

A SPATIO-TEMPORAL MODEL OF HOUSING PRICES BASED ON INDIVIDUAL SALES TRANSACTIONS OVER TIME[†]

Tony E. Smith^{*a} and Peggy Wu^b

June 18, 2008
(Revised January, 2009)

Abstract

A spatio-temporal model of housing price trends is developed that focuses on individual housing sales over time. The model allows for both the spatio-temporal lag effects of previous sales in the vicinity of each housing sale, and for general autocorrelation effects over time. A key feature of this model is the recognition of the unequal spacing between individual housing sales over time. Hence the residuals are modeled as a first-order autoregressive process with unequally spaced events. The maximum-likelihood estimation of this model is developed in detail, and tested in terms of simulations based on selected data. In addition, the model is applied to a small data set in the Philadelphia area.

Key Words: Housing Prices, Spatio-temporal models, Autocorrelation

[†] We owe special thanks to Gloria Guard (President) and Stephanie Wall (Outcomes Manager) at the Peoples' Emergency Center in Philadelphia for providing us with their data. We are also indebted to the referees of an earlier draft of this paper for many helpful comments.

^{*} Corresponding Author. Tel. : +1-215-898-9647, Fax: +1-215-573-2068,
E-mail Address: tesmith@seas.upenn.edu.

^a Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia., PA 19104

^b Department of Policy and Information Services, The Reinvestment Fund, Philadelphia, PA 19106.

1. Introduction

The present model grew out of an effort to identify the impacts of certain Community Development Corporation (CDC) housing projects on their local housing markets in Philadelphia. While no single measure can effectively capture neighborhood revitalization, it is becoming common practice to use changes in housing prices as a summary measure.¹ Hence this model focuses on the specific problem of identifying trends in housing prices within a given region based on a time series of individual housing sales transactions. Because our approach to this time-series problem appears to be new to the housing literature, the objective of the present paper is to present a self-contained development of this approach.² The resulting model is illustrated by a small example from the Philadelphia study. A presentation of the full study will be given in a subsequent paper [see Wu and Smith (2009)].

Since housing prices are well known to be influenced by the prices of recent house sales nearby, one must allow for possible spatio-temporal dependencies between such prices. In addition there are generally a host of other processes occurring over time that result in unobserved temporal autocorrelations among housing prices. But since individual housing sales do not occur at regular time intervals, it is difficult to model such processes in terms of standard discrete time series. An approach to unequally spaced temporal events has been developed for first-order autoregressive [AR(1)] processes by a number of authors [see Wansbeek and Kapteyn, (1985), McKenzie and Kapuscinski (1997), Batalgi and Wu (1999)]. A continuous version of these models [CAR(1)] has also been developed by Jones and Boadi-Boateng (1991) [and Jones (1993)]. The continuous version is certainly the most flexible one, but involves stochastic differential equations that rely on rather sophisticated analytical methods. Moreover, since housing sales are basically recorded on a daily basis, there is no need to consider finer time intervals. Hence we choose to develop a standard AR(1) model on daily time intervals, and then embed the observed sales events within this process. The approach we adopt is most closely related to Batalgi and Wu (1999). But since their formulation is in terms of panel data, it is convenient to give a self-contained development for the present case. Finally, an interesting alternative approach to modeling sales transactions with unequal time intervals was proposed by Pace et al. (2000). Because of its close similarity to the present paper, a detailed comparison of these two models is presented in Section 5 below.

We begin in the next section by developing the basic model, and then consider maximum likelihood estimation and testing of its parameters. This is followed by a small simulation study to examine the properties of the estimation procedure. The model is then applied to a selected example from the Philadelphia CDC study, and is compared with the approach

¹ The number of affordable housing units provided (built or renovated) by the CDC is also commonly used as a measure of “success”. However, it has been argued by many housing researchers that increased housing supply is only a measure of *input*, and thus is not a fair assessment of neighborhood revitalization as an *outcome* [see for example Smith (2003) and the many studies cited therein]. Hence the assumption implicit in the present approach is that improved neighborhood quality should increase local demand for housing, and thus local housing prices.

² Certain technical appendices have been omitted to save space, and can be found in the *Electronic Supplementary Material* online.

of Pace et al. (2000). Finally, a possible extension is considered in the concluding section of the paper.

2. Development of the Model

Consider a sequence of *sales prices*, $(y_i : i = 1, \dots, n)$, resulting from the sale of individual houses at distinct *time points*, $(t_i : i = 1, \dots, n)$ in a given metropolitan area.³ Such prices of course depend on a host of *attributes*, $x_i = (x_{i1}, \dots, x_{ik})$, of each house i , as well as the sales prices of houses recently sold in the immediate area of house i . Here we model such dependencies by *spatio-temporal lag weights*, w_{ij} , that are assumed to be positive only if house j is sold prior to house i ,⁴ and is “sufficiently close” to i in both time and space to be an influential factor. More precisely, it is here assumed that there is some *threshold time interval*, Δ , and *threshold distance*, d , beyond which other housing sales have no direct influence on the price of house i . Hence, if the relevant distance from i to j is denoted by d_{ij} , then is assumed that only houses in the *spatio-temporal neighborhood*,

$$N_i = \{j : (0 < t_i - t_j \leq \Delta) \& (d_{ij} \leq d)\} \quad (1)$$

of i have direct influence on the sales price of house i . If n_i denotes the cardinality (size) of N_i , then the corresponding lag weights, w_{ij} , are given by

$$w_{ij} = \begin{cases} 1/n_i & , j \in N_i \\ 0 & , j \notin N_i \end{cases} \quad (2)$$

While this simple spatial-threshold assumption plays no substantive role in the analysis to follow, it is used in both the simulations and empirical application below.⁵ With these conventions, our basic model of housing prices takes the following *spatio-temporal lag form*

$$y_i = \lambda \sum_{j=1}^n w_{ij} y_j + \beta_0 + \sum_{h=1}^k \beta_h x_{ih} + u_i \quad , \quad i = 1, \dots, n \quad (3)$$

In the first term, λ is an *intensity parameter* reflecting the strength of price dependencies. Note from the normalization assumption, $w_{ij} = 1/n_i$, in (2) that this term is

³ Alternatively, it may often be more appropriate to use the log of sales price as y_i .

⁴ Note that there may in fact be some minimal time lag required before a given sales price can influence subsequent prices (such as the time required for this sale to be published in the local paper). Hence the inclusion of all prior sales is a simplifying assumption.

⁵ It is worth noting that even if more elaborate spatial kernel functions were to be used, the *bandwidth*, d , of each kernel is well known to be the single most critical determinant of spatial dependence (see for example Silverman, 1986).

simply the average housing price in N_i weighted by λ .⁶ To ensure that variances in housing prices remain bounded over time, it is required that there be a diminished dependency of house prices i on these averages, i.e., that (Green, 2003, p.255):

$$|\lambda| < 1 \quad (4)$$

The second term in (3) involves the usual linearity assumption on housing attributes, where β_h is the relevant coefficient for attribute h . (Of particular relevance for our present purposes is the inclusion of time as an attribute of each sale, in order to capture price trend effects.) Finally, the *residuals*, $(u_i : i = 1, \dots, n)$, in (3) are assumed to be generated by an underlying autoregressive process which we now develop.

2.1 Autoregressive Process with Unequally Spaced Events

Consider a discrete process, $\{u(t) : t \in T\}$, over a sequence of consecutive days, $T = \{t_1, t_1 + 1, \dots, t_1 + k, \dots\}$, which is generated by the following AR(1) process, with *autocorrelation parameter*, ρ ,

$$u(t) = \rho \cdot u(t-1) + \varepsilon(t) \quad , \quad t > t_1 \quad (5)$$

where in addition it is assumed that $u(t_1) \sim N(0, v)$ and that all *innovations*, $\{\varepsilon(t) : t > t_1\}$, are independently and identically distributed normal variates, $\varepsilon(t) \sim N(0, \sigma^2)$. This process is also assumed to be *stationary*, so that $u(t) \sim N(0, v)$ holds for all $t \in T$. Since each $u(t)$ is necessarily normal with zero mean by (5), this is equivalent to requiring that variance stay constant, i.e., that $\text{var}[u(t)] \equiv v$. Moreover, since $u(t-1)$ and $\varepsilon(t)$ are independent, this yields the following well known stationarity condition:

$$\text{var}[u(t)] = \rho \cdot \text{var}[u(t-1)] + \text{var}[\varepsilon(t)] \Rightarrow v = \rho^2 v + \sigma^2 \Rightarrow v = \frac{\sigma^2}{1 - \rho^2} \quad (6)$$

In particular, it follows that [as a parallel to (4)] residual variance will be finite only if,

$$|\rho| < 1 \quad (7)$$

Within this standard AR(1) setting, we now assume that the first sales event occurs on day t_1 , so that the first residual in (3) is given by $u_1 = u(t_1)$. Next, if the second sale occurs on day $t_2 = t_1 + m$, then since $u(t_1 + 1) = \rho \cdot u(t_1) + \varepsilon(t_1 + 1)$, and

$$u(t_1 + 2) = \rho \cdot u(t_1 + 1) + \varepsilon(t_1 + 2) = \rho \cdot [\rho \cdot u(t_1) + \varepsilon(t_1 + 1)] + \varepsilon(t_1 + 2) \quad (8)$$

⁶ Note also that if housing prices are in log form, then this term corresponds to a *geometric average* of housing prices rather than an arithmetic average.

$$= \rho^2 \cdot u_1 + [\varepsilon(t_1 + 2) + \rho \cdot \varepsilon(t_1 + 1)] ,$$

it follows by successive substitutions into (5) for $m > 2$ that,

$$u_2 = \rho^m u_1 + \sum_{j=1}^m \rho^{m-j} \varepsilon(t_1 + j) \quad (9)$$

More generally, if we now replace (u_1, u_2, t_1, t_2) by $(u_{i-1}, u_i, t_{i-1}, t_i)$ and replace $m = t_2 - t_1$ by $\Delta_i = t_i - t_{i-1}$, then exactly the same argument shows that for all $i = 2, \dots, n$,

$$u_i = \rho^{\Delta_i} u_{i-1} + e_i \quad (10)$$

where the (*cumulative*) *innovations*, e_i , now have the form:

$$e_i = \sum_{j=1}^{\Delta_i} \rho^{\Delta_i - j} \varepsilon(t_{i-1} + j) \quad (11)$$

Moreover, since $\text{var}(u_i) = v$ for all $i = 1, \dots, n$, it also follows from (10) that, in a manner similar to (6),

$$\text{var}(u_i) = \rho^{2\Delta_i} \text{var}(u_{i-1}) + \text{var}(e_i) \Rightarrow v = \rho^{2\Delta_i} \cdot v + \text{var}(e_i) \quad (12)$$

$$\Rightarrow \text{var}(e_i) = v(1 - \rho^{2\Delta_i})$$

[which can also be verified by summing the variances in (11) and using the identity, $v = \sigma^2 / (1 - \rho^2)$]. Following Jones (1993) it is convenient to rescale the innovations, e_i , as,

$$\varepsilon_i = (1 - \rho^{2\Delta_i})^{-1/2} e_i \quad (13)$$

so that $\text{var}(\varepsilon_i) \equiv \text{var}(u_i) = v$. With this rescaling, it follows that the relevant *residual process* $(u_i : i = 1, \dots, n)$, for (3) can now be summarized as follows:

$$u_1 \sim N(0, v) \quad (14)$$

$$u_i = \rho^{\Delta_i} u_{i-1} + (1 - \rho^{2\Delta_i})^{1/2} \varepsilon_i \quad \text{with} \quad \varepsilon_i \sim N(0, v) , \quad i = 2, \dots, n \quad (15)$$

It is of interest to note that this *residual process*, [(14),(15)], is essentially identical to the continuous formulation in Jones (1993, p.62).⁷ Thus, by allowing time intervals to become arbitrarily small, the above formulation provides a somewhat more intuitive motivation of CAR(1) processes.

However, there is one additional restriction in the present model that is not shared by the continuous model. In particular, observe that with a minimal time unit (such as one day), it is quite possible that more than one event occurs in the same time interval (such as more than one house sold on the same day). It should be clear from (15) that this leads to implausible results, since the residuals u_i for simultaneous events must be *identical*.⁸

There are several ways to treat this problem. First, one can simply choose the time unit to be smaller than the closest pair of consecutive events in the given data. While this is possible in the present setting, intervals smaller than a day are at best artificial, and have little meaningful content. A much more satisfactory approach would be to introduce an additional “idiosyncratic” error term in (3) reflecting the unobserved attributes of individual houses that are *time independent*. This approach is discussed in more detail in the Concluding Remarks. For the present however, we choose to focus on a single set of errors driven by an underlying AR(1) process, and to examine the behavior of this model in detail. But since there are indeed a number of instances of houses sold on the same day in the application presented below, we have chosen to “jitter” the time sequence enough to allow a well-ordered sequence of sales events. Given that this error model [(14),(15)] is intended only to capture unobserved effects with some degree of stationary temporal dependency, such “tie breaking” conventions are deemed to have little affect on the overall behavior of the model.

2.2 Matrix Formulation of the Model

By combining (3), (14), and (15) we can give a more compact statement of the above model as follows. If we let $y = (y_1, \dots, y_n)'$, $W = (w_{ij} : i, j = 1, \dots, n)$, $\beta = (\beta_1, \dots, \beta_k)'$, $u = (u_1, \dots, u_n)'$, and let $X = [x_{ij} : i = 1, \dots, n, j = 0, 1, \dots, k]$ with $x_{i0} = 1, i = 1, \dots, n$, then (3) becomes

$$y = \lambda W y + X \beta + u \quad (16)$$

Similarly if we let $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$, then [(14),(15)] can also be written in matrix form as

$$u = D(\rho)u + C(\rho)\varepsilon, \quad \varepsilon \sim N(0, vI_n) \quad (17)$$

where $D(\rho)$ is a *lower triangular* matrix of the form

⁷ If time intervals, Δ_t , are allowed to become “arbitrarily small”, and are denoted by δt , then in the formulation of Jones (1993), $\phi(\delta t) = \exp(-\alpha_0 \delta t) = \rho^{2\delta t}$ where $\rho = \exp(-\alpha_0)$.

⁸ This does not arise in the continuous model where it is natural to assume that simultaneous events occur with probability zero.

$$D(\rho) = \begin{pmatrix} 0 & & & & \\ \rho^{\Delta_2} & 0 & & & \\ 0 & \rho^{\Delta_3} & 0 & & \\ \vdots & \ddots & \ddots & \ddots & \\ 0 & \dots & 0 & \rho^{\Delta_n} & 0 \end{pmatrix} \quad (18)$$

and $C(\rho)$ is a *diagonal* matrix of the form

$$C(\rho) = \begin{pmatrix} 1 & & & & \\ & (1-\rho^{2\Delta_2})^{1/2} & & & \\ & & \ddots & & \\ & & & & (1-\rho^{2\Delta_n})^{1/2} \end{pmatrix} \quad (19)$$

Hence the present *spatio-temporal model* is now summarized by (16) through (19) [together with the implicit parameter restrictions (4) and (7)].

Note that there is a strong similarity between the present model and the well-known “spatial autoregressive model with autoregressive disturbances” summarized, for example, in Anselin and Florax (1995, pp. 22–24). However, it should be clear that the present autocorrelation parameter, ρ , enters in a more complex manner than the autoregressive disturbance parameter of that model. On the other hand, the present model is in many ways simpler to analyze, since there are *no simultaneities* either in space or time. As a consequence, both the spatial weight matrix, W , and the temporal dependency matrix, $D(\rho)$, are lower triangular matrices, which greatly simplifies the analysis of this model.

