Adaptations of Conventional Spatial Econometric Models to Count Data

Kurt Brännäs

Department of Economics
Umeå School of Business and Economics, Umeå University
email: kurt.brannas@econ.umu.se

Umeå Economic Studies 883, 2014

Abstract

The paper suggests and studies count data models corresponding to previously studied spatial econometric models for continuous variables. A novel way of incorporating spatial weights is considered for both time and space dynamic models with or without simultaneity. The paper also contains a brief discussion about estimation issues.

Key Words – Integer-valued, Space, Time, Regional, Thinning, Estimation
JEL – C31, C32, C51, R12, R15, R23
MSC2010 – 62M10, 62M20, 62M30, 62P15, 62P20

1 Introduction

Count data are increasingly often found useful for empirical studies in many fields of economics. In regional economic settings with small areas and/or when counts (or frequencies) for other reasons are small it is particularly important to account for some of the key features of count data. Notably, counts are integer-valued, non-negative and in most models for counts, heteroskedasticity is an important feature.

In this paper we depart from some widely used linear spatial econometric models (e.g., Anselin, 1988; Anselin, Florax and Rey, 2004) and introduce and discuss specifications of spatial econometric models that account for count data features. The emphasis is on models that exhibit either or both time and spatial autoregressive lags.

Poisson and negative binomial regressions are leading examples of models that account for integer-valued, non-negative and heteroskedastic count data (e.g., Cameron and Trivedi, 1998; Winkelmann, 2008). The latter regression accounts for the empirically frequently found over-dispersion, i.e., that the sample variance is larger than the sample mean. The regressions contain observed heterogeneity possibly both in terms of current and lagged exogenous variables as well as spatial factors. Count data models may also be specified to account for unobserved heterogeneity to reflect any time and space dependencies (e.g., Zeger, 1988; Zhang, 2002; Sengupta and Cressie, 2013).

Another much studied count data model class stems from the independent works of McKenzie (1985) and Al-Osh and Alzaid (1987). They introduced and studied the integer-valued autoregressive model of order one (INAR(1)). A survey of the early literature offering, e.g., various extensions is given by McKenzie (2003) and partial textbook treatments are offered in, e.g., Cameron and Trivedi (1998, 2005). The INAR(1) model is written in the manner of a conventional autoregressive model of order one, except that a thinning operation here replaces multiplying the lagged endogenous variable by a parameter, see below. Otherwise integer-valued counts cannot be guaranteed. Still, INAR models share some basic properties with the conventional linear time series models.

In this paper a multivariate INAR(1) model (e.g., McKenzie, 1988; Brännäs, 1995; Berglund and Brännäs, 1996; Pedeli and Karlis, 2013) serves as a platform for developing time dynamic model extensions appropriate for spatial count data. We view the spatial configuration as given and constant across time. The incorporation of spatial effects through a weight matrix necessitates a novel treatment and it is to be done through the model parameters. We emphasise model characteristics and discuss some model properties in terms of low order moments. For robustness and technical reasons full distributional results are not given and therefore least squares and related estimators are briefly discussed but not the maximum likelihood estimator.

Section 2 develops the count data based spatial econometric models and gives some of their properties. In Section 3 we discuss approaches to the estimation of the unknown model parameters. In Appendix A a new approach to obtaining inversion results for thinning operations is introduced.

2 Model Specifications

Count data have some particular features that need to be recognised for the coherency between data generating processes (DGP) and spatial econometric models. Counts are obviously integer-valued and greater than or equal to zero. For large counts frequent use is made of normal approximations and then conventional models may be directly adopted. For smaller counts this may be a risky path to pursue as, e.g., forecasts may come out with an incorrect sign. In addition, by recognising key features of the DGP interpretational benefits may be brought to the empirical modelling exercise.

We start by giving some key results for the basic multivariate count data AR(1) model, before introducing spatial effects and exogenous variables in this setup. Later we consider the simultaneous and autoregressive equations model for count data as well as discuss some of its special cases.

2.1 The Multivariate AR(1)

The first order and M-variate count (integer-valued) data autoregressive model of order one (INAR(1)) can be written as

$$\mathbf{y}_t = \mathbf{A} \circ \mathbf{y}_{t-1} + \epsilon_t, \ \ t = 2, \ldots, T. \tag{1}$$

The count data AR(1) model in (1) is in the spatial context seen as having the elements of the $\mathbf{y}_t = (y_{1t}, \ldots, y_{Mt})'$ vector represent the same basic variable but with measurements representing the M different spatial units, such as municipalities or regions. The $M \times M$ matrix $\mathbf{A}$ has elements α_{ij}, and the symbol $\circ$ represents binomial thinning which replaces standard multiplication in order for the model to generate integer-valued outcomes. For instance, for a scalar integer-valued random y variable the thinning operation is defined as $\alpha \circ y = \sum_{i=1}^{y} u_i$, where $\{u_i\}_{i=1}^{y}$ is an iid sequence of $0-1$ random variables and $\Pr(u_i) = \alpha$. It follows that the integer-valued $\alpha \circ y \in [0, y]$ and that for a given y, $\alpha \circ y$ is binomially distributed with conditional mean αy and conditional variance $\alpha(1-\alpha)y$. This motivates the label binomial thinning. A few useful results for binomial thinning operations are given in Appendix A.

Hence, the parameters in the $\mathbf{A}$ matrix are interpreted as probabilities, so that $\alpha_{ij} \in [0, 1]$, for all relevant i, j. Thinning operations are performed element by element, such that for the ith equation we get from (1)

$$y_{it} = \sum_{j=1}^{M} \alpha_{ij} \circ y_{j,t-1} + \epsilon_{it}. \tag{2}$$

Here, the different thinning operations are assumed independent and independent of the disturbance term $\boldsymbol{\epsilon}_t$, for all t.[1] For the unobservable count data random $\boldsymbol{\epsilon}_t$ vector we have that $\boldsymbol{\epsilon}_t \geq 0$ and we assume that $E(\boldsymbol{\epsilon}_t) = \boldsymbol{\lambda} > \mathbf{0}$ and $E(\boldsymbol{\epsilon}_t \boldsymbol{\epsilon}_s') = \boldsymbol{\Sigma}$, for $t = s$, and equal to $\mathbf{0}$ when $t \neq s$. The $\boldsymbol{\epsilon}_t$ sequence is throughout assumed serially uncorrelated. In the important and parsimoniously parameterised special case of independently Poisson distributed $\epsilon_{it}, i = 1, \ldots, M$, we have $\text{diag}(\boldsymbol{\Sigma}) = \boldsymbol{\lambda}$ and zeroes elsewhere in $\boldsymbol{\Sigma}$. The Poisson case is an example of self-decomposability (Steutel and van Harn, 1979), while, e.g., the binomial is not self-decomposable.

Consider as an example, the population sizes of, for instance, individuals or firms, in the M regions that constitute the regional $\mathbf{y}_t$ vector. Then the diagonal α_{ii} elements reflect survival probabilities in the regions, and hence $1 - \alpha_{ii}$ is the probability of emigration from the ith region. An off-diagonal α_{ij} element corresponds to a migration

[1] Brännäs and Hellström (2001) consider the consequences of relaxing such independence assumptions in the univariate INAR(1) model. Most often only second and higher order moments will change when such assumptions are varied.

probability from a region j to region i. Births and immigration from any other (outside) regions are caught by the random ϵ_{it} term. If deaths are to be included, as say, an artificial region $M+1$ then $\alpha_{i,M+1} = 0$ while $\alpha_{M+1,j} \geq 0$, for all i and j. In fact, the model corresponds to the Markov model for open systems of Bartholomew (1982, ch 3). In this example it may be natural to enforce a column sum condition, i.e. $\sum_{j=1}^{M} \alpha_{ij} = 1$ and $\sum_{j=1}^{M+1} \alpha_{ij} = 1$ if deaths are included. In other contexts such as in studying the number of births across regions, row restrictions of this type appears less natural. Interpretations of the type illustrated here are not automatically provided by conventional VAR modelling exercises.

