

Intermediate Spatial Econometrics with Applications in R

Mauricio Sarrias
Universidad de Talca

May 17, 2023

Contents

I	Introduction to Spatial Dependence	1
1	Introduction to Spatial Econometric	3
1.1	Why do We Need Spatial Econometric?	3
1.1.1	Spatial Dependence	4
1.1.2	Spatial Autocorrelation	4
1.2	Spatial Weight Matrix	6
1.2.1	Weights Based on Boundaries	7
1.2.2	Weights Based on Distance	9
1.2.3	Row-Standardized Weights Matrix	11
1.2.4	Spatial Lagged Variables	12
1.2.5	Higher Order Spatial	13
1.3	Examples of Weight Matrices in R	14
1.3.1	Creating Contiguity Neighbors	15
1.3.2	Creating Distance-based Neighbors	20
1.3.3	Constructing a Spatially Lagged Variable	22
1.4	Testing for Spatial Autocorrelation	24
1.4.1	Global Spatial Autocorrelation: Moran's I	24
1.5	Application: Poverty in Santiago, Chile	29
1.5.1	Cloropeth Graphs	29
1.5.2	Moran's I Test	30
1.6	Exercises	33
2	Spatial Models	35
2.1	Taxonomy of Models	35
2.1.1	Spatial Lag Model	35
2.1.2	Spatial Durbin Model	40
2.1.3	Spatial Error Model	41
2.1.4	Spatial Autocorrelation Model	42
2.2	Motivation of Spatial Models	43
2.2.1	SLM as a Long-run Equilibrium	43
2.2.2	SEM and Omitted Variables Motivation	44

2.2.3	SDM and Omitted Variables Motivation	45
2.3	Interpreting Spatial Models	45
2.3.1	Measuring Spillovers	45
2.3.2	Marginal Effects	46
2.3.3	Partitioning Global Effects Estimates Over Space	52
2.4	Predictors for Spatial Models	53
2.5	Lesage's Book Example	53
2.5.1	Commuting Times and Congestion	53
2.5.2	Computing Effects in R	55
2.5.3	Cumulative Effects	59
2.6	Exercises	62

II Estimation Methods 63

3	Review of Asymptotic Theory	65
3.1	Convergence of Deterministic Sequences	65
3.2	Convergence in Probability	69
3.2.1	Convergence in Quadratic Mean	73
3.3	Law of Large Numbers	75
3.4	Convergence in Distribution	82
3.5	Central Limit Theorems	85
3.6	Orders in Probability	89
3.7	Triangular Arrays	92
3.8	Matrix	95
3.9	Matrix Norm	95
3.10	Bounded Matrices and Useful Lemmas for Spatial Econometrics	95
3.11	Quadratic forms	98
3.12	CLT for Spatial Models	100
3.13	Exercises	101
	Appendix 3.A Inequalities	102
4	Maximum Likelihood Estimation	103
4.1	What Are The Consequences of Applying OLS?	103
4.1.1	Finite and Asymptotic Properties	103
4.1.2	Illustration of Bias	106
4.2	Maximum Likelihood Estimation of SLM	108
4.2.1	Maximum Likelihood Function	108
4.2.2	Score Vector and Estimates	110
4.2.3	Hessian	112
4.2.4	Ord's Jacobian	114
4.3	Maximum Likelihood Estimation of SEM	114
4.3.1	What Are The Consequences of Applying OLS on a SEM Model?	114
4.3.2	Log-likelihood function	115
4.3.3	Score Function and ML Estimates	116
4.4	Asymptotic Properties of SLM	117
4.4.1	Consistency of QMLE	118