2.3 Likelihood and Concentrated Likelihood Functions

To estimate the present model it is convenient to combine (16) and (17) into a reduced model form as follows. First, for notational simplicity, we write $D = D(\rho)$ and $C = C(\rho)$ and solve for u in terms of ε to obtain

$$(I_n - D)u = C\varepsilon \Rightarrow u = (I_n - D)^{-1}C\varepsilon = B\varepsilon \quad (20)$$

where $B = (I_n - D)^{-1}C$.⁹ Next we set $A = I_n - \lambda W$ and write (16) as

$$(I_n - \lambda W)y = X\beta + B\varepsilon \Rightarrow Ay = X\beta + B\varepsilon \quad (21)$$

⁹ It is shown in (A1.2) of Appendix 1 (see footnote 2) that $\det(I_n - D) = 1$, so that $(I_n - D)^{-1}$ always exists and B is well defined.

Hence (20) and (21) are together equivalent to the reduced model¹⁰

$$y = A^{-1}X\beta + A^{-1}B\varepsilon, \quad \varepsilon \sim N(0, vI_n) \quad (22)$$

This in turn implies that y is multivariate normal with distribution

$$y \sim N\left(A^{-1}X\beta, vA^{-1}BB'(A')^{-1}\right) \quad (23)$$

Hence the log likelihood function for parameters $(\beta, v, \lambda, \rho)$ is given by

$$\begin{aligned} L(\beta, v, \lambda, \rho | y) &= -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln\{\det[vA^{-1}BB'(A')^{-1}]\} \\ &\quad - \frac{1}{2v}(y - A^{-1}X\beta)'[A^{-1}BB'(A')^{-1}]^{-1}(y - A^{-1}X\beta) \\ &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(v) - \frac{1}{2}\ln\{\det[A^{-1}BB'(A')^{-1}]\} \\ &\quad - \frac{1}{2v}(Ay - X\beta)'(BB')^{-1}(Ay - X\beta) \end{aligned} \quad (24)$$

where implicitly, $A = A(\lambda)$ and $B = B(\rho)$.

In Appendix 1 (see footnote 2 above) it is shown that the associated *concentrated likelihood function* of ρ is given by:

$$L_c(\rho | y) = -\frac{n}{2}[1 + \ln(2\pi)] - \frac{n}{2}\ln\{\hat{v}[\hat{\lambda}(\rho), \rho]\} - \frac{1}{2}\sum_{i=2}^n \ln(1 - \rho^{2\Delta_i}) \quad (25)$$

where

$$\hat{v} = \hat{v}[\hat{\lambda}(\rho), \rho] = \frac{1}{n}[\hat{A}y - X\hat{\beta}(\rho)]'(BB')^{-1}[\hat{A}y - X\hat{\beta}(\rho)] \quad (26)$$

with $\hat{A} = A[\hat{\lambda}(\rho)]$ and where in addition,

$$\hat{\beta} = \hat{\beta}(\rho) = \hat{\beta}[\hat{\lambda}(\rho), \rho] = [X'(BB')^{-1}X]^{-1}X'(BB')^{-1}\hat{A}y \quad (27)$$

and

$$\hat{\lambda} = \hat{\lambda}(\rho) = \frac{y'G'(BB')^{-1}GWy}{y'W'G'(BB')^{-1}GWy}, \quad (28)$$

¹⁰ It is also shown in (A1.2) of Appendix 1 that $\det(A) = 1$, so that A^{-1} always exists.

with

$$G = G(\rho) = I_n - X[X'(BB')^{-1}X]^{-1}X'(BB')^{-1}. \quad (29)$$

This *one-dimensional* function can in principle be maximized by a simple line search to obtain the maximum-likelihood estimate, $\hat{\rho}$. However, there are several practical considerations that should be mentioned at this point. First, in autocorrelation models where time intervals between events are unequal, the value of ρ is very dependent on the underlying *time unit*. In particular, since (15) shows that the only quantities used in the analysis are the powers, ρ^Δ (and their squares $\rho^{2\Delta}$), it is these values that determine the likelihood function in (25). So if time is rescaled by a factor, α , then the identity, $\rho^\Delta \equiv (\rho^{1/\alpha})^{\alpha\Delta}$, implies that $\hat{\rho} \rightarrow (\hat{\rho})^{1/\alpha}$. Hence even though positive autocorrelation estimates will always lie between zero and one, the actual value of $\hat{\rho}$ depends on the time units and can only be interpreted in this context.

A second related issue concerns the *sign* of ρ . While (7) allows negative as well as positive values of ρ , it should be clear that negative dependencies are somewhat problematic in the present setting. In particular, the type of oscillation behavior implied by negative ρ depends crucially on the choice of time unit which, as mentioned above, can be quite arbitrary in the present setting.¹¹ Moreover, since positive dependencies are of primary interest in the present setting, we simply restrict the relevant interval of ρ values to the interval $[0,1)$, and take all zero values of $\hat{\rho}$ to mean “no autocorrelation”.¹²

A final consideration that arises with unequally spaced events relates to the *degree* of inequality between time intervals. In particular, if the difference between the smallest and largest values of Δ is considerable, then computational overflows can result. For example, if time units are in days (as in the present application) and there is a 30 day lag between two consecutive sales, then standard double-precision computations of the quantity $(1 - \rho^{2\Delta_i})^{1/2} = \sqrt{1 - \rho^{60}}$ will be numerically equal to *one* for all values $\rho < .5$. While most cases are not this extreme, it should be clear that the concentrated likelihood function will tend to be very flat for small values of ρ . The consequences of this flatness are discussed further in Section 3.1 below.

Given an estimate, $\hat{\rho}$, of ρ , one can immediately obtain corresponding estimates $(\hat{\lambda}, \hat{\beta}, \hat{v})$ by successive substitution into (29), (28), (27) and (26). Hence whenever $\hat{\rho}$ is unique, it follows that all estimates will be unique. Here it should be noted that while the

¹¹ Indeed, such oscillation behavior loses all meaning in the *continuous* version of AR(1), where the autocorrelation parameter is *required* to be nonnegative [as is evident from the positivity restriction, $\alpha_0 > 0$, in Jones (1993) for the identity, $\rho = \exp(-\alpha_0)$, of footnote 7 above].

¹² Following Anselin and Moreno (2003), one might also interpret such negative estimates of ρ to be evidence that the present spatio-temporal specification is simply not supported by the data.

concentrated likelihood function, $L_c(\rho | y)$, can fail to be concave (as can be shown by numerical examples), the maximum-likelihood estimates, $\hat{\rho}$, have proved to be unique in all simulated cases studied to date. Thus it appears that non-uniqueness of parameter estimates is not a serious issue in the present model.

Finally, this set of estimates $(\hat{\beta}, \hat{\nu}, \hat{\lambda}, \hat{\rho})$ can be used to estimate asymptotic variances for testing purposes. A complete derivation of the asymptotic covariance matrix is given in Appendix 2 (see footnote 2 above).

3. Selected Simulation Results

Two types of simulation analyses were done for this model. The first involved an entirely artificial space-time process constructed on a square grid of housing sites. The purpose of these simulations was to examine the effect of sample size on the reliability of parameter estimates. As will be shown below, maximum-likelihood estimates for this model are quite sensitive to sample size, and can be extremely inefficient for small samples. In view of this, it was crucial to determine whether such effects were present in the CDC application developed below. So a second set of simulations was run for that model, using the empirical space-time structure and sample size, together with the parameter estimates obtained from the given data set. The results of these simulations, presented in Section 4.3 below, confirmed that the sample size of this application (400 sales) was sufficiently large to achieve reasonably efficient estimates. Moreover, since data sets for housing sales are typically much larger than this, the present results suggest that maximum-likelihood methods should produce reliable results in most applications. But given the wider range of potential applications of this model, it is of interest to consider its small sample behavior.¹³

3.1 A 100-Sample Case

Here “small sample” behavior is well illustrated by sample sizes as large as 100. To construct such a case, a population of 100 houses was placed on a 10 by 10 unit grid of locations, and these houses were sold in random order with time intervals sampled from an exponential (“memoryless”) distribution with a mean of 4 days. Time intervals were then rounded upward so that the minimal time interval was one day. To construct space time dependencies as in (1), a threshold time interval of $\Delta = 60$ days was chosen, along with a threshold distance of three units, $d = 3$. This produced a W matrix of space-time influences in which 65 % of the houses were influenced by the previous sale of at least one house nearby.¹⁴ Two housing attribute values (x_1, x_2) were randomly sampled from a uniform distribution on the unit interval, producing a 100×2 matrix, X , of housing data.

¹³ Here it should be emphasized that the following simulations are intended only to illustrate the “typical” small sample properties of this model in a single situation. Systematic simulation studies of model performance under a range of space-time structures and parameter values are left for future work.

¹⁴ Here parameter choices (d, Δ) for W were chosen to yield a degree of space-time interaction among housing sales that roughly matched that of the Philadelphia application below, where about 66% of the houses were influenced by previous sales (based on the space-time bandwidths used).

The parameter values chosen for this simulation were $\beta = (5, 1, 2)'$, $v = 4$, $\lambda = .4$, and $\rho = .2$. The variance, v , was chosen to be relatively large in comparison to the conditional mean. This was done to maintain some degree of comparability with the application below which also involves two explanatory variables, and thus resulted in a substantial unobserved variation. In addition, the value of ρ was chosen to be low enough to allow substantial unobserved variation between sales occurring only one day apart (as discussed at the end of Section 2.1 above).

In this setting, a set of 1000 sales price vectors, $y = (y_i : i = 1, \dots, 1000)$, were simulated. The resulting distributions of parameter estimates are summarized in Table 1 below.

	β_0	β_1	β_2	λ	ρ	v
Mean	5.032	0.991	1.963	0.397	0.206	3.817
Std Dev	.586	0.646	0.693	0.045	0.165	0.562

Table 1: Estimates for the Simulation of 100 Sales

Here the means of all parameters are close to their true values, but the standard deviations on most parameters are quite high. Recall that the high variance, $v = 4$, is a contributing factor here, especially for the β parameters. This can be seen in Figure 1(a) below, where even then most inefficient estimator, $\hat{\beta}_1$, is symmetrically distributed about its true value, $\beta_1 = 1$ (shown by the arrow head below the figure). The situation for $\hat{\lambda}$ is surprisingly much better than for the betas, and is seen in Figure 1(b) to be much more

Figure 1 here

concentrated around its true value $\lambda = .4$. But the situation for $\hat{\rho}$ is quite different, as shown in Figure 2(a) below.

Figure 2 here

Even though the sample mean of $\hat{\rho}$ seems reasonable, the actual sampling distribution exhibits extreme variation. Notice also that this variation combined with the nonnegativity restriction on ρ (discussed at the end of Section 2.3 above) produces a large spike of zero values. Hence it is clear that without this restriction there would also be a severe negative bias in the estimates of ρ . Even with the restriction, there is a slight downward bias in $\hat{\rho}$, which is directly inherited by \hat{v} as seen in Table 1 and Figure 2(b). This underestimation of variance is of particular significance for statistical inference, since the primary objective of this model is to account for the unobserved variation that is masked by space-time dependencies.

The main reason for the extreme variation in $\hat{\rho}$ values turns out to be the relative flatness of the concentrated likelihood function for ρ , as seen in Figure 3 below, where a case with $\hat{\rho}$ close to the mean [in (a)],

Figure 3 here

and a case with $\hat{\rho}$ truncated to zero [in (b)] are seen to have similar concentrated likelihood functions, both very flat in the range of low ρ values. This can in turn be attributed to the fact that for all time intervals, Δ , that are well above one (say at least five days), the values of ρ^Δ and $\sqrt{1-\rho^{2\Delta}}$ are approximately zero and one, respectively, for all low values of ρ . Hence these terms are essentially *constant* for small ρ . This flatness of course also results in small values of the Hessian for the concentrated likelihood function, and hence tends to inflate the asymptotic variance of $\hat{\rho}$ estimates. (Indeed, for the 1000 simulations above, only about 25% resulted in a significant value of ρ .)

Finally, since all other estimates are constructed from estimates of ρ , it should be clear that this instability in ρ estimates will be inherited by all other estimates.¹⁵ Thus it would appear that for samples this small, autocorrelated errors with unequal spacing can create substantial problems for parameter estimation.

3.2 A 400-Sample Case

Given these small-sample problems, it is of interest to extend the above example to a grid large enough to allow a sample size comparable to the sample size, $n = 400$, in the CDC application below. Hence a 20 by 20 grid was used to obtain 400 housing sites. These were again sold in random order, with lag times defined by rounded samples from an exponential distribution with mean equal to 4 days. A 400×2 matrix, X , of housing attributes was again sampled from the uniform distribution, and the same parameter values for $[\beta, v, \lambda, \rho]$ were used. The results are shown in Table 2 below.

	β_0	β_1	β_2	λ	ρ	v
Mean	5.012	1.001	1.985	0.401	0.196	3.963
Std Dev	.274	0.339	0.343	0.034	0.094	0.281

Table 2. Estimates for the Simulation of 400 Sales

¹⁵ In view of this, it is somewhat surprising that estimates of λ appear to behave quite well by comparison. Moreover, since this relation persists in all simulations studied thus far, it raises an interesting (open) question as to why the expression for $\hat{\lambda}(\hat{\rho})$ in (28) above should remain more stable than $\hat{\rho}$.

Here we see that there is considerable improvement with respect to all estimates. Not only are the means slightly better, but the sample standard deviations are roughly cut in half. The reason is made clear by the histogram for $\hat{\rho}$ in Figure 4(a) below, which exhibits much better statistical properties.

Figure 4 here

While a spike at zero is still evident, there is now a clear concentration around the true value, $\rho = .2$, with only a slight downward bias and a much smaller standard deviation. This reduction is again inherited by all parameter estimates constructed from $\hat{\rho}$. Of particular importance for statistical inference is the improvement in \hat{v} , as seen in Table 2 and Figure 4(b). Not only is the standard deviation cut in half, but also the downward bias has almost disappeared. Hence it appears that for sample sizes of 400 and larger, the maximum-likelihood estimates of all parameters are quite reliable. Additional confirmation of this will be given in Section 4.3 below.

4. Application to a Community Development Corporation Area in Philadelphia

The following application is taken from the Philadelphia study mentioned in the Introduction, and focuses on one of the CDCs in this larger study [which will be reported in a subsequent paper (Wu and Smith, 2009)]. The present CDC, designated as the *Peoples' Emergency Center* (PEC), consists of nine contiguous block groups located in West Philadelphia, as shown in Figure 5 below.

Figure 5 here

The specific objective of this study was to determine whether the overall trend of housing sales prices in each CDC area was significantly greater than that of comparable non-CDC areas in Philadelphia.¹⁶ Here a *control area* (CA), consisting of twenty five contiguous block groups in North Philadelphia was identified (also shown Figure 5) and used for purposes of comparison. The data for the present illustration consists of all housing sales in these two areas during the 20-month period from January 2004 to September 2005. During this period there were 64 housing sales in PEC, and 336 housing sales in the larger control area, CA.¹⁷

To compare sales trends, an instance of model (3) was constructed with y_i denoting the sales price *per square foot* of each house i (to control for variations in house sizes). To account for space-time dependencies among housing sales, the time and distance

¹⁶ During this time period there was a significant increase in housing prices throughout the entire Philadelphia area.