For stationarity we require that the largest eigenvalue of the positive $\mathbf{A}$ matrix is smaller than one. If all off-diagonal elements $\alpha_{ij} = 0$ then the model in (1) is stationary if $\max_i (\alpha_{ii}) < 1$. Note also that for this count data model specification with $\epsilon_{it} \geq 0$, non-stationarity can be rejected if for any time series i there is, at least, one negative change, i.e., $y_{it} - y_{it-1} < 0$, for all t. Obviously, stationarity only arises if there is a long history preceding the first observation at time $t = 1$.

Given stationarity the first two conditional and unconditional moments can be obtained as

$$E(\mathbf{y}_t | Y_{t-1}) = \mathbf{A}\mathbf{y}_{t-1} + \boldsymbol{\lambda} \tag{3}$$

$$E(\mathbf{y}_t) = (\mathbf{I} - \mathbf{A})^{-1}\boldsymbol{\lambda} \tag{4}$$

$$V(\mathbf{y}_t | Y_{t-1}) = \mathbf{H}_{t-1} + \boldsymbol{\Sigma} \tag{5}$$

$$V(\mathbf{y}_t) = \mathbf{A}V(\mathbf{y}_t)\mathbf{A}' + \mathbf{H}^* + \boldsymbol{\Sigma}, \tag{6}$$

where Y_{t-1} is the history of $\mathbf{y}_t$ up through time $t-1$. The $\mathbf{H}_{t-1}$ matrix is diagonal with $\operatorname{diag} \mathbf{H}_{t-1} = (\sum_{j=1}^{M} \alpha_{ij}(1-\alpha_{ij}) y_{jt-1}, i = 1, \ldots, M)$. Therefore, the diagonal matrix $\mathbf{H}^*$ has $\operatorname{diag} \mathbf{H}^* = (\sum_{j=1}^{M} \alpha_{ij}(1-\alpha_{ij}) E(y_{jt}), i = 1, \ldots, M)$ and it is time invariant. The diagonal elements of $\mathbf{H}_{t-1}$ in (5) highlights the autoregressive conditional heteroskedasticity (ARCH) property of the multivariate count data AR(1) model.

Various special cases of (1) have previously been considered in the literature. When $\mathbf{A} = \alpha\mathbf{I}$, i.e. the matrix is diagonal with a scalar parameter, and there is no dependence between the elements in the $\boldsymbol{\epsilon}_t$ vector and all elements have the same first two moments, so that, e.g., $\boldsymbol{\lambda} = \lambda_0 \mathbf{1}_M$, with λ_0 an unknown scalar parameter and $\mathbf{1}_M = (1, \ldots, 1)'$, and

$\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_M$, the model simplifies to a replicated INAR(1) (e.g., Silva, 2005). With also $\mathbf{A} = \mathbf{0}$ we then simply have M independent variables. Panel data applications of the full model with random effects are discussed by Brännäs (1995), Blundell, Griffith and Windmeijer (2002) and others.

In a case of a small number of spatial units M but with long time series, i.e. T is large, the off-diagonal elements of the $\mathbf{A}$ matrix can empirically be estimated and interpreted as representing spatial effects. For the more conventional situation of a large M and a small T we note that catching spatial effects may be empirically difficult due to the short time series. Specifying a model for α_{ij} in terms of exogenous variables and/or spatial effects may reduce the number of unknown parameters and make estimation feasible.

The model in (1) may be extended in, at least, two important directions. It is obviously possible to incorporate higher order lags of $\mathbf{y}_t$ as well as to incorporate simultaneous effects (cf. Brännäs, 2013). Importantly, exogenous variables potentially reflecting spatial effects and with or without lags, may be included through the parameters of the model. For α_{ij} we may adopt, say, a logistic specification (cf. Brännäs, 1995) to get

$$\alpha_{ij,t} = 1/(1 + \exp(\mathbf{x}'_{ij,t}\boldsymbol{\theta}_\alpha)) \in [0, 1], \tag{7}$$

with $\boldsymbol{\theta}_\alpha$ a k_α dimensional parameter vector. The exogenous variables are collected into the vector $\mathbf{x}_{ij,t}$. Since $\alpha_{ij,t}$ reflects a transition in the time interval $(t-1, t]$, $\mathbf{x}_{ij,t}$ is most often best taken to reflect what happened in previous time intervals, i.e. $(t-2, t-1]$ and before. For the $\boldsymbol{\lambda}$ vector we may let the elements be of exponential forms, so that $\lambda_{it} = \exp(\mathbf{z}'_{it}\boldsymbol{\theta}_\lambda) > 0$, with $\boldsymbol{\theta}_\lambda$ a $k_\lambda \times 1$ parameter vector.

Corresponding to (1) we may then write the model as

$$\mathbf{y}_t = \mathbf{A}_t \circ \mathbf{y}_{t-1} + \boldsymbol{\epsilon}_t, \; t = 2, \ldots, T, \tag{8}$$

where $E(\boldsymbol{\epsilon}_t) = \boldsymbol{\lambda}_t$. Such specifications may well reduce the number of unknown parameters in $\mathbf{A}$ and $\boldsymbol{\lambda}$ rather than to increase the number. Other ways of incorporating exogenous variables cannot in a general way guarantee a DGP that gives an endogenous and integer-valued $\mathbf{y}_t$ variable vector.

2.2 The Spatial AR(1)

In the spatial econometrics literature (e.g., Anselin, Florax and Rey, 2004) the spatial distance between observable units is an important ingredient. In the current model framework the spatial effects are best implicitly incorporated through the $\mathbf{A}$ matrix and the $\boldsymbol{\lambda}$ vector.

The spatial economics literature makes frequent use of a weight matrix $\mathbf{W}$, that may be time-varying, to reflect spatial distance and it can, e.g., have elements in the form of a simple gravity model $\mathbf{W}_{ij} = M_i M_j / D_{ij}$ or the even simpler $\mathbf{W}_{ij} = 1/D_{ij}^2$, for $i \neq j$ and $\mathbf{W}_{ii} = 0$, for all i. Here, M_i represents a measure of mass for unit i, and D_{ij} is a measure of distance between units i and j. The M_i and M_j may be measured by wealth or some other economic size variable and will therefore likely be time-varying, implying time-dependence also in $\mathbf{W}_{ij,t}$. As the distance increases $\mathbf{W}_{ij}$ will typically become smaller and this then implies a smaller spatial correlation.