4.4.2	Asymptotic Normality	123
4.5	Computing the Standard Errors For The Marginal Effects	124
4.6	Spillover Effects on Crime: An Application in R	125
4.6.1	Estimation of Spatial Models in R	125
4.6.2	Estimation of Marginal Effects in R	129
4.7	Programing the SLM in R	137
4.7.1	First approach	138
4.7.2	Second approach	142
4.8	Exercises	144
Appendix 4.A	Consistency of SLM Model	146
Appendix 4.B	Expected Value of Hessian for SLM	150
Appendix 4.C	Variance of the Score Function	152
Appendix 4.D	Proof of Asymptotic Normality	154
5	Hypothesis Testing	159
5.1	Test for Residual Spatial Autocorrelation Based on the Moran I Statistic	159
5.1.1	Cliff and Ord Derivation	159
5.1.2	Kelijan and Prucha (2001) Derivation of Moran's I	161
5.1.3	Example	161
5.2	Common Factor Hypothesis	162
5.3	Hausman Test: OLS vs SEM	163
5.4	Tests Based on ML	164
5.4.1	Likelihood Ratio Test	164
5.4.2	Wald Test	166
5.4.3	Lagrange Multiplier Test	168
5.4.4	Anselin and Florax Recipe	171
5.4.5	Lagrange Multiplier Test Statistics in R	171
5.5	Exercises	172
Appendix 5.A	Asymptotic Properties of Moran's I	173
6	Instrumental Variables and GMM	177
6.1	A Review of GMM	177
6.1.1	Model Specification	177
6.1.2	One-Step GMM Estimation	179
6.1.3	Two-Step GMM Estimation	180
6.2	Spatial Two Stage Estimation of SLM	182
6.2.1	Instruments in the Spatial Context	183
6.2.2	Defining the S2SLS Estimator	184
6.2.3	S2SLS Estimator as GMM	186
6.2.4	Additional Endogenous Variables	187
6.2.5	Consistency of S2SLS Estimator	188
6.2.6	Asymptotic Distribution of S2SLS Estimator	189
6.2.7	S2SLS Estimation in R	190
6.3	Generalized Moment Estimation of SEM Model	192
6.3.1	Spatially Weighted Least Squares	194
6.3.2	Moment Conditions	195
6.3.3	Feasible Generalized Least Squares Model	201

6.3.4	FGLS in R	204
6.4	Estimation of SAC Model: The Feasible Generalized Two Stage Least Squares estimator Procedure	205
6.4.1	Intuition Behind the Procedure	205
6.4.2	Moment Conditions Revised	208
6.4.3	Assumptions	211
6.4.4	Estimators and Estimation Procedure in a Nutshell	213
6.5	Application in R	219
6.5.1	SAC Model with Homokedasticity (GS2SLS)	220
6.5.2	SAC Model with Homokedasticity and Additional Endogeneity (GS2SLS)	221
6.6	Exercises	222
	Appendix 6.A Proof Theorem 3 in KP 1998	223
	Index	231

List of Figures

1.1	Environmental Externalities	3
1.2	Spatial Distribution of Poverty in Metropolitan Region, Chile	5
1.3	Spatial Autocorrelation	6
1.4	Rook Contiguity	8
1.5	Bishop Contiguity	8
1.6	Queen Contiguity	9
1.7	Higher-Order Neighbors	14
1.8	Plotting a Map in R	16
1.9	Commune with largest number of contiguities	18
1.10	Queen and Rook Criteria for MR	20
1.11	Different Spatial Weight Schemes for MR	23
1.12	Moran Scatterplot	26
1.13	Cloropleth map: Poverty in the Metropolitan Region	30
1.14	Moran Plot for Poverty	33
2.1	The SLM for two tegions	36
2.2	The SDM for Two Regions	40
2.3	The SEM for two regions	41
2.4	Taxonomy of spatial models	43
2.5	Regions east and west of the CBD	53
3.1	Convergence of sequence $2 + 3/n$	66
3.2	Bounded sequence	69
3.3	Illustration of convergnce in probability to a constant	70
3.4	Convergence of mean from normal distribution	78
3.5	Convergence of mean from binomial distribution	79
3.6	Chebychev's Convergence	82
3.7	Convergence of the sample mean and speed of convergence	92
4.1	Distribution of $\hat{\rho}$	107
4.2	Distances from R3 to all Regions	119
4.3	Spatial Distribution of Crime in Columbus, Ohio Neighborhoods	126

4.4	Effects of a Change in Region 30: Categorization	131
4.5	Effects of a Change in Region 30: Magnitude	132
6.1	Estimation steps for SAC model	214

List of Tables

4.1	Spatial Models for Crime in Columbus, Ohio Neighborhoods.	129
6.1	Spatial Models for Crime in Columbus: ML vs S2SLS	193
6.2	Spatial Models for Crime in Columbus: ML vs GM	206

Part I

Introduction to Spatial Dependence

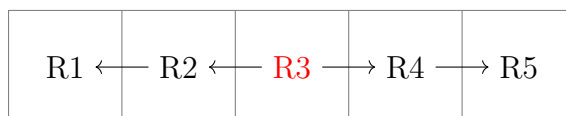
Introduction to Spatial Econometric

1.1 Why do We Need Spatial Econometric?

An important aspect of any study involving spatial units (cities, regions, countries, etc) is the potential relationships and interactions between them. For example, when modeling pollution at the regional level it is awkward to analyze each region as independent units. In fact, regions cannot be analyzed as isolated entities since they are spatially interrelated by ecological and economic interactions.