¹⁷ This address-level sales data was extracted from the Philadelphia Board of Revisions of Taxes (BRT) Properties File from 1990 to 2006.

thresholds in (1) above were chosen to be $\Delta = 60$ and $d = 500$, so that for any two houses, i and j , the corresponding space-time weight, w_{ij} , is nonzero if only if house j is within 500 feet of house i and is sold no more than 60 days prior to the sale of house i .¹⁸ Sales in each area were distinguished by a location dummy, δ_i ($= 1$ for PEC houses i), and the *sales time*, t_i , of each house i was used to capture (linear) sales-price trends. In this context, the relevant instance of model (3) for the present application takes the form:

$$y_i = \lambda \sum_{j=1}^n w_{ij} y_j + (\beta_0 + \beta_1 t_i + \beta_2 \delta_i + \beta_3 t_i \delta_i) + u_i, \quad i = 1, \dots, n \quad (30)$$

where $n = 400$ and where $t_i \delta_i$ denotes the *interaction effect* between time and location. For our present purposes, the coefficient of this interaction effect (β_3) is the key parameter of interest. This can be seen more clearly by rewriting (30) as follows:

$$y_i = \lambda \sum_{j=1}^n w_{ij} y_j + \beta_0 + \beta_2 \delta_i + (\beta_1 + \beta_3 \delta_i) t_i + u_i, \quad i = 1, \dots, n \quad (31)$$

Here it is clear that β_3 represents the difference in slopes between the linear sales-price trends in the PEC and CA areas. Hence a positive value of β_3 would at least be consistent with a positive local effect of the PEC housing projects.

4.1 OLS Estimation

To analyze this effect we start by treating (30) as a standard linear model. Here it is of interest to observe that if temporal autocorrelation effects are assumed to be absent, i.e., if $\rho = 0$, then (30) is *precisely* a standard linear model. To see this, consider the form of the full likelihood function in (24) and note first that if $\rho = 0$, then by definition, $C = B = I_n$, and in particular $\ln[\det(C)] = 0$. Hence if we let

$$\tilde{\beta} = \begin{pmatrix} \lambda \\ \beta \end{pmatrix}, \quad \tilde{X} = (Wy, X) \quad (32)$$

so that $Ay - X\beta = y - \lambda Wy - X\beta = y - \tilde{X}\tilde{\beta}$, and rewrite the reduced likelihood with $\rho = 0$ as $L(\beta, v, \lambda) \equiv L(\tilde{\beta}, v)$, then it follows at once from (24) that

$$L(\tilde{\beta}, v) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(v) - \frac{1}{2v} (y - \tilde{X}\tilde{\beta})'(y - \tilde{X}\tilde{\beta}) \quad (33)$$

which is precisely the standard linear model likelihood function for $(\tilde{\beta}, v)$.¹⁹

¹⁸ It should be clear from Figure 5 that the areas PEC and CA are sufficiently far apart to ensure that no space-time dependencies occur between houses in separate areas.

¹⁹ It should be noted here that since this model involves both lagged dependent variables and a time trend term, it is technically an instance of a ‘‘autoregressive process around a deterministic time trend’’. But while

Hence the ordinary least squares (OLS) estimates of $(\tilde{\beta}, v) = (\lambda, \beta, v)$ for this case are guaranteed to satisfy all the usual optimality properties of maximum-likelihood estimates.²⁰ More generally, if temporal autocorrelation effects are present but not too severe (as will be seen to be the case in this application), then OLS should continue to yield quite reasonable estimates. The results for OLS in the present case are summarized in Table 3 below:

Variable	Coefficient	<i>t</i> -value	<i>p</i> -value
Constant (β_0)	10.203	4.482	0.00001
Spatial Lag (λ)	0.123	2.314	0.021
Time (β_1)	0.014	2.419	0.016
PEC Indicator (β_2)	13.160	2.461	0.014
Time-PEC Interaction (β_3)	0.053	3.774	0.0002
Pseudo R-Squared	0.362		
AIC	3435.219		
Variance (<i>v</i>)	308.651		

Table 3. OLS Results for PEC

Here both the time parameter (β_1) and the key time-location interaction parameter (β_3) are significantly positive – with interaction being *very* significant. Moreover the relative size of these coefficients shows that price increases in PEC are considerably higher than those of CA, with a rate of increase above that of CA by more than 5 cents per square foot per day. Notice also that the spatial effect of nearby previous sales (λ) is significantly positive.

In the next section it is shown that our present spatio-temporal extension of OLS does *not* change these conclusions in any substantial way. Moreover, while the present analysis is based only on a small data set drawn from the larger Philadelphia CDC study, these same

the rates of convergence for OLS estimates are more delicate in this case, it can be shown that the standard model significance tests continue to be asymptotically valid. [See for example Hamilton (1994, Section 16.3)].

²⁰ In particular, OLS estimates in the presence of lagged dependent variables are consistent and asymptotically normally distributed about their true values. However, since the lagged dependent variables and residuals are not fully independent, these estimates are typically biased for small samples. [See for example Davidson and MacKinnon (2004, Section 3.2).]

conclusions follows from the larger study as well. Hence it does appear that prices have been rising faster in the CDC areas than can be accounted for by general housing price increases during this period. In so far as increased housing values can be taken to reflect neighborhood revitalization, our results thus suggest that these Philadelphia CDC projects have achieved some degree of success. For a fuller discussion of these points, see Wu and Smith (2009).

However, there are a number of shortcomings of the present analysis that carry over to the larger study as well. The most obvious is the absence of additional housing attributes that should be controlled for (as reflected by the low value of pseudo R-square²¹). But unfortunately, such attributes were not available in usable form for the current data set.²² The other key question for our present purposes relates to possible temporal autocorrelation among residuals that is not captured by the simple time trend in (30). As is well known, such autocorrelation tends to smooth residuals which can in principle inflate the statistical significance of key parameters. Indeed, this was the primary motivation for the present spatio-temporal model.

4.2 Spatio-Temporal Estimation

Hence we now re-examine this data in terms of present spatio-temporal model. The results of this estimation are summarized in Table 4 below. Notice first that the key parameter estimates and significance levels are strikingly similar to Table 3.

Variable	Coefficient	<i>t</i> -value	<i>p</i> -value
Constant (β_0)	10.203	4.482	0.00001
Spatial Lag (λ)	0.123	2.314	0.021
Time (β_1)	0.014	2.419	0.016
PEC Indicator (β_2)	13.160	2.461	0.014
Time-PEC Interaction (β_3)	0.053	3.774	0.0002

²¹ The standard R-square is known to be somewhat problematic in the case of lagged dependent variables. Hence for comparability with the spatio-temporal formulation below, we choose to define *pseudo R-square* here to be the squared correlation between y and its prediction, $\hat{y} = \hat{E}(y | X) = (I_n - \hat{\lambda}W)^{-1} X \hat{\beta}$ with OLS estimates $\hat{\lambda}$ and $\hat{\beta}$. However, it is also of interest to note that in this particular application the unadjusted R-square (0.372) is almost the same as the pseudo R-square (0.362).

²² Here it should be noted that CDCs are local non-profit organizations whose funding is devoted almost entirely to housing projects, and not to data collection. Hence all housing data was drawn from the Philadelphia BRT (footnote 17 above). Moreover, while this BRT data did include provisions for a number of key housing attributes (such as “number of bedrooms” and “interior and exterior condition”), most of this data was either missing or unusable for other reasons.

Pseudo R-Squared	0.362
AIC	3435.219
Variance (v)	308.651

Table 4. Spatio-Temporal Results for PEC

Moreover, as is seen from the pseudo R-square and AIC values, the overall goodness of fit for both models is also very similar.²³ Hence, in view of the discussion in Section 4.1 above, these similarities would seem to suggest that temporal autocorrelation is not present. However, the p-value (0.044) for temporal autocorrelation, ρ , does indicate some statistical significance here. But as noted in Section 2.3 above, the estimated *value* of ρ can only be interpreted relative to the *time units* used in the analysis. In the present case, $\hat{\rho} = .112$, with a time unit of one day. This implies from (10) that for consecutive housing sales, $i-1$ and i , that are more than one day apart, the influence of residual, u_{i-1} , on u_i is less than $\rho^2 u_{i-1} \approx (.013)u_{i-1}$. So in spite of its apparent statistical significance, the *autocorrelation impact* of ρ in the present context is actually minimal.

In summary, the present application of the spatio-temporal model has simply served to verify that the OLS estimates above do not appear to be severely influenced by unobserved temporal autocorrelation effects.

But given the small-sample estimation difficulties illustrated in Section 3.1 above, it is still of interest to ask whether the present value of $\hat{\rho}$ might actually be underestimating the true value. While no definitive answer can be given to this question, it is instructive to simulate the behavior of maximum-likelihood estimates for this application.

4.3 Simulation of Spatio-Temporal Estimates

Recall from Section 3.1 that the sample size, $n = 400$, was chosen specifically to be comparable with the present application. While the results there suggest that the present estimate of ρ should behave reasonably well, both the spatio-temporal dependency structure and data used for that simulation are considerably different than the present case. Hence it is of interest to carry out the same simulation procedure using the present data set and W matrix. In addition, to gage how well the model is doing in the region of the current parameter estimates, the “true” parameter values were chosen to be (slightly) rounded versions of the parameter estimates above ($\beta_0 = 10.52$, $\beta_1 = 0.013$, $\beta_2 = 12.38$,

²³ In the present context of maximum-likelihood estimation, the AIC measure is considered by many to yields more reliable goodness-of-fit comparisons than pseudo R-square. However, the latter is somewhat easier to interpret.

$\beta_3 = 0.05$, $\lambda = 0.12$, $\rho = 0.112$, and $v = 305.01$). The results of this simulation are summarized in Table 5 below.

	β_0	β_1	β_2	β_3	λ	ρ	v
Mean	10.466	0.0134	12.398	0.0545	0.1120	0.1067	301.51
Std Dev	2.459	0.0062	5.128	0.0141	0.0524	0.0542	21.69

Table 5. Estimates for the PEC Simulation

Here it is clear that the maximum-likelihood estimates are behaving quite well for this data and set of parameter values above. Of particular interest are the ρ -estimates, which are shown in more detail in Figure 6(a) below.

Figure 6 here

As in Figure 4(a) there is still a noticeable spike at zero. But again there is strong clustering about the true value, $\rho = .112$, resulting in only a small downward bias in $\hat{\rho}$. As expected, these results for ρ lead to even better results for the estimates of variance, v , as compared to Figure 4(b). In short, these results lend further credibility to the estimate of ρ above, and hence to the minimal nature of temporal autocorrelation in the present application.

5. Comparison with a Spatio-Temporal Lag Model Approach

An alternative approach to the spatio-temporal analysis of individual housing sales with unequal time intervals was proposed by Pace et al. (2000).²⁴ Here housing sales are again ordered by time of occurrence, and the full structure of time dependencies among these sales is specified by a nonnegative matrix, $T = (\tau_{ij})$, with $\tau_{ij} = 0$ for all $j \geq i$. Similarly, all spatial dependencies for a given sale are assumed to involve only previous sales, and the structure of such dependencies is represented by a nonnegative matrix, $S = (s_{ij})$, with $s_{ij} = 0$ for all $j \geq i$. These are combined into a general linear model designated as the *spatio-temporal linear model* (STLM). For our present purposes, it is convenient to focus on the special case of STLM given by the following *spatio-temporal lag model* paralleling (16) above:²⁵

$$y = (\lambda_T T + \lambda_S S + \lambda_{TS} TS + \lambda_{ST} ST) y + X \beta + u \quad , \quad u \sim N(0, vI) \quad (34)$$

²⁴ We are indebted to an anonymous referee for pointing out the close similarities between our current model and that of Pace et al. (2000).

²⁵ This is a special case of the STLM model in expression (7) of Pace et al. (2000), where the interaction terms (TX, SX, TSX, STX) are missing, along with matrix of non-lagged variables, Z .

Here the simple expression, $\lambda W y$, in (16) now has a more elaborate form, while the unobserved residuals, u , in (17) have a much simpler form. In essence, all spatio-temporal dependencies are here postulated to be among the observable sales prices, y , and *not* the residuals, u .²⁶ As the authors point out, the appeal of this model is its mathematical simplicity. In particular, since lower triangularity is preserved under products, the spatio-temporal lag matrix in (34) is lower triangular. This, together with the standard OLS specification of residuals, implies that all parameters can be consistently estimated using OLS. Hence this model can be applied to very large data sets, as is common in real estate markets.

To compare (34) with the present model in [(16),(17)], we begin by considering the types of spatio-temporal interactions that can be captured by the product matrices, TS and ST . Here we choose to focus on ST , and write STy more explicitly as:

$$\begin{pmatrix} 0 \\ s_{21} & 0 \\ s_{31} & s_{32} & 0 \\ \vdots & \ddots & \ddots & \ddots \\ s_{n1} & \cdots & s_{n,n-2} & s_{n,n-1} & 0 \end{pmatrix} \begin{pmatrix} 0 \\ \tau_{21} & 0 \\ \tau_{31} & \tau_{32} & 0 \\ \vdots & \ddots & \ddots & \ddots \\ \tau_{n1} & \cdots & \tau_{n,n-2} & \tau_{n,n-1} & 0 \end{pmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{bmatrix} \quad (35)$$

Hence the spatio-temporal influence on the i^{th} housing sale, say $\phi_i(y_1, \dots, y_{i-1})$, is seen to be of the form:

$$\phi_i(y_1, \dots, y_{i-1}) = \sum_{h=1}^{i-1} s_{ih} \left(\sum_{j=1}^{h-1} \tau_{hj} y_j \right) = \sum_{h=1}^{i-1} \sum_{j=1}^{h-1} (s_{ih} \tau_{hj} y_j) \quad (36)$$

If we examine a typical term, $s_{ih} \tau_{hj} y_j$, then the first point to notice is that by definition, $j < h < i$. Hence all influences on the price of housing sale, i , by the prices, y_j , of previous housing sales, j , involve *indirect* influences through some other housing sale, h , that is intermediate in time between i and j . To gain some insight as to the nature of such influences, suppose first that both s_{ih} and τ_{hj} are positive so event j does exert some influence on event i through h . Suppose further that $\tau_{ih} = \tau_{ij} = 0$ (so that events h and j both happen long before i), and similarly that $s_{ij} = s_{hj} = 0$ (so that events i and h are both very far from event j in space). Then this spatio-temporal interaction can be

²⁶ It should be remarked that Pace et al. (2000) motivate the form of their model by constructing [in expressions (1), (2) and (5)] a spatio-temporal extension of the spatial Durbin model (Anselin, 1988) which does indeed account for spatio-temporal autoregressive dependencies in the unobserved residuals. But this development is somewhat misleading in the sense that their final model, STLM [expression (7)], ignores the crucial “common factor” constraints on coefficients that preserve these autoregressive dependencies. Hence, while STLM could in principle be used to *test* this “common factor hypothesis” (as implied by their discussion on p.234), the model itself is simply a more elaborate version of the spatio-temporal lag model in (34) above.

given the following interpretation: Even though event h happened long before i , it is close enough in time to event j to share some common temporal influences with j . Moreover, even though event i is very far away from j in space, it is close enough to h to share some of the spatial effects of these earlier influences. Similar interpretations can be given to TS , and show that such spatio-temporal influences are indeed quite meaningful. But the crucial point here is that these influences are *necessarily indirect*.

In our present model, for example, it is assumed that housing sale i can be influenced by all previous sales j that are sufficiently close to i in *both* space and time, i.e., for which both $\tau_{ij} > 0$ and $s_{ij} > 0$. But with respect to the above interpretation, it is clear for example that if all other sales near i happened *after* i , so that $s_{ih}\tau_{hj} = 0$ for all $h < i$, then j can have neither an “ ST ” effect or a “ TS ” effect on i in (34). While it is true that j can still have additive “ T ” and “ S ” effects in (34), it should be clear that there is no way to model our particular *joint space-time interaction effect* in the STLM model.