Since, the $\mathbf{A}$ matrix in (1) contains probabilities a logistic specification may be usefully applied, cf. (7). Thanks to symmetry

$$\alpha_{ij} = 1/(1+\exp(\alpha_0 + \alpha_1 \mathbf{W}_{ij})) = \alpha_{ji}, \;\; i \neq j \tag{9}$$

is true also in the presence of time-dependence. The effect of this simple parametrisation is to reduce M^2 potentially unknown α_{ij} to two unknown parameters, α_0 and α_1. When $\alpha_1 = 0$ there is no spatial effect. If $\mathbf{W}$ is of some general form that contains unknown parameters we could instead write $\alpha_{ij} = 1/(1+\exp(\alpha_0 + \mathbf{W}_{ij}(\boldsymbol{\alpha}_1)))$. If $\mathbf{W}$ by an appropriate restriction on the parameter vector $\boldsymbol{\alpha}_1$ has no impact on α_{ij} it implies that $\alpha_{ii} = 1/(1+\exp(\alpha_0))$ is constant across i and j, and then indicative of a constant lag one time dependence in $y_{it}, i = 1, \ldots, M$. If we were to also include explanatory variables in α_{ij} the symmetry is likely to be lost. Obviously, the $\boldsymbol{\lambda}$ vector may also be specified to reflect spatial effects in some analogous manner.

A direct use of the more conventional spatial econometrics analogues, such as, $\mathbf{AW} \circ \mathbf{y}$ or $\mathbf{A} \circ \mathbf{Wy}$ are less suitable than the illustrated $\mathbf{A}(\mathbf{W}) \circ \mathbf{y}$ specification if we wish to adhere to the count data interpretation of $\mathbf{y}_t$. The reasons are that the elements of $\mathbf{AW}$ are not necessarily in unit intervals as required for probabilities, and $\mathbf{Wy}$ is not likely to have integer-valued elements.

Hence, we may write the preferred model representation as

$$y_{it} = \alpha_{ii} \circ y_{it-1} + \sum_{j=1, i \neq j}^{M} \alpha_{ij}(\mathbf{W}) \circ y_{j,t-1} + \epsilon_{it} = \sum_{j=1}^{M} \alpha_{ij}(\mathbf{W}) \circ y_{j,t-1} + \epsilon_{it}$$

or compactly as

$$\mathbf{y}_t = \mathbf{A}(\mathbf{W}) \circ \mathbf{y}_{t-1} + \boldsymbol{\epsilon}_t. \tag{10}$$

For small numbers of spatial units or when the spatial dependence is limited to the nearest neighbours it is not always necessary to include distance explicitly, but to instead use the α_{ij} parameters directly. For instance, Brännäs and Brännäs (1998) used a binomial INAR(1) model for fish visits in a closed experimental tank system, and Boudreault and Charpentier (2011) studied earthquake counts. Ghodsi, Shitan and Bakouch (2012) studied the moment properties of a space-time INAR model of order (1,1).

2.3 A Spatial Simultaneous Equations Model

A structural form of a simultaneous count data autoregressive model of order one can be written (cf. Brännäs, 2013) as an extension of the model in (1), i.e.

$$\mathbf{y}_t = \mathbf{A}_0 \circ \mathbf{y}_t + \mathbf{A}_1 \circ \mathbf{y}_{t-1} + \boldsymbol{\epsilon}_t, \; t = 2, \ldots, T, \tag{11}$$

where the $M \times M$ matrix $\mathbf{A}_0$ is of the general form

$$\mathbf{A}_0 = \begin{pmatrix} 0 & \alpha_{12}^0 & \alpha_{13}^0 & \cdots & \alpha_{1M}^0 \\ \alpha_{21}^0 & 0 & \alpha_{23}^0 & \cdots & \alpha_{2M}^0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & \alpha_{M-1,M}^0 \\ \alpha_{M1}^0 & \cdots & \cdots & \alpha_{M,M-1}^0 & 0 \end{pmatrix}.$$

The endogenous $\mathbf{y}_t = (y_{1t}, \ldots, y_{Mt})'$ vector and its lags are all integer-valued. The model contains simultaneity or interdependence across these y_{it} variables as reflected by the non-zero off-diagonal elements in the $\mathbf{A}_0$ matrix. The simultaneity is here and elsewhere seen as a consequence of a low sampling frequency. The parameters in the $\mathbf{A}_0$ and $\mathbf{A}_1$ matrices are interpreted as probabilities.

By this specification there can only be contemporaneous positive or no effects at all between the $y_{it}, i \ldots, M$ variables for the given specification of $\mathbf{A}_0$. This is beneficial in guaranteeing that $y_{it} \geq 0$, for all i and t. Even if this condition is to hold true we may still account for some smallish negative effects by using a minus sign for some of the α_{ij}^0 in $\mathbf{A}_0$. In such cases thinning is to be interpreted as $-(\alpha_{ij}^0 \circ y_{jt})$. Related to this representation, we may define $\mathbf{A}_* = \mathbf{I}_M - \mathbf{A}_0$, with $\mathbf{I}_M$ the $M \times M$ identity matrix. The model can then be written as

$$\mathbf{A}_* \circ \mathbf{y}_t = \mathbf{A}_1 \circ \mathbf{y}_{t-1} + \boldsymbol{\epsilon}_t. \tag{12}$$

This form reveals the closeness to structural VAR models.

With up to $2 \cdot M^2 - M$ potential parameters in the $\mathbf{A}_0$ and $\mathbf{A}_1$ matrices the general model in (11) is likely too rich in parameters for most practical purposes unless some additional restrictions are enforced beyond the zeroes of the diagonal of $\mathbf{A}_0$. These zeroes correspond to the normalisation convention (cf. the ones in the $\mathbf{A}_*$ matrix). Also note, that $\boldsymbol{\lambda}$ and $\boldsymbol{\Sigma}$, bring along $M + M(M+1)/2$ additional and potentially free parameters.

The specification that comes closest to the classical, static simultaneous equations model has $\mathbf{A}_1 = \mathbf{0}$, $\mathbf{A}_0$ time invariant, and with exogenous variables included only through $\boldsymbol{\lambda}_t$, i.e.

$$\mathbf{A}_* \circ \mathbf{y}_t = \boldsymbol{\lambda}_t + \boldsymbol{\epsilon}_t^*, \tag{13}$$

where $\boldsymbol{\epsilon}_t^* = \boldsymbol{\epsilon}_t - \boldsymbol{\lambda}_t$. If $\mathbf{y}_t$ takes on large numbers, $\boldsymbol{\lambda}_t$ can potentially be specified as linear in the exogenous variables vector $\mathbf{z}_t$ without violating the non-negativity constraint of $\mathbf{y}_t$. The general simultaneous equations model with time dependent parameters based on exogenous variables and the weight matrix $\mathbf{W}$, which may be time-varying, is written

$$\mathbf{y}_t = \mathbf{A}_{0t} \circ \mathbf{y}_t + \mathbf{A}_{1t} \circ \mathbf{y}_{t-1} + \boldsymbol{\lambda}_t + \boldsymbol{\epsilon}_t^*. \tag{14}$$

Brännäs (2013) gives some moment properties for the structural form representation in (11)-(12). For these models the literature does not offer any general results for the direct inversion or division of thinning operations $u = \theta \circ v$ that would result in some practical form of thickening or expansion operation, say, $v = \theta^* \odot u$, and much less so for the matrix case involved in obtaining general types of reduced forms. Hence, giving general reduced form results with explicit distributional properties and functional expressions for the $\mathbf{y}_t$ vector is difficult. One result related to this problem is due to Littlejohn (1994), but it appears difficult to handle in general setups. Some new results for the inversion in terms of moments are relatively easy to obtain and they are given in Appendix A. These results are supportive of the results obtained by a different approach below. While different, the current specification shares the feature of a non-exact relationship between distributions for general nonlinear simultaneous equation models and their reduced forms.

