of efforts to model spatial autocorrelation in both y and x and to construct improved inference procedures on this basis. These efforts have for the most part focused on direct tests of correlation between spatially autocorrelated processes, y and x, as originally studied by Bivand (1980) using simulation. From an analytical perspective, perhaps the best results to date are those of Dutilleul (1993), who proposed a method for modifying the degrees of freedom ("effective sample size") of the standard t test for correlation between y and x that compensates for such effects. While this method has also been applied by Dutilleul and Alpargu (2001) to regression with temporal autocorrelation in both y and x (by appealing to the rough analogy between correlation and linear regression), this method does not appear to be directly applicable to SEM. However, by employing the general strategy of modifying t tests in terms of their degrees of freedom, it does appear that improved inference procedures can be developed for SEM in the small-sample case. These results will be reported in a subsequent paper (Smith and Lee 2011b).

Acknowledgments The authors are indebted to Federico Martellosio for valuable comments and suggestions on an earlier draft of this paper. We are also grateful to the two referees for their constructive comments.

References

Alpargu G, Dutilleul P (2003a) To be or not to be valid in testing the significance of the slope in simple quantitative linear models with autocorrelated errors. J Stat Comput Simul 73(3):165–180

Alpargu G, Dutilleul P (2003b) Efficiency and validity analyses of two-stage estimation procedures and derived testing procedures in quantitative linear models with AR(1) errors. Commun Stat Simul Comput 32(3):799–833

Alpargu G, Dutilleul P (2006) Stepwise regression in mixed quantitative linear models with autocorrelated errors. Commun Stat Simul Comput 32:799–833

Anselin L (1988) Spatial econometrics: methods and models. Kluwer, Boston

Berman A, Plemmons RJ (1994) Nonnegative matrices in the mathematical sciences. Siam, Philadelphia

Bivand R (1980) A Monte Carlo study of correlation coefficient estimation with spatially autocorrelated observations. Quaest Geogr 6:5–10

Clifford P, Richardson S, Hémon D (1989) Assessing the significance of the correlation between two spatial processes. Biometrics 45(1):123–134

Davidson R, MacKinnon J (1993) Estimation and inference in econometrics. Oxford University Press, New York

Davidson R, MacKinnon J (2004) Econometric theory and methods, Oxford University Press, New York

DeJong P, Sprenger C, Van Veen F (1984) On extreme values of Moran's I and Geary's C. Geogr Anal 16(1):17–24

Dutilleul P (1993) Modifying the *t* test for assessing the correlation between two spatial processes. Biometrics 49(1):305–314

Dutilleul P (2008) A note on sufficient conditions for valid unmodified t testing in correlation analysis with autocorrelated and heteroscedastic sample data. Commun Stat Theory Method 37:137–145

Dutilleul P, Alpargu G (2001) Efficiency analysis of ten estimation procedures for quantitative linear models with autocorrelated errors. J Stat Comput Simul 69:257–275

Fingleton B (1999) Spurious spatial regression: some Monte Carlo results with a spatial unit root and spatial cointegration. J Reg Sci 39(1):1–19

Green WH (2003) Econometric analysis, 5th edn. New Jersey, Prentice Hall

Griffith D (1996) Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying geo-referenced data. Can Geogr 40(4):351–357

Huynh H, Feldt S (1970) Conditions under which mean square ratios in repeated measurements designs have exact F-distributions. J Am Stat Assoc 65(332):1582–1589

Kelejian HH, Prucha IR (2002) 2SLS and OLS in a spatial autoregressive model with equal spatial weights. Reg Sci Urban Econ 32(6):691–707

Kelejian HH, Prucha IR (2010) Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. J Econ 157(1):53–67

Kramer W (2003) The robustness of the *F*-test to spatial autocorrelation among Regression disturbances. Statistica 63(3):435–440

Krämer W, Baltagi B (1996) A general condition for an optimal limiting efficiency of OLS in the general linear regression model. Econ Lett 50(1):13–17

Krämer W, Donninger C (1987) Spatial autocorrelation among errors and the relative efficiency of OLS in the linear regression model. J Am Stat Assoc 82(398):577–579

Lauridsen J, Kosfeld R (2006) A test strategy for spurious spatial regression, spatial nonstationarity, and spatial cointegration. Pap Reg Sci 85(3):363–377

Legendre P, Dale MRT, Fortin M-J, Gurevitch J, Hohn M, Myers D (2002) The consequences of spatial structure for the design and analysis of ecological field surveys. Ecography 25(5):601–615

Martellosio F (2010) Power properties of invariant tests for spatial autocorrelation in linear regression. Econ Theory 26(1):152–186

Martellosio F (2011) Non-testability of equal weights spatial dependence. Econ Theory. doi: 10.1017/S0266466611000089

Mur J, Trívez FJ (2003) Unit roots and deterministic trends in spatial econometric models. Int Reg Sci Rev 26(3):289–312

Smith TE, Lee KL (2011a) The effects of spatial autoregressive dependencies on inference in OLS: A geometric approach. Working Paper available on line at: http://www.seas.upenn.edu/~tesmith/Geometry_of_Spurious_Correlation.pdf, in progress

Smith TE, Lee KL (2011b) Small sample inference for spatial error models, in progress

Tilke C (1993) The relative efficiency of OLS in the linear regression model with spatially autocorrelated errors. Stat Pap 34(3):263–270