However, there is another matrix product that does allow such effects. In particular, if one considers the *Hadamard product*, $S \cdot T$, defined by simple component-wise multiplication, $[S \cdot T]_{ij} = (s_{ij}) \cdot (\tau_{ij})$, then this clearly encompasses the desired joint interactions since,

$$[S \cdot T]_{ij} > 0 \Leftrightarrow (s_{ij} > 0) \text{ and } (\tau_{ij} > 0) \quad (37)$$

Of equal importance in the present context is the fact that Hadamard products obviously preserve lower triangularity. Hence by broadening the spatio-temporal lag operator in (34) to

$$y = (\lambda_T T + \lambda_S S + \lambda_{TS} TS + \lambda_{ST} ST + \lambda_{S \cdot T} S \cdot T)y + X\beta + u \quad (38)$$

one can encompass both *direct* and *indirect* spatio-temporal interactions in a manner that still permits consistent OLS estimation.

Finally it is of interest to ask whether one can also use these simple weight matrices, S and T , to model spatio-temporal autocorrelation effects in the unobserved residuals, u . As in model [(16),(17)] we focus here on temporal autocorrelation and, as an alternative to (10) and (11), now consider the much simpler *linear autoregressive model*,

$$u_i = \varepsilon_i \quad (39)$$

$$u_i = \rho \sum_{j=1}^{i-1} \tau_{ij} u_j + \varepsilon_i, \quad i = 2, \dots, n \quad (40)$$

with independent innovations, $\varepsilon_i \sim N(0, \sigma^2)$, for all i . Here the unequal time intervals between i and j are assumed to be captured by appropriately chosen values of τ_{ij} . This yields the matrix form,

$$u = \rho T u + \varepsilon \Rightarrow u = (I - \rho T)^{-1} \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I). \quad (41)$$

which is seen an instance of expressions (1) and (5) in Pace et al. (2000).

From a theoretical perspective the major difficulty with this process is that *stationarity is not possible* without further modification. In particular, if it is assumed that

$$\text{var}(u_i) = v > 0, \quad i = 1, \dots, n \quad (42)$$

then it is easily shown that (41) and (42) are inconsistent, unless either $T = \mathbf{O}$ or $\rho \equiv 0$. The basic idea can be seen from (39) plus the first instance of (40). For by (39) it follows that

$$\text{var}(u_1) = \text{var}(\varepsilon_1) \Rightarrow v = \sigma^2. \quad (43)$$

Hence by setting $i = 2$ in (40) we see that

$$u_2 = \rho \tau_{21} u_1 + \varepsilon_2 \Rightarrow \text{var}(u_2) = (\rho \tau_{21})^2 \text{var}(u_1) + \text{var}(\varepsilon_2) \quad (44)$$

$$\Rightarrow v = (\rho \tau_{21})^2 v + v = v[1 + (\rho \tau_{21})^2]$$

But if $\rho \neq 0$ then $\tau_{21} \neq 0$ would imply that $(\rho \tau_{21})^2 > 0$, so that (44) can only hold if $v = 0$. Since this contradicts (42), we must have $\tau_{21} = 0$. Proceeding by induction, this forces $T = \mathbf{O}$ whenever $\rho \neq 0$.²⁷

However, it is also well known that by a simple relaxation of the variance of the initial innovation, ε_1 , this problem can be overcome for the special case of the *standard AR(1) model* [given by (10) and (11) above with $\Delta_i \equiv 1$].²⁸ Here $T = (\tau_{ij})$ takes the special form with $\tau_{i,i-1} = 1$ for all $i = 2, \dots, n$, and $\tau_{ij} = 0$ elsewhere, so that by relaxing (43) and setting $\tau_{21} = 1$, it now follows from the argument in (44) that [in a manner paralleling (12)],

$$v = \rho^2 v + \sigma^2 \Rightarrow v = \sigma^2 / (1 - \rho^2) \quad (45)$$

Hence by restricting the admissible values of ρ to the open interval $(-1, 1)$ [so that v is defined], and assuming only that the initial variance in (39) takes the form

$$\text{var}(\varepsilon_1) = \sigma^2 / (1 - \rho^2) \equiv v, \quad (44)$$

²⁷ A full argument is given in Appendix 3 (see footnote 2 above).

²⁸ In fact this “special case” provides the motivation for essentially *all* linear autoregressive models, both in time and space. For the spatial case, this is clear from the motivating examples in the original papers of Whittle (1954) and Ord (1975).

a simple inductive argument shows this initialization of the AR(1) model yields a well defined stationary process as in (42), with $\text{var}(\varepsilon_i) = \sigma^2$ for all $i > 1$.²⁹ From a practical viewpoint, this larger initial variance is taken to reflect the entire history of the unobserved process prior to $i = 1$.³⁰

Hence, there remains the interesting question of whether this simple modification for the AR(1) model might not allow variance stationarity (42) for other possible specifications of T . More precisely, one may ask whether there exist (nonzero) specifications of $T = (\tau_{ij})$ other than AR(1) [together with an appropriate choice of σ^2] such that the model:

$$u_1 = \varepsilon_1 \quad , \quad \varepsilon_1 \sim N(0, v) \quad (45)$$

$$u_i = \rho \sum_{j < i} \tau_{ij} u_j + \varepsilon_i \quad , \quad \varepsilon_i \sim N(0, \sigma^2) \quad , \quad i = 2, \dots, n \quad (46)$$

with independent innovations $(\varepsilon_i : i = 1, \dots, n)$ satisfies stationarity condition (42). Here the answer is unfortunately negative. In particular, if one imposes the reasonable assumption that more distant time influences are never greater than more recent influences, i.e., that

$$h < j < i \Rightarrow 0 \leq \tau_{ih} \leq \tau_{ij} \quad (47)$$

and requires only that model [(45),(46)] hold for ρ in some sufficiently small open interval containing zero (so that full independence is allowed), then it is shown in Appendix 3 (see footnote 2 above) that AR(1) is indeed the *only* possibility here.³¹ So while model (41) [or its relaxation in (45) and (46)] is very appealing from an analytical viewpoint, such models must always involve *non-stationary* residuals. Moreover, since the structure of these non-stationarities will depend critically on the particular specification of T , one must justify why this specification together with its implicit non-stationarities is appropriate for the particular time interval (and irregular event sequence) under study.³²

6. Concluding Remarks

In this paper, we have developed a spatio-temporal model that is particularly suitable for the analysis of address-level events occurring sporadically in time. In particular, this

²⁹ Proofs of this result can be found in any standard text, such as Hamilton (1994, Section 3.4).

³⁰ For further discussion of this point see Green (2003, Section 12.2).

³¹ It is also worth noting that this result depends only on the first two moments of the independent innovations, $(\varepsilon_i : i = 1, \dots, n)$ so that the normality assumption in (7) is not required.

³² For example, if $\rho > 0$ then it is clear from the cumulative nature of (46) that sales residuals u_i with many sales in the recent past, i.e., with many positive dependencies $(\tau_{ij} : j < i)$, will tend to have much higher variances than those with very few sales in the recent past.

model not only avoids the need for temporal aggregation (that is typical of most spatio-temporal regression models), but also allows for the possibility of short-run temporal dependencies (such as changes in asking prices based on very recent sales). Moreover, while the above analysis suggests that more robust estimation procedures may be needed for small-sample applications, the present maximum-likelihood framework does appear to be well suited for the analysis of larger data sets, as typified by housing sales transactions in major urban areas.

There is however a more subtle limitation of the present model that does not arise in more standard temporal aggregation schemes. In particular, when events are aggregated with respect to regular time intervals, as in the standard AR(1) model, these aggregate events are by definition well separated in time. But in our present extension of the AR(1) model to individual events, it is quite possible for such events to occur at almost the same time. In particular, it is possible for several house sales to occur on the same day. Moreover, as discussed in Section 2.3 above, the steady-state conditions for this model imply that the unobserved residuals for such events must be *identical*. This would make perfectly good sense if the unobserved variation captured by u_i was due *entirely* to time dependent phenomena affecting all events occurring at time t_i . For then it could be argued by simple continuity that events occurring close in time to t_i would be similarly affected. But in actuality each residual, u_i , necessarily includes any unobserved attributes of individual house i that influence its sales price but are not shared by other houses sold at times close to t_i .

Hence one important extension of the present model would be to incorporate such effects by the addition of an *idiosyncratic residual*, u_{oi} , for each sales event i . If it is assumed that these residuals are *iid* normal variates, say with $u_o = (u_{oi} : i = 1, \dots, n)' \sim N(0, v_o I_n)$, and in addition that u_o is independent of the temporal effects captured by u , then the model in (16) could be extended as follows:

$$y = \lambda W y + X \beta + u + u_o, \quad u \sim N(0, v B B'), \quad u_o \sim N(0, v_o I_n) \quad (48)$$

Here events occurring simultaneously would exhibit distinct idiosyncratic residuals even though they shared a common temporal residual.

From a conceptual viewpoint, this extension appears to be rather straightforward. But analytically it is considerably more complex. In particular, it is no longer possible to reduce the estimation problem to a single dimension. However, if this model is reparameterized by the standard technique of setting $\theta = v / v_o$, then it can be shown that (34) still exhibits many of the analytical properties developed for model (16) above. This extension will be presented in a subsequent paper.

References

- Anselin, L. (1988) *Spatial Econometrics: Methods and Models*, Kluwer, Dordrecht.
- Anselin, L. and R.J.G.M. Florax, (1995) “Small sample properties of tests for spatial dependence in regression models: Some further results”, in Anselin, L. and Florax R.J.G.M. (eds.) *New directions in spatial econometrics*, Springer: New York.
- Anselin, L. and R. Moreno (2003) “Properties of Tests for Spatial Error Components”, *Regional Science and Urban Economics*, 33: 595-618.
- Batalgi, B.H. and P.X.Wu (1999) “Unequally spaced panel data regressions with AR(1) disturbances”, *Econometric Theory*, 15: 814-823.
- Davidson, R. and J.G. MacKinnon (2004) *Econometric Theory and Methods*, Oxford University Press: New York.
- Green, W.H. (2003) *Econometric Analysis*, 5th ed., Prentice Hall: Upper Saddle River, New Jersey.
- Hamilton, J.D. (1994) *Time Series Analysis*, Princeton University Press: Princeton.
- Hinkley, D.V. and B. Efron, (1978) “Assessing the Accuracy of the Maximum Likelihood Estimator: Observed Versus Expected Fisher Information”, *Biometrika*, 65: 457-482.
- Jones, R.H., (1993) *Longitudinal Data with Serial Correlation: A State-Space Approach*, Chapman & Hall: New York.
- Jones, R.H. and F. Boadi-Boateng (1991) “Unequally spaced longitudinal data with AR(1) serial correlation”, *Biometrics*, 47: 161-175.
- Lindsay, B.G. and B. Li, (1997) “On second-order optimality of the observed Fisher information”, *The Annals of Statistics*, 25: 2172-2199
- McKenzie, C.R. and C.A. Kapuscinski, (1997) “Estimation in a linear model with serially correlated errors when observations are missing”, *Mathematics and Computers in Simulation*, 44: 1-9.
- Ord, K. (1975) “Estimation Methods for Models of Spatial Interaction”, *Journal of the American Statistical Association*, 70: 120-126.
- Pace, R.K., R. Barry, O.W. Gilley, and C.F. Sirmans, (2000) “A method for spatial-temporal forecasting with an application to real estate prices”, *International Journal of Forecasting*, 16: 229-246.
- Searle, S.R., (1971) *Linear Models*, Wiley: New York.

- Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*, Boca Raton, Florida: Chapman & Hall.
- Smith, Brent C. (2003) “The impact of community development corporations on neighborhood housing markets modeling appreciation”, *Urban Affairs Review*, 39:2, 181-204
- Wansbeek, T. and A. Kapteyn, (1985) “Estimation in a linear model with serially correlated errors when observations are missing”, *International Economic Review*, 26: 469-490.
- Whittle, P. (1954) “On Stationary Processes in the Plane”, *Biometrika*, 41: 434-439.
- Wu, P. and T.E. Smith (2009) “A spatio-temporal analysis of neighborhood change: impacts of community development projects on local property values”, *Working Paper* (in progress), Department of Electrical and Systems Engineering, University of Pennsylvania.

Figures for the Text.

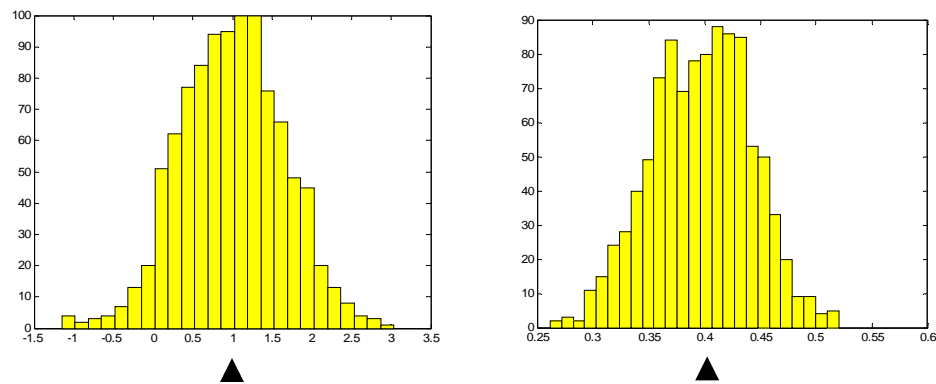


Figure 1. (a) Histogram for $\hat{\beta}_1$ (b) Histogram for $\hat{\lambda}$

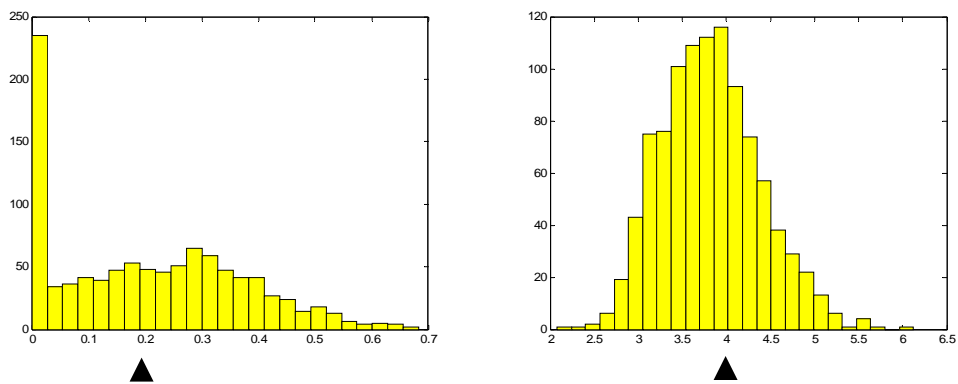


Figure 2. (a) Histogram for $\hat{\rho}$ (b) Histogram for $\hat{\nu}$

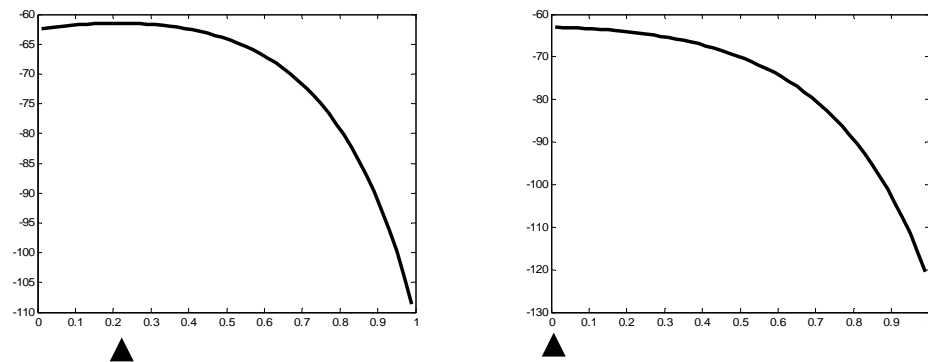


Figure 3. (a) Likelihood for $\hat{\rho} = .212$ (b) Likelihood for $\hat{\rho} \approx 0$