We start by conditioning the model in (11) on past observations Y_{t-1} and on the set of exogenous variables to get

$$E(\mathbf{y}_t|Y_{t-1}) = \mathbf{A}_{0t}E(\mathbf{y}_t|Y_{t-1}) + \mathbf{A}_{1t}\mathbf{y}_{t-1} + \boldsymbol{\lambda}.$$

The matrix $\mathbf{A}_{*t} = \mathbf{I} - \mathbf{A}_{0t}$ is assumed invertible in this complete system, so that the conditional expectation is

$$E(\mathbf{y}_t|Y_{t-1}) = \mathbf{A}_{*t}^{-1}\mathbf{A}_{1t}\mathbf{y}_{t-1} + \mathbf{A}_{*t}^{-1}\boldsymbol{\lambda}_t = \mathbf{C}_t\mathbf{y}_{t-1} + \boldsymbol{\lambda}_t^*.$$

This is the key part of the reduced form and a full reduced form can now be written

$$\mathbf{y}_t = E(\mathbf{y}_t|Y_{t-1}) + \boldsymbol{\xi}_t = \mathbf{C}_t\mathbf{y}_{t-1} + \boldsymbol{\lambda}_t^* + \boldsymbol{\xi}_t, \tag{15}$$

where $E(\boldsymbol{\xi}_t|Y_{t-1}) = \mathbf{0}$. This way of writing the model is useful for model analysis, though it is not automatically useful as a description of the data generating process. There is no guarantee that integer-valued $\mathbf{y}_t$ can be generated unless the distribution of $\boldsymbol{\xi}_t$ can be ascertained. The structural form in (11)-(12) is mostly seen as the ideal interpretational description of the data generating process with its explicit direct and indirect effect interpretations of the parameterisation. The reduced form (15) only gives total effects, but it is the cornerstone for, e.g., distributional properties and forecasting.

The model in (15) is of a VAR(1) form, which makes model based analysis by analogy to the VAR literature straightforward. To obtain the corresponding conditional variance $V(\mathbf{y}_t|Y_{t-1})$ we rewrite the structural form in (14) as:

$$\begin{aligned}\mathbf{y}_t &= \mathbf{A}_{0t}\mathbf{y}_t + \mathbf{A}_{1t}\mathbf{y}_{t-1} + \boldsymbol{\lambda}_t + \boldsymbol{\epsilon}_t^* + (\mathbf{A}_{0t}\circ\mathbf{y}_t - \mathbf{A}_{0t}\mathbf{y}_t) + (\mathbf{A}_{1t}\circ\mathbf{y}_{t-1} - \mathbf{A}_{1t}\mathbf{y}_{t-1}) \\ &= \mathbf{A}_{0t}\mathbf{y}_t + \mathbf{A}_{1t}\mathbf{y}_{t-1} + \boldsymbol{\lambda}_t + \boldsymbol{\epsilon}_t^{**}, \end{aligned} \tag{16}$$

where the composite disturbance terms $\boldsymbol{\epsilon}_t^* = \boldsymbol{\epsilon}_t - \boldsymbol{\lambda}_t$ and $\boldsymbol{\epsilon}_t^{**}$ both have zero means, but obviously the latter contains both current values and lags of $\mathbf{y}_t$. The corresponding reduced form is identical to the one in (15) with $\boldsymbol{\xi}_t = (\mathbf{I} - \mathbf{A}_{0t})^{-1}\boldsymbol{\epsilon}_t^{**} = \mathbf{A}_{*t}^{-1}\boldsymbol{\epsilon}_t^{**}$. The important ingredient for the conditional variance is the one-step-ahead prediction error $\tilde{\mathbf{y}}_t = \mathbf{y}_t - E(\mathbf{y}_t|Y_{t-1})$. We get

$$\begin{aligned}\tilde{\mathbf{y}}_t &= \mathbf{A}_{0t}\mathbf{y}_t + \mathbf{A}_{1t}\mathbf{y}_{t-1} + \boldsymbol{\lambda}_t + \boldsymbol{\epsilon}_t^{**} - \mathbf{A}_{0t}E(\mathbf{y}_t|Y_{t-1}) - \mathbf{A}_{1t}\mathbf{y}_{t-1} - \boldsymbol{\lambda}_t, \\ &= \mathbf{A}_{0t}\tilde{\mathbf{y}}_t + \boldsymbol{\epsilon}_t^{**} = \mathbf{A}_{*t}^{-1}\boldsymbol{\epsilon}_t^{**}\end{aligned}$$

and therefore we get the conditional variance expression

$$V(\mathbf{y}_t|Y_{t-1}) = E(\tilde{\mathbf{y}}_t\tilde{\mathbf{y}}_t'|Y_{t-1}) = \mathbf{A}_{*t}^{-1}E(\boldsymbol{\epsilon}_t^{**}\boldsymbol{\epsilon}_t'^{**}|Y_{t-1})(\mathbf{A}_{*t}')^{-1}, \tag{17}$$

where

$$\begin{aligned}E(\boldsymbol{\epsilon}_t^{**}\boldsymbol{\epsilon}_t'^{**}|Y_{t-1}) &= \boldsymbol{\Theta}_{0,t-1} - \boldsymbol{\Theta}_{1,t-1} - \boldsymbol{\Sigma} + [E(\boldsymbol{\epsilon}_t\mathbf{y}_t'|Y_{t-1}) - \boldsymbol{\lambda}_t[E(\mathbf{y}_t'|Y_{t-1})](\mathbf{I} - \mathbf{A}_{0t})' \\ &\quad +(\mathbf{I} - \mathbf{A}_{0t})[E(\mathbf{y}_t\boldsymbol{\epsilon}_t'|Y_{t-1}) - E(\mathbf{y}_t|Y_{t-1})\boldsymbol{\lambda}_t'].\end{aligned}$$

Here, $\boldsymbol{\Theta}_{0,t-1}$ is a diagonal matrix with diagonal elements $\sum_{j=1}^{M} \alpha_{ij}^0(1 - \alpha_{ij}^0)E(y_{jt}|Y_{t-1})$, for $i = 1, \ldots, M$, and $\mathrm{diag}(\boldsymbol{\Theta}_{1,t-1}) = (\sum_{j=1}^{M} \alpha_{ij}^1(1 - \alpha_{ij}^1)y_{j,t-1}, i = 1, \ldots, M)$. The latter matrix corresponds to the $\mathbf{H}_{t-1}$ in (5). In these expressions, $E(\mathbf{y}_t|Y_{t-1})$ is given in (15), and $E(\boldsymbol{\epsilon}_t\mathbf{y}_t'|Y_{t-1}) = \mathbf{A}_{1t}\mathbf{y}_{t-1}\boldsymbol{\lambda}_t' + \boldsymbol{\Sigma} - (\mathbf{I} - \mathbf{A}_{0t})E(\mathbf{y}_t|Y_{t-1})\boldsymbol{\lambda}_t'$ from which its transpose can also be obtained. Derivations for the different parts are given in Appendix B.

2.4 Static Spatial Model

The classical econometric approach to incorporating spatial correlation into a static regression model is through the disturbance term in $y_{it} = \lambda_{it} + \epsilon_{it} + u_i$, where λ_{it} is a linear or, say, an exponential function of exogenous variables and possibly of spatial effects, and u_i is a random spatial effect which is taken to be iid and with mean μ. The spatial correlation in the random error term arises through the model $\epsilon_{it} = \gamma\sum_{j=1}^{M}\mathbf{W}_{ij}\epsilon_{jt} + v_t$ (cf. Anselin, 1988). For all spatial units at time t the M-variate error is written as $\boldsymbol{\epsilon}_t = \gamma\mathbf{W}\boldsymbol{\epsilon}_t + \mathbf{v}_t$ and with $\mathbf{v}_t = v_t\mathbf{1}$.

One count data parametrisation is to use $\mathbf{A}(\mathbf{W})$ as in (10) and to write the count data analogue corresponding to the ϵ_{it} part above as

$$\boldsymbol{\epsilon}_t = \boldsymbol{\Gamma}(\mathbf{W}) \circ \boldsymbol{\epsilon}_t + \mathbf{v}_t, \tag{18}$$

where, e.g., $\boldsymbol{\Gamma}(\mathbf{W})_{ij} = 1/(1+\exp(\gamma_0+\gamma\mathbf{W}_{ij}))$, for $i \neq j$, and equal to zero for the diagonal elements. Conditioning on previous $\mathbf{y}_{t-k}, k \geq 1$, we get $E(\boldsymbol{\epsilon}_t|Y_{t-1}) = [\mathbf{I} - \boldsymbol{\Gamma}(\mathbf{W})]^{-1}\boldsymbol{\kappa}_v$, where $\boldsymbol{\kappa}_v = E(\mathbf{v}_t)$. Then the model can using $\mathbf{y}_t = E(\mathbf{y}_t|Y_{t-1}) + \boldsymbol{\xi}_t$ with $E(\boldsymbol{\xi}_t|Y_{t-1}) = \mathbf{0}$ be written

$$\mathbf{y}_t = \boldsymbol{\lambda}_t + [\mathbf{I} - \boldsymbol{\Gamma}(\mathbf{W})]^{-1}\boldsymbol{\kappa}_v + \boldsymbol{\mu} + \boldsymbol{\xi}_t, \tag{19}$$

where $\boldsymbol{\mu} = E(\mathbf{u}|Y_{t-1}) = \mu\mathbf{1}$.

An alternative is to proceed as in the classical count data regression approach (e.g., Cameron and Trivedi, 1998; Winkelmann, 2008). In this multivariate context we may start from $E(y_{it}|\boldsymbol{\epsilon}_t) = \epsilon_{it}\lambda_{it}$ to get $E(y_{it}) = E(\epsilon_{it})\lambda_{it} = \lambda_{it}$ when λ_{it} contains a constant term. The variance is typically of the form $V(y_{it}) = \lambda_{it} + \sigma_i^2\lambda_{it}^2$. The general specification of Brännäs and Johansson (1996) that accounts for dependence across time and spatial units could be used as a platform to explicitly include spatial effects.

3 Remarks on Estimation

The presence of the $\mathbf{y}_t$ variables in the right hand side of (11) or (14) implies a dependence with the disturbance term $\boldsymbol{\epsilon}_t$. This dependence renders, e.g., the ordinary (conditional) least squares estimator inconsistent. Such a dependence is not present in the AR(1) representations (1) or (8). Section 2.3 indicated that obtaining a distributionally well-defined reduced form is nontrivial in general. For that reason maximum likelihood estimation appears to be beyond reach in most cases, and it will not be discussed in this introductory account on estimation.

3.1 AR(1) Models

The focus is first on directly estimating the unknown parameters of the autoregressive models (1) and (8). The presence of the thinning operations seemingly complicates a direct use of consistent estimation approaches such as conventional conditional or ordinary least squares (OLS), instrumental variable (IV) or generalised method of moments (GMM). However, the operators disappear when we consider the prediction error $\mathbf{e}_t = \mathbf{y}_t - E(\mathbf{y}_t|Y_{t-1}) = \mathbf{y}_t - \mathbf{A}_t\mathbf{y}_{t-1} - \boldsymbol{\lambda}_t$. The estimators are therefore best viewed as minimising sums of squares of prediction errors. We consider both single equation as well as joint estimation of the full system.

The single equation OLS estimator for spatial unit i and for constant parameters uses (2) to obtain the prediction error $e_{it} = y_{it} - \mathbf{y}'_{t-1}\boldsymbol{\alpha}_{i.} - \lambda_i$, where $\boldsymbol{\alpha}_{i.} = (\alpha_{i1}, \ldots, \alpha_{iM})'$ is the transpose of the ith row of the $\mathbf{A}$ matrix. For all time series observations for spatial unit i we may write $\mathbf{e}_i = \mathbf{y}_i - \mathbf{y}_{(-1)}\boldsymbol{\alpha}_{i.} - \lambda_i \mathbf{1}$, where $\mathbf{y}_i = (y_{i2}, \ldots, y_{iT})'$ and $\mathbf{y}_{(-1)} = (\mathbf{y}_{1(-1)}, \ldots, \mathbf{y}_{M(-1)})$ is a $T-1 \times M$ matrix with ith column $\mathbf{y}_{i(-1)} = (y_{i1}, \ldots, y_{i,T-1})'$. The OLS estimator for equation i is then

$$\begin{pmatrix} \hat{\boldsymbol{\alpha}}_i \\ \hat{\lambda}_i \end{pmatrix} = (\mathbf{Y}'\mathbf{Y})^{-1}\mathbf{Y}'\mathbf{y}_i, \tag{20}$$

where $\mathbf{Y} = (\mathbf{y}_{(-1)}, \mathbf{1})$ is unchanged across the $i = 1, \ldots, M$ spatial units.

If the $\boldsymbol{\alpha}_i$ and λ_i contain spatial weighing matrices and/or other exogenously determined variables, as discussed previously, the conditional expectation and the prediction error are nonlinear in parameters. Then, a nonlinear least squares (NLS) estimator also minimises the criterion function $S_i = \mathbf{e}'_i\mathbf{e}_i$, but the estimator is now by necessity of an iterative type.

In the constant parameter case, the lagged y variables together with a constant vector are the only used instrumental variables (IV). Additional and higher order lags of y can be used as additional IVs to obtain an asymptotically more efficient estimator by the GMM estimation approach. When exogenous variables and spatial weights are incorporated, at least, the former type of variables can also be used for GMM estimation.

For all equations jointly and with constant parameters, the OLS estimator can in matrix form be written

$$\begin{pmatrix} \hat{\mathbf{A}}' \\ \hat{\lambda}' \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{\alpha}}_1 & \ldots & \hat{\boldsymbol{\alpha}}_M \\ \hat{\lambda}_1 & \ldots & \hat{\lambda}_M \end{pmatrix} = (\mathbf{I} \otimes \mathbf{Y}'\mathbf{Y})^{-1}(\mathbf{I} \otimes \mathbf{Y}')\mathbf{y} \tag{21}$$

with $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_M)$ and where $\otimes$ is the Kronecker matrix product. Note, that the SURE (the feasible generalised least squares estimator using information about any covariance matrix $\boldsymbol{\Sigma}$) estimator of Zellner (1962) based on identical regressors is identical to single equation or full system OLS. With exogenous influence through the parameters this simplification may no longer be justified. The NLS estimator now minimises

$S = \mathbf{e}'\mathbf{e}$, with respect to the hyper-parameters of $\mathbf{A}_t$ and $\boldsymbol{\lambda}_t$. Here, the $M(T-1) \times 1$ systemwide prediction error vector is $\mathbf{e} = (\mathbf{e}_1', \ldots, \mathbf{e}_M')' = \mathbf{y} - (\mathbf{I}_M \otimes \mathbf{y}_{(-1)}) \operatorname{vech} \mathbf{A}_t' - (\mathbf{I}_M \otimes \mathbf{I}_{t-1}) \boldsymbol{\Lambda}_t$ with $\boldsymbol{\Lambda}_t = (\boldsymbol{\lambda}_{1t}', \ldots, \lambda_{Mt}')'$, and where $\boldsymbol{\lambda}_{it} = (\lambda_{i2}, \ldots, \lambda_{iT})'$.