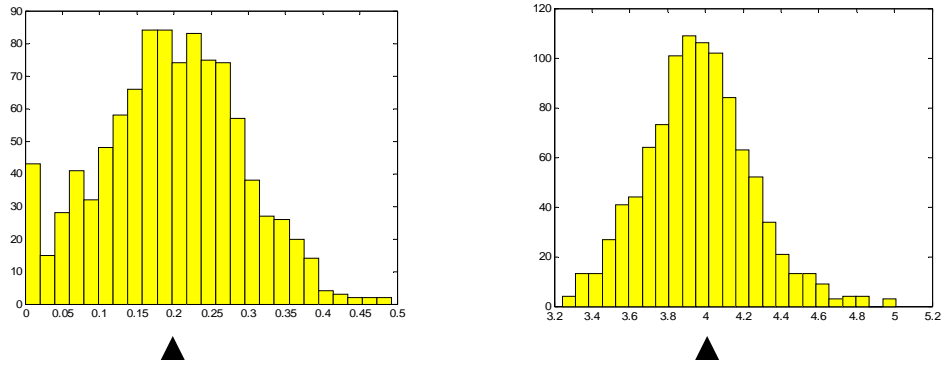


Figure 4. (a) Histogram for $\hat{\rho}$ (b) Histogram for $\hat{\nu}$

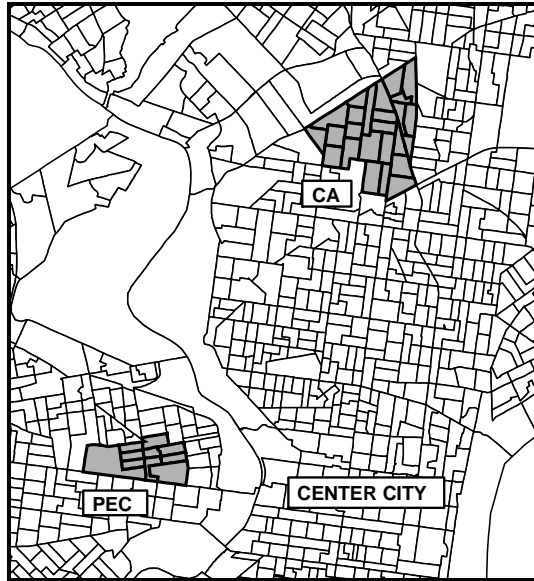


Figure 5. CDC Area (PEC) and Control Area (CA)

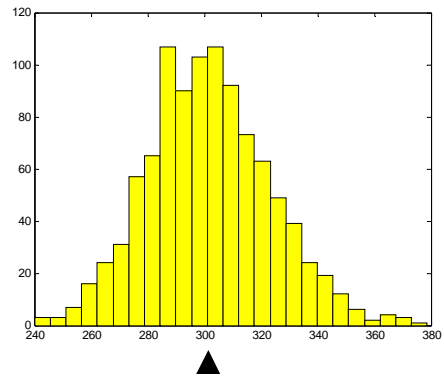
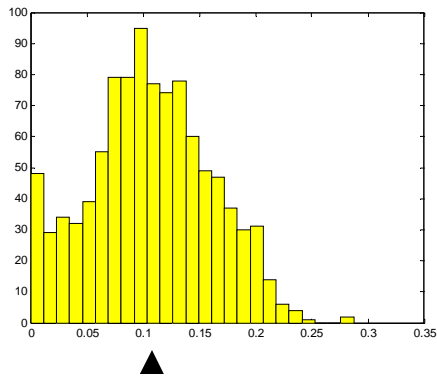


Figure 6. (a) Histogram for $\hat{\rho}$ (b) Histogram for $\hat{\nu}$

Appendix 1. Analysis of Likelihood and Concentrated Likelihood Functions.

We begin by recalling from expression (24) in the text that the full log likelihood function is given by

$$(A1.1) \quad L(\beta, v, \lambda, \rho | y) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(v) - \frac{1}{2} \ln\{\det[A^{-1}BB'(A')^{-1}]\} \\ - \frac{1}{2v} (Ay - X\beta)'(BB')^{-1}(Ay - X\beta)$$

where again $A = A(\lambda)$ and $B = B(\rho)$. Next we observe that significant reduction of the log-determinant term in (A1.1) is possible. In particular, since both W and D are lower triangular with zero diagonals, it follows that both A and $I_n - D$ are lower triangular with *unit* diagonals, and thus that

$$(A1.2) \quad \det(A) = 1 = \det(I_n - D)$$

This together with the definition of B [in (20) of the text] in turn implies that

$$(A1.3) \quad \ln\{\det[A^{-1}BB'(A')^{-1}]\} = -2\ln[\det(A)] + 2\ln[\det(B)] = (0) + 2\ln[\det\{(I_n - D)^{-1}C\}] \\ = -2\ln[\det(I_n - D)] + 2\ln[\det(C)] = (0) + 2\ln[\det(C)]$$

Hence (A1.1) reduces to

$$(A1.4) \quad L(\beta, v, \lambda, \rho | y) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(v) - \ln[\det(C)] \\ - \frac{1}{2v} (Ay - X\beta)'(BB')^{-1}(Ay - X\beta)$$

As with all general linear models of this type, the standard approach to maximum-likelihood estimation is to start by fixing the autocorrelation parameters (λ, ρ) and solving for the conditional estimates of β and v in terms of their respective first-order conditions. If we now denote the first and second order partial derivatives of $\phi(x, \dots)$ by $\nabla_x \phi$ and $\nabla_{xx} \phi$, then

$$(A1.5) \quad 0 = \nabla_{\beta} L = \frac{1}{v} X'(BB')^{-1}(Ay - X\beta) \Rightarrow (BB')^{-1}Ay = X'(BB')^{-1}X\beta \\ \Rightarrow \hat{\beta} = \hat{\beta}(\lambda, \rho) = [X'(BB')^{-1}X]^{-1}X'(BB')^{-1}Ay$$

Similarly, for v we again obtain the standard result:

$$(A1.6) \quad 0 = \nabla_v L = -\frac{n}{2v} + \frac{1}{2v^2}(Ay - X\beta)'(BB')^{-1}(Ay - X\beta)$$

$$\Rightarrow v = \frac{1}{n}(Ay - X\beta)'(BB')^{-1}(Ay - X\beta)$$

so that by substituting $\hat{\beta}$ into (A1.6) we obtain the corresponding estimate of v :

$$(A1.7) \quad \hat{v} = \hat{v}(\lambda, \rho) = \frac{1}{n}(Ay - X\hat{\beta})'(BB')^{-1}(Ay - X\hat{\beta})$$

By substituting (A1.5) and (A1.7) into (A1.4) we then obtain the following *concentrated likelihood function*, $L_c(\lambda, \rho | y)$, for (λ, ρ) :

$$(A1.8) \quad L_c(\lambda, \rho | y) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{v}) - \ln[\det(C)] - \frac{n}{2}$$

$$= -\frac{n}{2}[1 + \ln(2\pi)] - \frac{n}{2}\ln[\hat{v}(\lambda, \rho)] - \ln\{\det[C(\rho)]\}$$

where dependencies on (λ, ρ) are now made explicit. Observe in particular that λ appears only in the log-quadratic term of (A1.8). This shows that further reduction is possible in the present model by considering the first-order condition of (A1.8) with respect to λ :

$$(A1.9) \quad 0 = \nabla_\lambda L_c = -\frac{n}{2}\hat{v}(\lambda, \rho)^{-1}\nabla_\lambda \hat{v}(\lambda, \rho) \Leftrightarrow \nabla_\lambda \hat{v}(\lambda, \rho) = 0$$

To evaluate the right hand side, observe first from (A1.5),

$$(A1.10) \quad Ay - X\hat{\beta} = Ay - X[X'(BB')^{-1}X]^{-1}X'(BB')^{-1}Ay$$

$$= \{I_n - X[X'(BB')^{-1}X]^{-1}X'(BB')^{-1}\}Ay = GAy$$

where

$$(A1.11) \quad G = G(\rho) = I_n - X[X'(BB')^{-1}X]^{-1}X'(BB')^{-1}$$

so that by (A1.7),

$$(A1.12) \quad \hat{v}(\lambda, \rho) = \frac{1}{n}y'A'G'(BB')^{-1}GAy = \frac{1}{n}y'(I_n - \lambda W')G'(BB')^{-1}G(I_n - \lambda W)y$$

$$= \frac{1}{n}[y'G'(BB')^{-1}Gy - 2\lambda y'G'(BB')^{-1}GWy + \lambda^2 y'W'G'(BB')^{-1}GWy]$$

Hence the first order condition

$$(A1.13) \quad 0 = \nabla_{\lambda} \hat{v}(\lambda, \rho) = \frac{1}{n} [-2y'G'(BB')^{-1}GWy + 2\lambda y'W'G'(BB')^{-1}GWy]$$

yields the following explicit solution for λ ,

$$(A1.14) \quad \hat{\lambda} = \hat{\lambda}(\rho) = \frac{y'G'(BB')^{-1}GWy}{y'W'G'(BB')^{-1}GWy}$$

with G given by (A1.11). Finally, by substituting (A1.14) into (A1.8), we obtain the following *concentrated likelihood function* for the autocorrelation parameter, ρ ,

$$(A1.15) \quad L_c(\rho | y) = -\frac{n}{2}[1 + \ln(2\pi)] - \frac{n}{2} \ln\{\hat{v}[\hat{\lambda}(\rho), \rho]\} - \ln\{\det[C(\rho)]\}$$

This can be simplified even further by noting from (19) in the text that

$$(A1.16) \quad \det[C(\rho)] = \prod_{i=2}^n (1 - \rho^{2\Delta_i})^{1/2}$$

and hence that

$$(A1.17) \quad L_c(\rho | y) = -\frac{n}{2}[1 + \ln(2\pi)] - \frac{n}{2} \ln\{\hat{v}[\hat{\lambda}(\rho), \rho]\} - \frac{1}{2} \sum_{i=2}^n \ln(1 - \rho^{2\Delta_i})$$

Appendix 2: Derivation of the Asymptotic Covariance Matrix

First we note that in this Appendix, references will be made to expressions in Appendix 1 as well as in the text. If the full parameter vector for the log likelihood function in (24) of the text is denoted by $\theta = (\beta, v, \lambda, \rho)$, then (as with all maximum-likelihood estimates), one can approximate the asymptotic covariance matrix, $\text{cov}(\hat{\theta})$, for $\hat{\theta} = (\hat{\beta}, \hat{v}, \hat{\lambda}, \hat{\rho})$, by calculating the *Hessian matrix*,

$$(A2.1) \quad H(\hat{\theta}) = \nabla_{\theta\theta} L(\hat{\theta} | y)$$

of second partial derivatives of the full likelihood function in (A1.4) evaluated at $\hat{\theta}$, and the associated *expected Fisher information matrix*:

$$(A2.2) \quad \mathcal{I}(\hat{\theta}) = E[-H(\hat{\theta})]$$

The desired asymptotic covariance matrix is then estimated by the inverse,

$$(A2.3) \quad \text{cov}(\hat{\theta}) = \mathcal{I}(\hat{\theta})^{-1}$$

Alternatively, one can simply take

$$(A2.4) \quad \mathcal{I}_0(\hat{\theta}) = -H(\hat{\theta})$$

to be the *observed Fisher information matrix*, and estimate covariance by

$$(A2.5) \quad \text{cov}(\hat{\theta}) = \mathcal{I}_0(\hat{\theta})^{-1}$$

The second approach has the obvious advantage of avoiding the need for calculating expectations.¹ Moreover, it has been argued [for example in Hinkley and Efron (1978) and Lindsay and Li (1997)] that in many cases this observed form actually has better asymptotic properties. Finally, since one must obtain an analytical expression for the Hessian matrix,

$$(A2.6) \quad H = \begin{pmatrix} H_{\beta\beta} & H_{\beta v} & H_{\beta\lambda} & H_{\beta\rho} \\ H_{v\beta} & H_{vv} & H_{v\lambda} & H_{v\rho} \\ H_{\lambda\beta} & H_{\lambda v} & H_{\lambda\lambda} & H_{\lambda\rho} \\ H_{\rho\beta} & H_{\rho v} & H_{\rho\lambda} & H_{\rho\rho} \end{pmatrix} = \begin{pmatrix} \nabla_{\beta\beta} L & \nabla_{\beta v} L & \nabla_{\beta\lambda} L & \nabla_{\beta\rho} L \\ \nabla_{v\beta} L & \nabla_{vv} L & \nabla_{v\lambda} L & \nabla_{v\rho} L \\ \nabla_{\lambda\beta} L & \nabla_{\lambda v} L & \nabla_{\lambda\lambda} L & \nabla_{\lambda\rho} L \\ \nabla_{\rho\beta} L & \nabla_{\rho v} L & \nabla_{\rho\lambda} L & \nabla_{\rho\rho} L \end{pmatrix}$$

¹ It is also worth noting that many gradient procedures for maximizing likelihood, such as Gauss-Newton procedures, involve the computation of this matrix. Hence $H(\hat{\theta})$ is automatically available at the termination of such algorithms.

(where for example, $\nabla_{\beta v} L = \nabla_v[\nabla_{\beta} L]$), in order to calculate the associated expected Fisher information matrix,

$$(A2.7) \quad \mathcal{I} = \begin{pmatrix} \mathcal{I}_{\beta\beta} & \mathcal{I}_{\beta v} & \mathcal{I}_{\beta\lambda} & \mathcal{I}_{\beta\rho} \\ \mathcal{I}_{v\beta} & \mathcal{I}_{vv} & \mathcal{I}_{v\lambda} & \mathcal{I}_{v\rho} \\ \mathcal{I}_{\lambda\beta} & \mathcal{I}_{\lambda v} & \mathcal{I}_{\lambda\lambda} & \mathcal{I}_{\lambda\rho} \\ \mathcal{I}_{\rho\beta} & \mathcal{I}_{\rho v} & \mathcal{I}_{\rho\lambda} & \mathcal{I}_{\rho\rho} \end{pmatrix} = \begin{pmatrix} E(-H_{\beta\beta}) & E(-H_{\beta v}) & E(-H_{\beta\lambda}) & E(-H_{\beta\rho}) \\ E(-H_{v\beta}) & E(-H_{vv}) & E(-H_{v\lambda}) & E(-H_{v\rho}) \\ E(-H_{\lambda\beta}) & E(-H_{\lambda v}) & E(-H_{\lambda\lambda}) & E(-H_{\lambda\rho}) \\ E(-H_{\rho\beta}) & E(-H_{\rho v}) & E(-H_{\rho\lambda}) & E(-H_{\rho\rho}) \end{pmatrix}$$

we shall derive both matrices explicitly. To do so, we begin by recalling from (A1.4) and (A1.16) that the full likelihood function is given by

$$(A2.8) \quad L(\beta, v, \lambda, \rho | y) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(v) - \frac{1}{2} \sum_{i=2}^n \ln(1 - \rho^{2\Delta_i}) \\ - \frac{1}{2v} (Ay - X\beta)' (BB')^{-1} (Ay - X\beta)$$

and that the partials of L with respect to β and v are given from (A1.5) and (A1.6) by

$$(A2.9) \quad \nabla_{\beta} L = \frac{1}{v} X' (BB')^{-1} (Ay - X\beta) = \frac{1}{v} X' (BB')^{-1} Ay - \frac{1}{v} X' (BB')^{-1} X\beta$$

$$(A2.10) \quad \nabla_v L = -\frac{n}{2v} + \frac{1}{2v^2} (Ay - X\beta)' (BB')^{-1} (Ay - X\beta)$$

Hence

$$(A2.11) \quad H_{\beta\beta} = \nabla_{\beta\beta} L = -\frac{1}{v} X' (BB')^{-1} X$$

which in turn implies that

$$(A2.12) \quad \mathcal{I}_{\beta\beta} = E(-H_{\beta\beta}) = \frac{1}{v} X' (BB')^{-1} X$$

Similarly,

$$(A2.13) \quad H_{v\beta} = \nabla_{v\beta} L = -\frac{1}{v^2} X' (BB')^{-1} (Ay - X\beta)$$

together with the expectation, $E(Ay) = AA^{-1}X\beta = X\beta$, implies that

$$(A2.14) \quad \mathcal{I}_{v\beta} = E(-H_{v\beta}) = \frac{1}{v^2} X' (BB')^{-1} [E(Ay) - X\beta] = 0$$

Finally,

$$(A2.15) \quad H_{vv} = \nabla_{vv} L = \frac{n}{2v^2} - \frac{1}{v^3} (Ay - X\beta)' (BB')^{-1} (Ay - X\beta)$$

implies that

$$(A2.16) \quad \mathcal{I}_{vv} = E(-H_{vv}) = -\frac{n}{2v^2} + \frac{1}{v^3} E[(Ay - X\beta)'(BB')^{-1}(Ay - X\beta)]$$

To calculate the remaining expectation in (A2.14), we begin with the following result on the expectation of quadratic forms [see also Searle (1971, p.55)]. For any random n -vector, z , and $n \times n$ matrix, M ,

$$(A2.17) \quad \begin{aligned} z'Mz = \text{tr}(Mzz') &\Rightarrow E(z'Mz) = \text{tr}[M E(zz')] = \text{tr}[M \{\text{cov}(z) + E(z)E(z)'\}] \\ &\Rightarrow E(z'Mz) = \text{tr}[M \text{cov}(z)] + \text{tr}[ME(z)E(z)'] \\ &= \text{tr}[M \text{cov}(z)] + E(z)'M E(z) \end{aligned}$$

Hence, letting $z = Ay - X\beta$, $M = (BB')^{-1}$, and noting from (22) in the text that $E(z) = 0$ and $\text{cov}(z) = \text{cov}(Ay) = vBB'$, it follows from (A2.17) that

$$(A2.18) \quad \begin{aligned} E[(Ay - X\beta)'(BB')^{-1}(Ay - X\beta)] &= E[z'Mz] = \text{tr}[M \text{cov}(z)] + E(z)'M E(z) \\ &= \text{tr}[M(vBB')] + (0) = v \cdot \text{tr}((BB')^{-1}BB') = v \cdot \text{tr}(I_n) = nv \end{aligned}$$

and we may conclude that

$$(A2.19) \quad \mathcal{I}_{vv} = -\frac{n}{2v^2} + \frac{1}{v^3}(nv) = \frac{n}{2v^2}$$

Turning next to λ , observe first from (A2.8) that

$$(A2.20) \quad \begin{aligned} \nabla_{\lambda} L &= -\frac{1}{2v} \nabla_{\lambda} [(y - \lambda Wy - X\beta)'(BB')^{-1}(y - \lambda Wy - X\beta)] \\ &= -\frac{2}{2v} (y - \lambda Wy - X\beta)'(BB')^{-1}(-Wy) = \frac{1}{v} (y - \lambda Wy - X\beta)'(BB')^{-1}Wy \end{aligned}$$

implies

$$(A2.21) \quad H_{\lambda\lambda} = \nabla_{\lambda} [\nabla_{\lambda} L] = -\frac{1}{v} y'W'(BB')^{-1}Wy$$

so that again by (A2.17) with $M = W'(BB')^{-1}W$ and $z = y$, together with (22),

$$(A2.22) \quad \begin{aligned} \mathcal{I}_{\lambda\lambda} = E(-H_{\lambda\lambda}) &= \frac{1}{v} E[y'W'(BB')^{-1}Wy] \\ &= \frac{1}{v} \{ \text{tr}[W'(BB')^{-1}W \text{cov}(y)] + E(y)'[W'(BB')^{-1}W]E(y) \} \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{v} \{ \text{tr}[W'(BB')^{-1}W\{vA^{-1}BB'(A')^{-1}\}] + (A^{-1}X\beta)'[W'(BB')^{-1}W]A^{-1}X\beta \} \\
&= \text{tr}[W'(BB')^{-1}WA^{-1}BB'(A')^{-1}] + \frac{1}{v} \beta'X'(A')^{-1}W'(BB')^{-1}WA^{-1}X\beta \\
&= \text{tr}[(BB')^{-1}(WA^{-1})BB'(WA^{-1})'] + \frac{1}{v} \beta'X'(WA^{-1})'(BB')^{-1}(WA^{-1})X\beta
\end{aligned}$$

Next by (A2.9),

$$(A2.23) \quad H_{\beta\lambda} = \nabla_{\beta\lambda} L = \frac{1}{v} \nabla_{\lambda} [X'(BB')^{-1}(y - \lambda Wy)] = -\frac{1}{v} X'(BB')^{-1}Wy$$

which implies that

$$(A2.24) \quad \mathcal{I}_{\beta\lambda} = E(-H_{\beta\lambda}) = \frac{1}{v} X'(BB')^{-1}WE(y) = \frac{1}{v} X'(BB')^{-1}WA^{-1}X\beta$$

Similarly by (A2.10)

$$\begin{aligned}
(A2.25) \quad H_{v\lambda} &= \nabla_{\lambda} [\nabla_v L] = \frac{1}{2v^2} \nabla_{\lambda} [(Ay - X\beta)'(BB')^{-1}(Ay - X\beta)] \\
&= \frac{2}{2v^2} (y - \lambda Wy - X\beta)'(BB')^{-1}(-Wy) = -\frac{1}{v^2} (Ay - X\beta)'(BB')^{-1}Wy
\end{aligned}$$

which implies

$$(A2.26) \quad \mathcal{I}_{v\lambda} = E(-H_{v\lambda}) = \frac{1}{v^2} E[(Ay - X\beta)'(BB')^{-1}Wy]$$

To evaluate this expectation, first observe that again by (A2.17) with $M = A'(BB')^{-1}W$ and $z = y$,

$$\begin{aligned}
(A2.27) \quad E[(Ay - X\beta)'(BB')^{-1}Wy] &= E[y'A'(BB')^{-1}Wy] - E[\beta'X'(BB')^{-1}Wy] \\
&= \text{tr}[A'(BB')^{-1}W \text{cov}(y)] + E(y)'[A'(BB')^{-1}W]E(y) - \beta'X'(BB')^{-1}WE(y)
\end{aligned}$$

But since

$$\begin{aligned}
(A2.28) \quad E(y)'[A'(BB')^{-1}W]E(y) - \beta'X'(BB')^{-1}WE(y) \\
&= \beta'X'(A')^{-1}[A'(BB')^{-1}W]A^{-1}X\beta - \beta'X'(BB')^{-1}WA^{-1}X\beta \\
&= \beta'X'[(A')^{-1}A'](BB')^{-1}WA^{-1}X\beta - \beta'X'(BB')^{-1}WA^{-1}X\beta = 0
\end{aligned}$$

and $\text{cov}(y) = vA^{-1}BB'(A')^{-1}$, it follows that (A2.27) reduces to

$$\begin{aligned}
(A2.29) \quad E[(Ay - X\beta)'(BB')^{-1}Wy] &= v \cdot tr[\{A'(BB')^{-1}\}WA^{-1}BB'(A')^{-1}] \\
&= v \cdot tr[WA^{-1}BB'(A')^{-1}\{A'(BB')^{-1}\}] \\
&= v \cdot tr(WA^{-1})
\end{aligned}$$

Finally, since W is by construction lower triangular with zero diagonal, it follows that $tr(WA^{-1}) = 0$. To see this, observe first that for any lower triangular matrix, T , with zero diagonal, the determinant $|I - \theta T|$ is a product of ones and hence is unity. But the characteristic equation for WA^{-1} is then of the form

$$\begin{aligned}
(A2.30) \quad 0 &= |I - \theta WA^{-1}| = |A^{-1}| \cdot |A - \theta W| = |A|^{-1} |A - \theta W| \\
&= |I - \lambda W|^{-1} |(I - \lambda W) - \theta W| \\
&= (1) \cdot |(I - \lambda W) - \theta W| = |I - (\lambda + \theta)W| = 1
\end{aligned}$$

Hence the characteristic equation for WA^{-1} can have no non-zero solutions, and it follows that all eigenvalues of WA^{-1} are *zero*. Finally, since the trace of any matrix is the sum of its eigenvalues, this implies in particular that $tr(WA^{-1}) = 0$ and hence that

$$(A2.31) \quad \mathcal{I}_{v\lambda} = 0$$

Finally we turn to those terms involving the key parameter, ρ , which are by far the most complex to compute. Again we start with the partial of L in (A2.8) with respect to ρ :

$$(A2.32) \quad \nabla_{\rho} L = -\frac{1}{2} \sum_{i=2}^n \left[\nabla_{\rho} \ln(1 - \rho^{2\Delta_i}) \right] - \frac{1}{2v} (Ay - X\beta)' [\nabla_{\rho} (BB')^{-1}] (Ay - X\beta)$$

Here we consider each term in turn. First

$$\begin{aligned}
(A2.33) \quad \nabla_{\rho} \ln(1 - \rho^{2\Delta_i}) &= -2\Delta_i \rho^{2\Delta_i-1} (1 - \rho^{2\Delta_i})^{-1} \\
\Rightarrow -\frac{1}{2} \sum_{i=2}^n \left[\nabla_{\rho} \ln(1 - \rho^{2\Delta_i}) \right] &= \sum_{i=2}^n \Delta_i \rho^{2\Delta_i-1} (1 - \rho^{2\Delta_i})^{-1}
\end{aligned}$$

Next to calculate $\nabla_{\rho} (BB')^{-1}$ it is convenient to recall that $B = (I_n - D)^{-1}C$, so that by letting $M = C^{-2}$ and $G = I_n - D$, it follows (from the symmetry of C) that

$$(A2.34) \quad (BB')^{-1} = (B^{-1})' B^{-1} = [C^{-1}(I_n - D)]' [C^{-1}(I_n - D)]$$

$$= (I_n - D)'C^{-2}(I_n - D) = G'MG$$

Similarly, if we let $\nabla_\rho(G) = G_\rho$ and $\nabla_\rho(M) = M_\rho$ then by the chain rule for matrix derivatives,

$$(A2.35) \quad \begin{aligned} \nabla_\rho(BB')^{-1} &= \nabla_\rho[G'MG] = [\nabla_\rho(G')]MG + G'[\nabla_\rho(M)]G + G'M[\nabla_\rho(G)] \\ &= G'_\rho MG + G'M_\rho G + G'MG_\rho \end{aligned}$$

Hence it suffices to compute G_ρ and M_ρ . To do so, it is convenient to give explicit representations of $D = [D(i, j) : i, j = 1, \dots, n]$ and $C = [C(i, j) : i, j = 1, \dots, n]$ as follows. First by (18) in the text, it follows that

$$(A2.36) \quad D(i, j) = \begin{cases} \rho^{\Delta_i}, & i = 2, \dots, n, j = i - 1 \\ 0, & \text{otherwise} \end{cases}$$

so that $G = [G(i, j) : i, j = 1, \dots, n]$ is of the form

$$(A2.37) \quad G(i, j) = \begin{cases} 1 & , i = j = 1, \dots, n \\ -\rho^{\Delta_i} & , i = 2, \dots, n, j = i - 1 \\ 0 & , \text{elsewhere} \end{cases}$$

Similarly, by (19) in the text,

$$(A2.38) \quad C(i, j) = \begin{cases} 1 & , i = j = 1 \\ (1 - \rho^{2\Delta_i})^{1/2} & , i = j = 2, \dots, n, \\ 0 & , \text{elsewhere} \end{cases}$$

so that $M = [M(i, j) : i, j = 1, \dots, n]$ is of the form

$$(A2.39) \quad M(i, j) = \begin{cases} 1 & , i = j = 1 \\ (1 - \rho^{2\Delta_i})^{-1} & , i = j = 2, \dots, n, \\ 0 & , \text{elsewhere} \end{cases}$$

Hence, the gradient condition,

$$(A2.40) \quad G_\rho = \nabla_\rho(I_n - D) = -\nabla_\rho(D)$$

together with $\nabla_\rho(\rho^{\Delta_i}) = \Delta_i \rho^{\Delta_i - 1}$ implies that $G_\rho = [G_\rho(i, j) : i, j = 1, \dots, n]$ is of the form,

$$(A2.41) \quad G_\rho(i, j) = \begin{cases} -\Delta_i \rho^{\Delta_i-1}, & i = 2, \dots, n, j = i-1 \\ 0, & \text{otherwise} \end{cases}$$

Similarly, (A2.39) together with

$$(A2.42) \quad \nabla_\rho \left[(1 - \rho^{2\Delta_i})^{-1} \right] = -(1 - \rho^{2\Delta_i})^{-2} (-2\Delta_i \rho^{2\Delta_i-1}) = 2\Delta_i \rho^{2\Delta_i-1} (1 - \rho^{2\Delta_i})^{-2}$$

implies that $M_\rho = [M_\rho(i, j) : i, j = 1, \dots, n]$ is of the form

$$(A2.43) \quad M_\rho(i, j) = \begin{cases} 2\Delta_i \rho^{2\Delta_i-1} (1 - \rho^{2\Delta_i})^{-2}, & i = j = 2, \dots, n \\ 0, & \text{otherwise} \end{cases}$$

Hence, $\nabla_\rho (BB')^{-1}$ is given by (A2.35) together with [(A2.37), (A2.39), (A2.41), (A2.43)].