The mentioned estimators are all consistent and asymptotically normally distributed. To obtain a consistent estimator of the covariance matrix of the estimator it is important to recall that the models are characterised by conditional heteroskedasticity (cf. eq. (5)) and that the prediction error is sharing this property. Hence, the use of a sandwich or robust covariance matrix estimator is called for. Eq. (5) is an important ingredient in this. Note though, that (5) is based on strong assumptions about independent thinning and that using e_{it}^2 to replace $V(y_{it}|Y_{t-1})$ as in the White estimator for linear regressions may be an even more robust alternative (see also Brännäs, 1995).

3.2 Simultaneous Equations

To estimate simultaneous equation models, we make use of the rewritten simultaneous equation model in (16), i.e. $\mathbf{y}_t = \mathbf{A}_{0t}\mathbf{y}_t + \mathbf{A}_{1t}\mathbf{y}_{t-1} + \boldsymbol{\lambda}_t + \boldsymbol{\epsilon}_t^{**}$, which appears to be the most convenient starting point. Conventional OLS or NLS estimation will in this case be inconsistent due to the right hand side endogenous variables. Therefore, for consistent estimation we consider IV type limited information estimators for single equations as well as full information estimation for all equations jointly. In these approaches an important first step is the one of finding instruments for the right hand side endogenous variables.

Given that the simultaneous equations model contains lagged endogenous variables which by assumption are independent of the random disturbance term $\boldsymbol{\epsilon}_t$ in (11) and (14), it is natural to consider $\mathbf{y}_{t-k}, k \geq 1$ as potentially valid instrumental variables for the right hand side $\mathbf{y}_t$.

Recall that $\boldsymbol{\epsilon}_t^{**}$ contains both current and lagged endogenous variables, but since,

$$
\begin{aligned}
E[(\mathbf{A}_{0t} \circ \mathbf{y}_t - \mathbf{A}_{0t}\mathbf{y}_t)\mathbf{y}_{t-k}|Y_{t-k}] &= \mathbf{0} \\
E[(\mathbf{A}_{1t} \circ \mathbf{y}_{t-1} - \mathbf{A}_{1t}\mathbf{y}_{t-1})\mathbf{y}_{t-k}|Y_{t-k}] &= \mathbf{0}
\end{aligned}
$$

it follows that vectors $\mathbf{y}_{t-k}, k \geq 1$ are also unconditionally uncorrelated with the composite disturbance term $\boldsymbol{\epsilon}_t^{**}$. In addition, it follows from the model specification that the instrumental variables are correlated with the right hand side y_{it} variables. Therefore, with these instrumental variables the IV or GMM estimators based on a single equation or on all equations jointly will, at least, in the constant parameter case, i.e. when the model contains $\mathbf{A}_0$, $\mathbf{A}_1$ and $\boldsymbol{\lambda}$, be consistent and asymptotically, normally distributed. The use of the two stage least squares estimator, with (15) estimated in a first step, will also give a consistent estimator.

Since the number of available instrumental variables will typically be large in this setting, the GMM estimator will be quite efficient. In fact, the current context is very close, in how IV matrices are constructed, to the one treated in the literature on the estimation of the first order dynamic panel data model.

When exogenous variables and weight matrices are nonlinearly included in the $\mathbf{A}_{0t}$ and $\mathbf{A}_{1t}$ matrices and in the $\boldsymbol{\lambda}_t$ vector, the estimation of the hyper-parameter vector is to be made by some nonlinear version of the IV or GMM estimators. In such a case, the exogenous variables and their lags can also be used as instrumental variables. Depending on the parametrisation, there may be cases where systemwide rather than single equation estimation should be pursued. This is, e.g., the case when some parameters are viewed as constant across equations.

Next, we consider estimation based on the prediction error or alternatively on the reduced form. The ith equation of (11) is $y_{it} = \sum_{j=1,j\neq i}^{M} \alpha_{ij}^{0} \circ y_{jt} + \sum_{j=1,j}^{M} \alpha_{ij}^{1} \circ y_{j,t-1} + \epsilon_{it}$ and simultaneity implies that the right hand side current y_{jt} variables are correlated with ϵ_{it}. The prediction error is $\tilde{y}_{it} = y_{it} - E(y_{it}|Y_{t-1})$, which by specialisation of (2) can be written

$$\tilde{y}_{it} = y_{it} - \sum_{j=1}^{M} c_{ij} y_{j,t-1} - \lambda_i^*,$$

where c_{ij} is the jth element in the ith row of $\mathbf{C} = (\mathbf{I} - \mathbf{A}_0)^{-1}\mathbf{A}_1$ and λ_i^* is the ith element of the $M \times 1$ vector $(\mathbf{I} - \mathbf{A}_0)^{-1}\boldsymbol{\lambda}$. The prediction error has mean zero, but its variance is heteroskedastic, cf. (17). For the full system, the prediction error is

$$\tilde{\mathbf{y}}_t = \mathbf{y}_t - (\mathbf{I} - \mathbf{A}_0)^{-1}(\mathbf{A}_1\mathbf{y}_{t-1} - \boldsymbol{\lambda}).$$

Whether we consider limited or full information estimation methods, nonlinearity is a key ingredient of any least squares estimator based on this prediction error. Obviously, it will also be important to consider the identifiability of individual equations.

Appendix A: Some results for binomial thinning operators

The binomial thinning operator is defined as

$$\alpha \circ y = \sum_{i=1}^{y} u_i,$$

where y is an integer-valued random variable, and the $\{u_i\}$ sequence is made up of $0-1$ iid random variables. The probability $\Pr(u_i = 1) = \alpha$ is constant across i. Hence, the integer-valued $\alpha \circ y \in [0, y]$, and for a given y it follows a binomial distribution. If, e.g., y is Poisson distributed the distribution of $\alpha \circ y$ is Poisson distributed as well.

Results are most easily obtained using the probability generating function (pgf). For the case of a given y we have $E(t^{\alpha\circ y}|y) = E(t^{u_1+u_2+\ldots+u_y}|y) = E^y(t^u) = [t^0 \cdot \Pr(u = 0) + t^1 \cdot \Pr(u = 1)]^y = [1-\alpha+\alpha t]^y$. The conditional expectation is then $E(\alpha \circ y|y) = \partial E(t^{\alpha\circ y}|y)/\partial t_{|t=1} = \alpha y$ and unconditionally we get $E(\alpha \circ y) = E_y[E(\alpha \circ y|y)] = \alpha E(y)$.

The corresponding second order moments are $E[(\alpha \circ y)^2|y] = \alpha y + \alpha^2 y(y-1)$ and unconditionally $E[(\alpha \circ y)^2] = \alpha^2 E(y^2) + \alpha(1-\alpha)E(y)$. From these results it follows that the conditional variance is $V(\alpha \circ y|y) = \alpha(1-\alpha)y$ and that the unconditional variance is $V(\alpha \circ y) = \alpha(1-\alpha)E(y) + \alpha^2 V(y)$. In addition, $E[(\alpha \circ y_i)(\alpha \circ y_j)|y_i, y_j] = \alpha^2 y_i y_j$ and $E(\alpha \circ y_i)(\alpha \circ y_j) = \alpha^2 E(y_i y_j)$.

Among other useful results we mention the following equalities in distribution: $\alpha \circ (\beta \circ y) = (\alpha\beta) \circ y$; $\alpha \circ \beta \circ y = \beta \circ \alpha \circ y$ and $\alpha \circ (y + x) = \alpha \circ y + \alpha \circ x$. Note however, that $\alpha \circ y + \beta \circ y$ and $(\alpha + \beta) \circ y$ are not equal in distribution.

Some inversion results

Some inversion results can easily be obtained in a univariate framework using the pgf. With equality in distribution for $\alpha \circ y = \sum_{i=1}^{y} u_i = x$ we ideally wish to characterise the distribution of y in terms of the one for x. The pgf of x satisfies $E(t^x) = E_y[E(t^u)^y]$ and we already gave the result $E(y) = \alpha^{-1}E(x)$, above. We also gave the result $E(x^2) = E[(\alpha \circ y)^2] = \alpha^2 E(y^2) + \alpha(1-\alpha)E(y)$ which then directly gives, e.g., $V(y) = \alpha^{-2}V(x) - \alpha^{-2}(1-\alpha)E(x)$. Taking higher order derivatives of $E(t^x)$ with respect to t and setting

$t = 1$ gives that all moments up to order k can be obtained from the equality

$$E\left[\prod_{i=0}^{k}(y-i)\right] = \alpha^{-k} E\left[\prod_{i=0}^{k}(x-i)\right].$$

We may proceed by using the assumption of independence between thinning operations to study $x_i = \sum_{j=1}^{M} \alpha_{ij} \circ y_j$, for $i = 1, \ldots, M$. Conditional on $\mathbf{y}$ we have for every $x_i, i = 1, \ldots, M$, that $E(t^{\sum_{j=1}^{M} \alpha_{ij} \circ y_j} | \mathbf{y}) = \prod_{j=1}^{M}(1 - \alpha_{ij} + \alpha_{ij} t)^{y_j}$. The conditional expectation of x_i conditional on $\mathbf{y}$ is then

$$E(x_i|\mathbf{y}) = \partial E(t^{x_i}|\mathbf{y})/\partial t_{|t=1} = \sum_{j=1}^{M} \alpha_{ij} y_j$$

so that the unconditional expectation is $E(x_i) = \sum_{j=1}^{M} \alpha_{ij} E(y_j)$. A first inversion result then follows as

$$E(\mathbf{y}) = \mathbf{A}^{-1} E(\mathbf{x}).$$

To obtain an inversion result for the simultaneous equation model containing the $\mathbf{A}_*$ matrix with its negative elements, note that for $x = -(\alpha \circ y)$ we get $E(x|y) = -\alpha y$. Therefore, for (12) the inversion is of the form $E(\mathbf{y}_t) = \mathbf{A}_*^{-1} E(\mathbf{A}_1 \circ \mathbf{y}_{t-1} + \boldsymbol{\epsilon}_t)$ which gives $E(\mathbf{y}_t) = (\mathbf{A}^* - \mathbf{A}_1)^{-1} \boldsymbol{\lambda}$, and for (13) it is $E(\mathbf{y}_t) = \mathbf{A}_*^{-1} \boldsymbol{\lambda}_t$.

Additional results for the multivariate case $\mathbf{x} = \mathbf{A} \circ \mathbf{y}$ can be obtained by using the full multivariate probability generating function $\Psi = E(t_1^{x_1} t_2^{x_2}, \cdots t_M^{x_m})$. For instance, conditional on $\mathbf{y}$ the conditional pgf can be written as $\Psi_y = \prod_{j=1}^{M} E(t_j^{x_j}|\mathbf{y})$. The conditional and unconditional results given above can then be obtained from $\partial \Psi_y / \partial t_i = \Psi_y \, \partial \ln \Psi_y / \partial t_i$ evaluated at $t_1 = \ldots = t_M = 1$. For higher order moment results higher order derivatives of Ψ_y are required.

Appendix B: Brief derivation of the structural form conditional variance

The conditional expecion of $\boldsymbol{\epsilon}_t^{**} = (\boldsymbol{\epsilon}_t - \boldsymbol{\lambda}_t) + (\mathbf{A}_{0t} \circ \mathbf{y}_t - \mathbf{A}_{0t}\mathbf{y}_t) + (\mathbf{A}_{1t} \circ \mathbf{y}_{t-1} - \mathbf{A}_{1t}\mathbf{y}_{t-1})$ is zero, so that $V(\mathbf{y}_t|Y_{t-1}) = \mathbf{A}_{*t}^{-1}E(\boldsymbol{\epsilon}_t^{**}\boldsymbol{\epsilon}_t^{\prime **}|Y_{t-1})(\mathbf{A}_{*t}^{\prime})^{-1}$. We use short hand notations and drop the subindex t to write $\boldsymbol{\epsilon}_t^{**} = \bar{\boldsymbol{\epsilon}} + \bar{\mathbf{A}}_0 + \bar{\mathbf{A}}_1$. The expression to simplify is then

$$\begin{aligned} E(\boldsymbol{\epsilon}_t^{**}\boldsymbol{\epsilon}_t^{\prime **}|Y_{t-1}) &= E[\bar{\boldsymbol{\epsilon}}\bar{\boldsymbol{\epsilon}}' + \bar{\boldsymbol{\epsilon}}\bar{\mathbf{A}}_0' + \bar{\boldsymbol{\epsilon}}\bar{\mathbf{A}}_1' + \bar{\mathbf{A}}_0\bar{\boldsymbol{\epsilon}}' + \bar{\mathbf{A}}_0\bar{\mathbf{A}}_0' + \bar{\mathbf{A}}_0\bar{\mathbf{A}}_1' \\ &\quad + \bar{\mathbf{A}}_1\bar{\boldsymbol{\epsilon}}' + \bar{\mathbf{A}}_1\bar{\mathbf{A}}_0' + \bar{\mathbf{A}}_1\bar{\mathbf{A}}_1'|Y_{t-1}]. \end{aligned}$$

It follows directly from the assumptions that $E(\bar{\boldsymbol{\epsilon}}\bar{\boldsymbol{\epsilon}}'|Y_{t-1}) = \boldsymbol{\Sigma}$ and that $E(\bar{\boldsymbol{\epsilon}}\bar{\mathbf{A}}_1'|Y_{t-1}) = \mathbf{0}$. Obviously, for the transposed matrix $E(\bar{\mathbf{A}}_1\bar{\boldsymbol{\epsilon}}'|Y_{t-1}) = \mathbf{0}'$.

We get $E(\bar{\mathbf{A}}_1\bar{\mathbf{A}}_1'|Y_{t-1}) = E((\mathbf{A}_{1t} \circ \mathbf{y}_{t-1})(\mathbf{A}_{1t} \circ \mathbf{y}_{t-1})'|Y_{t-1}) - \mathbf{A}_{1t}\mathbf{y}_{t-1}\mathbf{y}_{t-1}'\mathbf{A}_{1t}'$ which after manipulation comes out as a diagonal matrix with diagonal elements as in $\boldsymbol{\Theta}_{1,t-1}$.