These results complete the calculation of $\nabla_\rho L$ in (A2.32). Hence to calculate the Hessian element $H_{\rho\rho} = \nabla_{\rho\rho} L$ we see from (A2.32) that

$$(A2.44) \quad \nabla_{\rho\rho} L = -\frac{1}{2} \sum_{i=2}^n \nabla_{\rho\rho} \left[\ln(1 - \rho^{2\Delta_i}) \right] - \frac{1}{2v} (Ay - X\beta)' [\nabla_{\rho\rho} (BB')^{-1}] (Ay - X\beta)$$

and that from (A2.33)

$$(A2.45) \quad \begin{aligned} \nabla_{\rho\rho} \left[\ln(1 - \rho^{2\Delta_i}) \right] &= \nabla_\rho \left[\Delta_i \rho^{2\Delta_i-1} (1 - \rho^{2\Delta_i})^{-1} \right] \\ &= \Delta_i \rho^{2\Delta_i-2} (1 - \rho^{2\Delta_i})^{-1} \left[(2\Delta_i - 1) + 2\Delta_i \rho^{2\Delta_i} (1 - \rho^{2\Delta_i})^{-1} \right] \end{aligned}$$

Hence it remains to calculate the second partial derivative, $\nabla_{\rho\rho} (BB')^{-1}$. As with the first partial derivative, we start with the chain rule applied to (A2.35):

$$(A2.46) \quad \begin{aligned} \nabla_{\rho\rho} (BB')^{-1} &= \nabla_\rho [G'_\rho M G + G' M_\rho G + G' M G_\rho] = [G'_{\rho\rho} M G + G'_\rho M_\rho G + G'_\rho M G_\rho] \\ &\quad + [G'_\rho M_\rho G + G' M_{\rho\rho} G + G' M_\rho G_\rho] + [G'_\rho M G_\rho + G' M_\rho G_\rho + G' M G_{\rho\rho}] \end{aligned}$$

which shows that it suffices to calculate $\nabla_{\rho\rho} G$ and $\nabla_{\rho\rho} M$. But since

$$(A2.47) \quad \nabla_\rho [-\Delta_i \rho^{\Delta_i-1}] = -\Delta_i (1 - \Delta_i) \rho^{\Delta_i-2}$$

it follows from (A2.41) that $G_{\rho\rho} = [G_{\rho\rho}(i, j) : i, j = 1, \dots, n]$ is given by

$$(A2.48) \quad G_{\rho\rho}(i, j) = \begin{cases} -\Delta_i(1-\Delta_i)\rho^{\Delta_i-2}, & i = 2, \dots, n, j = i-1 \\ 0 & , \text{ otherwise} \end{cases}$$

Similarly,

$$(A2.49) \quad M_{\rho\rho}(i, j) = \begin{cases} \nabla_{\rho}[2\Delta_i\rho^{2\Delta_i-1}(1-\rho^{2\Delta_i})^{-2}] & , i = j = 2, \dots, n \\ 0 & , \text{ otherwise} \end{cases}$$

where

$$(A2.50) \quad \nabla_{\rho}[2\Delta_i\rho^{2\Delta_i-1}(1-\rho^{2\Delta_i})^{-2}] = 8\Delta_i^2\rho^{4\Delta_i-2}(1-\rho^{2\Delta_i})^{-3} + 2\Delta_i(2\Delta_i-1)\rho^{2\Delta_i-2}(1-\rho^{2\Delta_i})^{-2}$$

Hence $\nabla_{\rho\rho}(BB')^{-1}$ is given by (A2.44) together with [(A2.37), (A2.39), (A2.41), (A2.43)] and [(A2.48), (A2.49), (A2.50)]. In turn this implies that $H_{\rho\rho}$ is given by (A2.44), (A2.45) and this calculation of $\nabla_{\rho\rho}(BB')^{-1}$. Turning now to $\mathcal{I}_{\rho\rho}$, it follows from (A2.44) that,

$$(A2.51) \quad \begin{aligned} \mathcal{I}_{\rho\rho} &= E(-H_{\rho\rho}) \\ &= \frac{1}{2} \sum_{i=2}^n \nabla_{\rho\rho} [\ln(1-\rho^{2\Delta_i})] + \frac{1}{2\nu} E\{(Ay - X\beta)'[\nabla_{\rho\rho}(BB')^{-1}](Ay - X\beta)\} \end{aligned}$$

But by setting $M = \nabla_{\rho\rho}(BB')^{-1}$ in (A2.18), it now follows that

$$(A2.52) \quad E\{(Ay - X\beta)'[\nabla_{\rho\rho}(BB')^{-1}](Ay - X\beta)\} = \nu \cdot \text{tr}[\nabla_{\rho\rho}(BB')^{-1}BB']$$

and hence that

$$(A2.53) \quad \mathcal{I}_{\rho\rho} = \frac{1}{2} \sum_{i=2}^n \nabla_{\rho\rho} [\ln(1-\rho^{2\Delta_i})] + \frac{1}{2} \text{tr}[\nabla_{\rho\rho}(BB')^{-1}BB']$$

Next, to calculate $H_{\rho\lambda}$ we observe from (A2.32) that

$$(A2.54) \quad \begin{aligned} H_{\rho\lambda} &= \nabla_{\rho\lambda} L = -\frac{1}{2\nu} \nabla_{\lambda} \{(Ay - X\beta)'[\nabla_{\rho}(BB')^{-1}](Ay - X\beta)\} \\ &= -\frac{1}{2\nu} \{2(Ay - X\beta)'[\nabla_{\rho}(BB')^{-1}]\nabla_{\lambda}(y - \lambda Wy - X\beta)\} \\ &= \frac{1}{\nu} (Ay - X\beta)'[\nabla_{\rho}(BB')^{-1}]Wy \end{aligned}$$

Hence,

$$(A2.55) \quad \mathcal{I}_{\rho\lambda} = E(-H_{\rho\lambda}) = -\frac{1}{\nu} E\{(Ay - X\beta)'[\nabla_{\rho}(BB')^{-1}]Wy\}$$

$$\begin{aligned}
&= -\frac{1}{v} \left[E\{y'A'[\nabla_{\rho}(BB')^{-1}]Wy\} - \beta'X'[\nabla_{\rho}(BB')^{-1}]WE(y) \right] \\
&= -\frac{1}{v} \left[E\{y'A'[\nabla_{\rho}(BB')^{-1}]Wy\} - \beta'X'[\nabla_{\rho}(BB')^{-1}]W(A^{-1}X\beta) \right]
\end{aligned}$$

But by setting $M = A'[\nabla_{\rho}(BB')^{-1}]W$ and $z = y$ in (A2.17) it follows that

$$\begin{aligned}
\text{(A2.56)} \quad &E\{y'A'[\nabla_{\rho}(BB')^{-1}]Wy\} \\
&= \text{tr}\{A'[\nabla_{\rho}(BB')^{-1}]W \text{cov}(y)\} + E(y)'A'[\nabla_{\rho}(BB')^{-1}]WE(y) \\
&= \text{tr}\{A'[\nabla_{\rho}(BB')^{-1}]W[vA^{-1}BB'(A')^{-1}]\} + \beta'X'[(A')^{-1}A'][\nabla_{\rho}(BB')^{-1}]WA^{-1}X\beta \\
&= \text{tr}\{[\nabla_{\rho}(BB')^{-1}]WvA^{-1}BB'[(A')^{-1}A']\} + \beta'X'[\nabla_{\rho}(BB')^{-1}]WA^{-1}X\beta \\
&= v \cdot \text{tr}\{[\nabla_{\rho}(BB')^{-1}]WA^{-1}BB'\} + \beta'X'[\nabla_{\rho}(BB')^{-1}]WA^{-1}X\beta
\end{aligned}$$

Hence by substituting this back into (A2.55), it is seen that the last two terms cancel, leaving

$$\text{(A2.57)} \quad \mathcal{I}_{\rho\lambda} = -\text{tr}\{[\nabla_{\rho}(BB')^{-1}]WA^{-1}BB'\}$$

To complete the calculations, we observe first from (A2.9) that

$$\text{(A2.58)} \quad H_{\beta\rho} = \nabla_{\beta\rho}L = \frac{1}{v} X'[\nabla_{\rho}(BB')^{-1}](Ay - X\beta)$$

which is directly obtainable from the above calculation of $\nabla_{\rho}(BB')^{-1}$. Moreover,

$$\begin{aligned}
\text{(A2.59)} \quad \mathcal{I}_{\beta\rho} &= E(-H_{\beta\rho}) = \frac{1}{v} X'[\nabla_{\rho}(BB')^{-1}][E(Ay) - X\beta] \\
&= \frac{1}{v} X'[\nabla_{\rho}(BB')^{-1}][X\beta - X\beta] = 0
\end{aligned}$$

Finally, to obtain $H_{v\rho}$, observe from (A2.10) that

$$\text{(A2.60)} \quad H_{v\rho} = \nabla_{v\rho}L = \frac{1}{2v^2}(Ay - X\beta)'[\nabla_{\rho}(BB')^{-1}](Ay - X\beta)$$

which is again obtainable from the calculation of $\nabla_{\rho}(BB')^{-1}$. Moreover, by the same argument as in (A2.52) it now follows that

$$\begin{aligned}
\text{(A2.61)} \quad \mathcal{I}_{\lambda\rho} &= E(-H_{\lambda\rho}) = -\frac{1}{2v^2} E\{(Ay - X\beta)'[\nabla_{\rho}(BB')^{-1}](Ay - X\beta)\} \\
&= -\frac{1}{2v^2} (v \cdot \text{tr}\{[\nabla_{\rho}(BB')^{-1}]BB'\}) \\
&= -\frac{1}{2v^2} \text{tr}\{[\nabla_{\rho}(BB')^{-1}]BB'\}
\end{aligned}$$

This together with the observation $H_{ji} = H'_{ij}$ and $\mathcal{I}_{ji} = \mathcal{I}'_{ij}$ for all $i, j \in \{\beta, v, \lambda, \rho\}$ with $i \neq j$ is seen to complete the calculation of both H and \mathcal{I} in (A2.6) and (A2.7), respectively.

Appendix 3. Stationarity in Linear Autoregressive Models

As in the text, we assume throughout this appendix that the matrix, $T = (\tau_{ij} : i, j = 1, \dots, n)$, is nonnegative and lower triangular, i.e., that $\tau_{ij} \geq 0$ for all $j < i = 2, \dots, n$ and $\tau_{ij} = 0$ otherwise. In this context, we begin by recalling the linear autoregressive scheme of (39) and (40) in the text,

$$(A3.1) \quad u_1 = \varepsilon_1$$

$$(A3.2) \quad u_i = \rho \sum_{j=1}^{i-1} \tau_{ij} u_j + \varepsilon_i, \quad i = 2, \dots, n$$

with independent innovations, $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, \dots, n$ [which are implicitly assumed to be ordered in time, so that ε_1 is the first innovation, and so on]. This model can be summarized in matrix form as:

$$(A3.3) \quad u = \rho T u + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I),$$

Our first result is to show that this model can be stationary only under *complete independence*,¹ i.e., that

Proposition 1. *Model (A3.3) is compatible with the variance stationarity condition,*

$$(A3.4) \quad \text{var}(u_i) = v > 0, \quad i = 1, \dots, n$$

iff it is true that either $\rho = 0$ or $T = O$.

Proof: If either $\rho = 0$ or $T = O$ then the result holds trivially by setting $v = \sigma^2$. Hence it suffices to establish the necessity of these conditions. Here it is enough to show that if (A3.4) holds with $\rho \neq 0$ then $T = O$. We proceed by induction, and for sake of completeness, begin by repeating the argument for $i = 2$ in (43) and (44) of the text. Since (A3.1) yields the implication,

$$(A3.5) \quad \text{var}(u_1) = \text{var}(\varepsilon_1) \Rightarrow v = \sigma^2,$$

it follows by setting $i = 2$ in (A3.2) that

$$(A3.6) \quad u_2 = \rho \tau_{21} u_1 + \varepsilon_2 \Rightarrow \text{var}(u_2) = (\rho \tau_{21})^2 \text{var}(u_1) + \text{var}(\varepsilon_2)$$

¹ It should be noted that none of the results in this Appendix require either normality of innovations, ε , or nonnegativity of T . However, since these are natural parts of the models being considered, we shall retain these assumptions throughout.

$$\Rightarrow v = (\rho\tau_{21})^2 v + v = v[1 + (\rho\tau_{21})^2].$$

Hence if $\rho \neq 0$ then $\tau_{21} \neq 0$ would imply that $(\rho\tau_{21})^2 > 0$, and thus that $v = 0$ in (A3.6). But since this contradicts (A3.4), it follows that (A3.2) can only hold for $i = 2$ if $\tau_{21} = 0$, which in turn implies that $u_2 = \varepsilon_2$. More generally, we may proceed by induction as follows. If for any $k \geq 2$ it is true that $(\tau_{ij} = 0 : j \leq i = 1, \dots, k)$, so that $u_j = \varepsilon_j$ for all $j = 1, \dots, k$ then by (A3.2)

$$(A3.7) \quad u_{k+1} = \rho \sum_{j=1}^k \tau_{k+1,j} u_j + \varepsilon_{k+1} = \rho \sum_{j=1}^k \tau_{k+1,j} \varepsilon_j + \varepsilon_{k+1}$$

$$\Rightarrow \text{var}(u_{k+1}) = \rho^2 \sum_{j=1}^k \tau_{k+1,j}^2 \text{var}(\varepsilon_j) + \text{var}(\varepsilon_{k+1})$$

$$\Rightarrow v = \rho^2 \sum_{j=1}^k \tau_{k+1,j}^2 v + v = v \left(1 + \sum_{j=1}^k (\rho\tau_{k+1,j})^2 \right)$$

But since $(\rho\tau_{k+1,j})^2 \geq 0$ for all $j = 1, \dots, k$, and since $\rho \neq 0$ must imply that $(\rho\tau_{k+1,j})^2 > 0$ whenever $\tau_{k+1,j} \neq 0$, it again follows that $v > 0$ is only possible if $(\tau_{k+1,j} = 0 : j < k+1)$. Hence by induction (together with the lower-triangular definition of T) we may conclude that if $\rho \neq 0$ then $T = O$. ■

To broaden the range of variance stationary models to include certain models with autoregressive dependencies, we now relax the initial variance of ε_1 and consider the model,

$$(A3.8) \quad u_1 = \varepsilon_1 \sim N(0, v)$$

$$(A3.9) \quad u_i = \rho \sum_{j=1}^{i-1} \tau_{ij} u_j + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 2, \dots, n$$

with $(\varepsilon_i : i = 1, \dots, n)$ independent. Here we require some additional structural assumptions to ensure that variance stationarity (A3.4) is at least possible, and also to eliminate unwanted cases. First observe from the argument in (A3.6) above that if (A3.4) holds then

$$(A3.10) \quad \text{var}(u_2) = (\rho\tau_{21})^2 \text{var}(u_1) + \text{var}(\varepsilon_2) \Rightarrow v = (\rho\tau_{21})^2 v + \sigma^2$$

$$\Rightarrow v = \frac{\sigma^2}{1 - (\rho\tau_{21})^2}$$

Hence v is only finite and positive if $|\rho\tau_{21}| < 1$. Moreover, if $\tau_{21} = 0$, so that $v = \sigma^2$ [as in (A3.5) above] then the argument of Proposition 1 goes through. Thus to avoid restricting the model to independent processes with $T = 0$, we now require that

$$(A3.11) \quad \tau_{21} > 0 .$$

This ensures that variance stationarity is at least possible for sufficiently small values of ρ , i.e., with $|\rho| < 1/\tau_{21}$. In this context, we require that the set, R , of *admissible values* for ρ contain all such values that are sufficiently close to 0, i.e., that

$$(A3.12) \quad (-\rho_0, \rho_0) \subset R \text{ for some } \rho_0 \in (0, 1/\tau_{21}) .$$

This ensures in particular that $0 \in R$, and hence that “complete independence” continues to be one possible model.² Next recall from (A3.9) that τ_{ij} reflects the relative influence on event i of prior event j . Hence it is natural to assume that more recent events, j , are at least as influential as those in the distant past, i.e., that

$$(A3.13) \quad h < j < i \Rightarrow \tau_{ih} \leq \tau_{ij}$$

The set of conditions [(A3.8),(A3.9),(A3.11),(A3.12),(A3.13)] is here said to define a *linear autoregressive model*.

It is convenient to designate the special case with $\tau_{i,i-1} = \tau_{21}$ for all $i = 3, \dots, n$ and $\tau_{ij} = 0$ elsewhere as the *AR(1) model*. Notice that one can always set $\tau_{21} = 1$ by simply rescaling the autoregressive parameter, ρ , to the value $\tilde{\rho} = \rho/\tau_{21}$ [with admissible range, $\tilde{R} = \{\rho/\tau_{21} : \rho \in R\}$]. But in the arguments to follow it is convenient to represent all appearances of τ_{21} explicitly. With these conventions, our objective is to show that AR(1) is the *only* linear autoregressive model satisfying variance stationarity (A3.4). Before doing however, it is instructive to observe that without condition (A3.13) this is not true. For example, consider the linear autoregressive model with $\tau_{i,1} = 1$, $i = 2, \dots, n$ and $\tau_{ij} = 0$ otherwise. Here all events after the first event depend only ε_1 , so that for all $i = 2, \dots, n$

$$(A3.14) \quad u_i = \rho u_1 + \varepsilon_i = \rho \varepsilon_1 + \varepsilon_i \Rightarrow \text{var}(u_i) = \rho^2 \text{var}(u_1) + \text{var}(\varepsilon_i)$$

$$\Rightarrow v = \rho^2 v + \sigma^2 \Rightarrow v = \frac{\sigma^2}{1 - \rho^2}$$

² From an analytical perspective, condition (A3.12) will also be seen to ensure that all variances, $\text{var}(u_i)$, are continuously differentiable functions of ρ in an open neighborhood of $\rho = 0$.