To obtain $E(\bar{\mathbf{A}}_0\bar{\mathbf{A}}_1'|Y_{t-1})$ we rewrite the model as $\mathbf{A}_{0t} \circ \mathbf{y}_t = \mathbf{y}_t - \mathbf{A}_{1t} \circ \mathbf{y}_{t-1} - \boldsymbol{\epsilon}_t$ to get $\bar{\mathbf{A}}_0 = (\mathbf{I} - \mathbf{A}_{0t})\mathbf{y}_t - \bar{\mathbf{A}}_1 - \boldsymbol{\epsilon}_t - \mathbf{A}_{1t}\mathbf{y}_{t-1}$. It then follows that

$$\begin{aligned} E(\bar{\mathbf{A}}_0\bar{\mathbf{A}}_1'|Y_{t-1}) &= (\mathbf{I} - \mathbf{A}_{0t})[E(\mathbf{y}_t|Y_{t-1})\mathbf{y}_{t-1}'\mathbf{A}_{1t}' - E(\mathbf{y}_t|Y_{t-1})\mathbf{y}_{t-1}'\mathbf{A}_{1t}'] - \boldsymbol{\Theta}_{1,t-1} \\ &= -\boldsymbol{\Theta}_{1,t-1}. \end{aligned}$$

Next to get $E(\bar{\boldsymbol{\epsilon}}\bar{\mathbf{A}}_0'|Y_{t-1})$ we rewrite $\bar{\mathbf{A}}_0'$ as above and can then write

$$\begin{aligned} E(\bar{\boldsymbol{\epsilon}}\bar{\mathbf{A}}_0'|Y_{t-1}) &= E((\boldsymbol{\epsilon}_t - \boldsymbol{\lambda}_t)(\mathbf{y}_t - \mathbf{A}_{0t}\mathbf{y}_t)'|Y_{t-1}) - E((\boldsymbol{\epsilon}_t - \boldsymbol{\lambda}_t)(\mathbf{A}_{0t} \circ \mathbf{y}_{t-1})'|Y_{t-1}) \\ &\quad - E((\boldsymbol{\epsilon}_t - \boldsymbol{\lambda}_t)\boldsymbol{\epsilon}_t'|Y_{t-1}) \\ &= [E(\boldsymbol{\epsilon}_t\mathbf{y}_t'|Y_{t-1}) - \boldsymbol{\lambda}_t E(\mathbf{y}_t|Y_{t-1}](\mathbf{I} - \mathbf{A}_{0t})' - \boldsymbol{\Sigma}. \end{aligned}$$

Finally, $E(\bar{\mathbf{A}}_0\bar{\mathbf{A}}_0'|Y_{t-1}) = E(\mathbf{A}_{0t} \circ \mathbf{y}_t(\mathbf{A}_{0t} \circ \mathbf{y}_t)'|Y_{t-1}) - \mathbf{A}_{0t}E(\mathbf{y}_t\mathbf{y}_t'|Y_{t-1})\mathbf{A}_{0t}'$. After some tedious manipulation we write the result as $\mathrm{diag}(\boldsymbol{\Theta}_{0,t-1}) = \sum_{j=1}^{M} \alpha_{ij}^0(1 - \alpha_{ij}^0)E(y_{jt}|Y_{t-1})$, for $i = 1, \ldots, M$.

Collecting parts we obtain the result given in Section 2.3.

References

Al-Osh, M.A. and Alzaid, A.A. (1987). First Order Integer-valued Autoregressive INAR(1) Process. *Journal of Time Series Analysis* **8**, 261-275.

Anselin, L. (1988). *Spatial Econometrics. Methods and Models*. Kluwer, Boston.

Anselin, L., Florax, R.J.G.M. and Rey, S.J. (eds) (2004). *Advances in Spatial Econometrics*. Springer, Berlin.

Bartholomew, D.J. (1982). *Stochastic Models for Social Processes*. 3rd edition, Wiley, New York.

Berglund, E. and Brännäs, K. (1996). Entry and Exit of Plants: A Study Based on Swedish Panel Count Data for Municipalities. In *Yearbook of the Finnish Statistical Society 1995*, 95-111, Helsinki.

Berglund, E. and Brännäs, K. (2001). Plants' Entry and Exit in Swedish Municipalities. *Annals of Regional Science* **35**, 431-448.

Blundell, R., Griffith, R. and Windmeijer, F. (2002). Individual Effects and Dynamics in Count Data Models. *Journal of Econometrics* **108**, 113-131.

Boudreault, M. and Charpentier, A. (2011). Multivariate Integer-Valued Autoregressive Models Applied to Earthquake Counts. arXiv:1112.0929 [stat.AP].

Brännäs, K. (1995). Explanatory Variables in the AR(1) Count Data Model. Umeå Economic Studies 381.

Brännäs, K. (2013). Simultaneity in the Multivariate Count Data Autoregressive Model. Umeå Economic Studies 870.

Brännäs, E. and Brännäs, K. (1998). A Model of Patch Visit Behaviour in Fish. *Biometrical Journal* **40**, 717-724.

Brännäs, K. and Hellström, J. (2001). Generalized Integer-Valued Autoregression. *Econometric Reviews* **20**, 425-443.

Brännäs, K. and Johansson, P. (1996). Panel Data Regression for Counts. *Statistical Papers* **37**, 191-213.

Cameron, A.C. and Trivedi, P.K. (1998). *Regression Analysis of Count Data*. Cambridge University Press, Cambridge.

Cameron, A.C. and Trivedi, P.K. (2005). *Microeconometrics. Methods and Applications.*

Cambridge University Press, Cambridge.

Ghodsi, A., Shitan, M. and Bakouch, H. (2012). A First-Order Spatial Integer-Valued Autoregressive SINAR(1,1) Model. *Communications in Statistics - Theory and Methods* **41**, 2773-2787.

Littlejohn, R.P. (1994). An Operation which Inverts Bernoulli Multiplication and Associated Stationary Reversible Markov Processes. *Journal of Applied Probability* **29**, 234-238.

McKenzie, E. (1985). Some Simple Models for Discrete Variate Time Series. *Water Resources Bulletin* **21**, 645-650.

McKenzie, E. (1988). Some ARMA models for Dependent Sequences of Poisson Counts. *Advances in Applied Probability* **20**, 822-835.

McKenzie, E. (2003) Discrete Variate Time Series. In *Handbook of Statistics*, Volume 21, Shanbhag, D.N. and Rao, C.R. (eds), Elsevier, Amsterdam, pp. 573-606.

Pedeli, X. and Karlis, D. (2013). Some Properties of Multivariate INAR(1) Processes. *Computational Statistics & Data Analysis* **67**, 213-225.

Rudholm, N. (2001). Entry and the Number of Firms in the Swedish Pharmaceuticals Market. *Review of Industrial Organization* **19**, 351-364.

Sengupta, A. and Cressie, N. (2013). Empirical Hierarchical Modelling for Count Data using the Spatial Random Effects Model. *Spatial Economic Analysis* **8**, 389-418.

Silva, I. (2005). Contributions to the Analysis of Discrete-Valued Time Series. PhD thesis. Department of Applied Mathematics, University of Porto. (Published as Analysis of discrete-valued time series. LAP Lambert Academic Publishing, 2012).

Steutel, F. W. and K. van Harn, K. (1979). Discrete Analogues of Self-Decomposability and Stability. *Annals of Probability* **7**, 893-899.

Winkelmann, R. (2008). *Econometric Analysis of Count Data*. 5th edition, Springer, Berlin.

Zeger, S.L. (1988). A Regression Model for Time Series of Counts. *Biometrika* **75**, 621-629.

Zellner, A. (1962). An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests of Aggregation Bias. *Journal of the American Statistical Association* **57**, 500-509.

Zhang, H. (2002). On Estimation and Prediction for Spatial Generalized Linear Models. *Biometrics* **58**, 129-136.