Hence for $\text{var}(\varepsilon_1) = \sigma^2 / (1 - \rho^2) = v$, this model is trivially stationary. But dependencies of this type are hardly what one thinks of as being “autoregressive”. Hence requirement (A3.13) is designed to eliminate such cases. With these preliminary observations, our main result is to show that:

Proposition . *The only linear autoregressive model that is compatible with variance stationarity (A3.4) is the AR(1) model.*

Proof: As mentioned in the text, the proof that an AR(1) process is variance stationary is completely standard (using the rescaling, $\rho \rightarrow \rho / \tau_{21}$, if necessary), and is omitted. Hence the objective of this proof is to show that variance stationarity (A3.4) implies that a linear autoregressive process must be AR(1). To do so, we now assume that (A3.4) holds, and proceed by induction. If for notational simplicity, we set $\tau = \tau_{21} > 0$, then again by (A3.10) it follows for $i = 2$ that $u_2 = \rho \tau \varepsilon_1 + \varepsilon_2$ and $\sigma^2 = v(1 - \rho^2 \tau^2)$. Hence for $i = 3$ we have

$$\begin{aligned}
\text{(A3.15)} \quad u_3 &= \rho(\tau_{31}u_1 + \tau_{32}u_2) + \varepsilon_3 = \rho\tau_{31}\varepsilon_1 + \rho\tau_{32}(\rho\tau\varepsilon_1 + \varepsilon_2) + \varepsilon_3 \\
&= (\rho\tau_{31} + \rho^2\tau_{32}\tau)\varepsilon_1 + \rho\tau_{32}\varepsilon_2 + \varepsilon_3 \\
&\Rightarrow \text{var}(u_3) = (\rho\tau_{31} + \rho^2\tau_{32}\tau)^2 \text{var}(\varepsilon_1) + (\rho\tau_{32})^2 \text{var}(\varepsilon_2) + \text{var}(\varepsilon_3) \\
&\Rightarrow v = (\rho\tau_{31} + \rho^2\tau_{32}\tau)^2 v + [(\rho\tau_{32})^2 + 1] v(1 - \rho^2\tau^2) \\
&\Rightarrow 1 = (\rho\tau_{31} + \rho^2\tau_{32}\tau)^2 + [(\rho\tau_{32})^2 + 1] (1 - \rho^2\tau^2)
\end{aligned}$$

for all $\rho \in R$. Since this last expression is continuously differentiable on the open interval $(-\rho_0, \rho_0)$ in (A3.12), we may differentiate this identity with respect to ρ and obtain the new identity,

$$\text{(A3.16)} \quad 0 = 2(\rho\tau_{31} + \rho^2\tau_{32}\tau)(\tau_{31} + 2\rho\tau_{32}\tau) + (2\rho\tau_{32}^2)(1 - \rho^2\tau^2) - (2\rho\tau^2)(\rho^2\tau_{32}^2 + 1)$$

for $\rho \in (-\rho_0, \rho_0)$. Next dividing by any $\rho \in (0, \rho_0)$ we obtain the reduced identity

$$\text{(A3.17)} \quad 0 = 2(\tau_{31} + \rho\tau_{32}\tau)(\tau_{31} + 2\rho\tau_{32}\tau) + (2\tau_{32}^2)(1 - \rho^2\tau^2) - (2\tau^2)(\rho^2\tau_{32}^2 + 1)$$

for $\rho \in (0, \rho_0)$. Finally, letting $\rho \downarrow 0$ and appealing to the continuity of this expression at $\rho = 0$, we obtain the limiting relation:

$$\text{(A3.18)} \quad 0 = 2\tau_{31}^2 + 2\tau_{32}^2 - 2\tau^2 \Rightarrow \tau^2 = \tau_{31}^2 + \tau_{32}^2$$

Next, by differentiating (A3.17) once more (and dividing by 2) we obtain the new identity

$$(A3.19) \quad 0 = (\tau_{32}\tau)(\tau_{31} + 2\rho\tau_{32}\tau) + (\tau_{31} + \rho\tau_{32}\tau)(2\tau_{32}\tau) - 4\rho\tau_{32}^2\tau^2$$

for $\rho \in (0, \rho_0)$. Hence letting $\rho \downarrow 0$ and again appealing to continuity, we see that at $\rho = 0$,

$$(A3.20) \quad 0 = \tau_{31}\tau_{32}\tau + 2\tau_{31}\tau_{32}\tau = 3\tau_{31}\tau_{32}\tau$$

$$\Rightarrow \tau\tau_{32}\tau_{31} = 0$$

But since $\tau = \tau_{21} > 0$ by (A3.11) we must have $\tau_{32}\tau_{31} = 0$. Moreover, since $0 \leq \tau_{31} \leq \tau_{32}$ by nonnegativity together with (A3.13), it then follows that $\tau_{31} = 0$.³ Finally, since this in turn implies from (A3.18) that $\tau_{32} = \tau$, we see from (A3.15) that

$$(A3.21) \quad u_3 = \rho\tau u_2 + \varepsilon_3 .$$

Hence the given model is AR(1) up to $i = 3$. We now extend this argument by induction to all i . To do so, suppose that for some $k > 3$ it is true that

$$(A3.22) \quad u_i = \rho\tau u_{i-1} + \varepsilon_i , \quad i = 2, \dots, k .$$

Then it follows by successive substitution [starting with (A3.8)] that

$$(A3.23) \quad u_i = \sum_{j=1}^i (\rho\tau)^{i-j} \varepsilon_j , \quad i = 2, \dots, k .$$

Hence for $i = k + 1$ we must have

$$(A3.24) \quad u_{k+1} = \rho \sum_{i=1}^k \tau_{k+1,i} u_i + \varepsilon_{k+1} = \rho \sum_{i=1}^k \tau_{k+1,i} \left[\sum_{j=1}^i (\rho\tau)^{i-j} \varepsilon_j \right] + \varepsilon_{k+1}$$

$$= \rho \sum_{i=1}^k \sum_{j=1}^i [\tau_{k+1,i} (\rho\tau)^{i-j} \varepsilon_j] + \varepsilon_{k+1} .$$

But by letting $a_{ij} = \tau_{k+1,i} (\rho\tau)^{i-j} \varepsilon_j$, and using the summation identity,

$$(A3.25) \quad \sum_{i=1}^k \sum_{j=1}^i a_{ij} = \sum_{j \leq i} a_{ij} = \sum_{j=1}^k \sum_{i=j}^k a_{ij}$$

³ Here we use nonnegativity of τ -values. However, one could equally well assume $|\tau_j| \leq |\tau_{i,j-1}|$ in (A3.13), and the same argument would go through.

it follows that (A3.24) can be rewritten as

$$\begin{aligned}
\text{(A3.26)} \quad u_{k+1} &= \rho \sum_{j=1}^k \sum_{i=j}^k [\tau_{k+1,i}(\rho\tau)^{i-j} \varepsilon_j] + \varepsilon_{k+1} \\
&= \rho \sum_{j=1}^k \left[\sum_{i=j}^k \tau_{k+1,i}(\rho\tau)^{i-j} \right] \varepsilon_j + \varepsilon_{k+1} \\
&= \sum_{j=1}^k \left[\sum_{i=j}^k \tau_{k+1,i} \rho^{i+1-j} \tau^{i-j} \right] \varepsilon_j + \varepsilon_{k+1} \\
&= \left[\sum_{i=1}^k \tau_{k+1,i} \rho^i \tau^{i-1} \right] \varepsilon_1 + \sum_{j=2}^k \left[\sum_{i=j}^k \tau_{k+1,i} \rho^{i+1-j} \tau^{i-j} \right] \varepsilon_j + \varepsilon_{k+1} \\
\Rightarrow \text{var}(u_{k+1}) &= \left[\sum_{i=1}^k \tau_{k+1,i} \rho^i \tau^{i-1} \right]^2 \text{var}(\varepsilon_1) \\
&\quad + \sum_{j=2}^k \left[\sum_{i=j}^k \tau_{k+1,i} \rho^{i+1-j} \tau^{i-j} \right]^2 \text{var}(\varepsilon_j) + \text{var}(\varepsilon_{k+1}) \\
\Rightarrow v &= \left[\sum_{i=1}^k \tau_{k+1,i} \rho^i \tau^{i-1} \right]^2 v \\
&\quad + \sum_{j=2}^k \left[\sum_{i=j}^k \tau_{k+1,i} \rho^{i+1-j} \tau^{i-j} \right]^2 v(1-\rho^2\tau^2) + v(1-\rho^2\tau^2) \\
\Rightarrow 1 &= \left[\sum_{i=1}^k \tau_{k+1,i} \rho^i \tau^{i-1} \right]^2 + \left\{ 1 + \sum_{j=2}^k \left[\sum_{i=j}^k \tau_{k+1,i} \rho^{i+1-j} \tau^{i-j} \right]^2 \right\} (1-\rho^2\tau^2)
\end{aligned}$$

for all $\rho \in R$ [as in (A3.15)]. Hence differentiating with respect to ρ on $(-\rho_0, \rho_0)$, we now have

$$\begin{aligned}
\text{(A3.27)} \quad 0 &= 2 \left[\sum_{i=1}^k \tau_{k+1,i} \rho^i \tau^{i-1} \right] \left(\sum_{i=1}^k \tau_{k+1,i} (i) (\rho\tau)^{i-1} \right) \\
&\quad + \left\{ \sum_{j=2}^k 2 \left[\sum_{i=j}^k \tau_{k+1,i} \rho^{i+1-j} \tau^{i-j} \right] \left(\sum_{i=j}^k \tau_{k+1,i} (i+1-j) (\rho\tau)^{i-j} \right) \right\} (1-\rho^2\tau^2) \\
&\quad + \left\{ 1 + \sum_{j=2}^k \left[\sum_{i=j}^k \tau_{k+1,i} \rho^{i+1-j} \tau^{i-j} \right]^2 \right\} (-2\rho\tau^2)
\end{aligned}$$

for all $\rho \in (-\rho_0, \rho_0)$ [as in (A3.16)]. Again dividing by $\rho \in (0, \rho_0)$, we obtain

$$\text{(A3.28)} \quad 0 = 2 \left[\sum_{i=1}^k \tau_{k+1,i} (\rho\tau)^{i-1} \right] \left(\sum_{i=1}^k \tau_{k+1,i} (i) (\rho\tau)^{i-1} \right)$$

$$\begin{aligned}
& + \left\{ \sum_{j=2}^k 2 \left[\sum_{i=j}^k \tau_{k+1,i} (\rho\tau)^{i-j} \right] \left(\sum_{i=j}^k \tau_{k+1,i} (i+1-j) (\rho\tau)^{i-j} \right) \right\} (1 - \rho^2 \tau^2) \\
& - 2\tau^2 \left\{ 1 + \sum_{j=2}^k \left[\sum_{i=j}^k \tau_{k+1,i} \rho^{i+1-j} \tau^{i-j} \right]^2 \right\}
\end{aligned}$$

for all $\rho \in (0, \rho_0)$, so that by letting $\rho \downarrow 0$ we now have the following limiting relation at $\rho = 0$,

$$\begin{aligned}
\text{(A3.29)} \quad 0 &= 2(\tau_{k+1,1})(\tau_{k+1,1}) + 2 \sum_{j=2}^k (\tau_{k+1,j})(\tau_{k+1,j}) - 2\tau^2 \left\{ 1 + \sum_{i=2}^k (0) \right\} \\
&\Rightarrow \tau^2 = \sum_{j=1}^k \tau_{k+1,j}^2
\end{aligned}$$

which is seen to be a direct generalization of (A3.18). Proceeding in a similar manner, we differentiate (A3.28) once again to obtain

$$\begin{aligned}
\text{(A3.30)} \quad 0 &= 2 \left[\sum_{i=2}^k \tau_{k+1,i} (i-1) \rho^{i-2} \tau^{i-1} \right] \left(\sum_{i=1}^k \tau_{k+1,i} (i) (\rho\tau)^{i-1} \right) \\
&+ 2 \left[\sum_{i=1}^k \tau_{k+1,i} (\rho\tau)^{i-1} \right] \left(\sum_{i=2}^k \tau_{k+1,i} (i) (i-1) \rho^{i-2} \tau^{i-1} \right) \\
&- 2\rho\tau^2 \left\{ \sum_{j=2}^k 2 \left[\sum_{i=j}^k \tau_{k+1,i} (\rho\tau)^{i-j} \right] \left(\sum_{i=j}^k \tau_{k+1,i} (i+1-j) (\rho\tau)^{i-j} \right) \right\} \\
&+ (1 - \rho^2 \tau^2) \cdot \\
&\left\{ \sum_{j=2}^k 2 \left[\sum_{i=j+1}^k \tau_{k+1,i} (i-j) \rho^{i-j-1} \tau^{i-j} \right] \left(\sum_{i=j}^k \tau_{k+1,i} (i+1-j) (\rho\tau)^{i-j} \right) \right. \\
&\quad \left. + \sum_{j=2}^k 2 \left[\sum_{i=j}^k \tau_{k+1,i} (\rho\tau)^{i-j} \right] \left(\sum_{i=j+1}^k \tau_{k+1,i} (i+1-j) (i-j) \rho^{i-j-1} \tau^{i-j} \right) \right\} \\
&- 4\tau^2 \sum_{j=2}^k \left[\sum_{i=j}^k \tau_{k+1,i} \rho^{i+1-j} \tau^{i-j} \right] \left(\sum_{i=j}^k \tau_{k+1,i} (i+1-j) (\rho\tau)^{i-j} \right)
\end{aligned}$$

for $\rho \in (0, \rho_0)$. Finally, dividing by 2, and letting $\rho \downarrow 0$ one last time, we obtain the following additional relation at $\rho = 0$,

$$\text{(A3.31)} \quad 0 = (\tau_{k+1,2} \tau)(\tau_{k+1,1}) + (\tau_{k+1,1})[\tau_{k+1,2}(2)\tau] - (0)$$

$$\begin{aligned}
& + (1) \left\{ \sum_{j=2}^k [\tau_{k+1,j+1} \tau] [\tau_{k+1,j} (2)] \right. \\
& \quad \left. + \sum_{j=2}^k [\tau_{k+1,j}] (\tau_{k+1,j+1} (2) \tau) \right\} \\
& - 2\tau^2 \sum_{j=2}^k (0) \\
& = \tau(\tau_{k+1,2} \tau_{k+1,1}) + 2\tau(\tau_{k+1,2} \tau_{k+1,1}) \\
& \quad + \tau \sum_{j=2}^k \tau_{k+1,j+1} \tau_{k+1,j} + 2\tau \sum_{j=2}^k \tau_{k+1,j+1} \tau_{k+1,j} \\
& = 3\tau \sum_{j=1}^k \tau_{k+1,j+1} \tau_{k+1,j} \\
& \Rightarrow 0 = \tau \sum_{j=1}^k \tau_{k+1,j+1} \tau_{k+1,j}
\end{aligned}$$

which is seen to be a generalization of (A3.20). But the positivity of τ and the nonnegativity of all τ_{ij} then implies that $\tau_{k+1,j+1} \tau_{k+1,j} = 0$ for all $j = 1, \dots, k$, which together with the inequalities, $0 \leq \tau_{k+1,j} \leq \tau_{k+1,j+1}$ for all $j = 1, \dots, k-1$ yields the conclusion that

$$(A3.32) \quad \tau_{k+1,j} = 0, \quad j = 1, \dots, k-1.$$

Moreover, since (A3.32) together with (A3.29) implies that $\tau = \tau_{k+1,k}$, we must have

$$(A3.33) \quad u_{k+1} = \rho \tau u_k + \varepsilon_{k+1}.$$

Hence it follows by induction that (A3.22) must hold for all $i = 2, \dots, n$, so that the given model is AR(1). ■