Consider Figure 1.1, where region 3 (R3) is highly industrialized, whereas region 1, 2, 4 and 5 are residential areas. If region 3 increases its economic activity, then pollution not only will increase in that region, but also in the neighbor regions. It is also expected that contamination will increase in region 1 and 5 but in lower magnitudes. These spatial externalities of R3 can be generated by both spatial-economic interactions (e.g. transportation of input and output from region 3) and spatial-ecological interactions (e.g. carbon emissions).

Figure 1.1: Environmental Externalities



Similarly, if we study crime at the city level then somehow we should incorporate the possibility that crime is localized. For example identification of concentration or cluster of greater criminal activity has emerged as a central mechanism to targeting a criminal justice and crime prevention response to crime problem. These clusters of crime are commonly referred to as **hotspots**: geographic locations of high crime concentration, relative to the distribution of crime across the whole region of interest.

Both examples implicitly state that geography location and distance matter. In fact, they reflect the importance of the first law of geography. According to Waldo Tobler: *“everything is related to everything else”, but near things are more related than distant things*. This first law is the foundation of the fundamental concepts of **spatial dependence** and **spatial autocorrelation**.

1.1.1 Spatial Dependence

Spatial dependence reflects a situation where values observed at one location or region, say observation i , depend on the values of neighboring observations at nearby locations. Formally, we might state

$$y_i = f(y_j), \quad i = 1, \dots, n, j \neq i.$$

In words, what happens in region i , depends on what happens in region j for all $j \neq i$. Using our previous example, we would like to estimate

$$\begin{aligned} y_1 &= \beta_{21}y_2 + \beta_{31}y_3 + \beta_{41}y_4 + \beta_{51}y_5 + \epsilon_1 \\ y_2 &= \beta_{12}y_1 + \beta_{32}y_3 + \beta_{42}y_4 + \beta_{52}y_5 + \epsilon_2 \\ y_3 &= \beta_{13}y_1 + \beta_{23}y_2 + \beta_{43}y_4 + \beta_{53}y_5 + \epsilon_3 \\ y_4 &= \beta_{14}y_1 + \beta_{24}y_2 + \beta_{34}y_3 + \beta_{54}y_5 + \epsilon_4 \\ y_5 &= \beta_{15}y_1 + \beta_{25}y_2 + \beta_{35}y_3 + \beta_{45}y_4 + \epsilon_5 \end{aligned}$$

where β_{ji} is the effect of pollution of region j on region i . However, it is easy to see that this would be of little practical usefulness, since it would result in a system with many more parameters than observations: we have $n = 5$ observations, but 20 parameters to be estimated, which implies that we do not have sufficient degrees of freedom. Intuitively, once we allow for dependence relation between a set of n observations/locations, there are potentially $n^2 - n$ relations that could arise. We subtract n from the potential n^2 dependence relations because we rule out dependence of an observation on itself.

The key point is that, under standard econometric modeling, it is impossible to model spatial dependency. However, as we will see in the next sections, we might be able to incorporate spatial relationships more efficiently using the so-called spatial weight matrix.

1.1.2 Spatial Autocorrelation

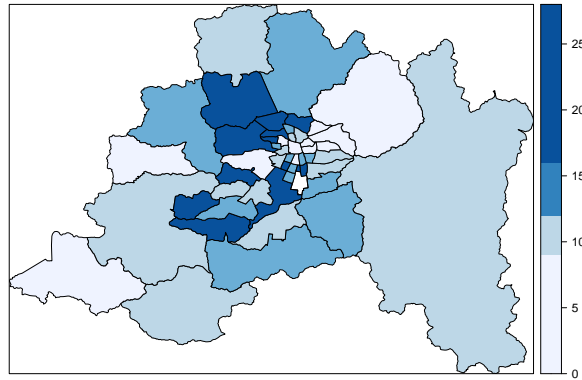
Another important concept is **spatial autocorrelation**. In space, the term autocorrelation refers to the correlation between the value of the variable at two different locations. Other ways of defining the same concept are: (1) correlation between the same attribute at two (or more) different locations, or (2) coincidence of values similarity with location similarity. Essentially, spatial autocorrelation is concerned with establishing whether the presence of a variable in one region in a regional system makes the presence of that variable in neighboring regions more, or less, likely.

The counterpart of spatial autocorrelation (and spatial dependency) is spatial randomness. Spatial randomness means that we cannot observe any spatial pattern in the data. That is, the value we observe in some spatial unit is equally likely as in any other spatial unit. Spatial randomness is important because it will form the null hypothesis later. If rejected, then there is evidence of spatial structure.

As an example, Figure 1.2 plots the spatial distribution of poverty in the Metropolitan Region, Chile. It can be observed that there is some spatial pattern where communes with similar rate of poverty are clustered.

Formally, the existence of spatial autocorrelation may be expressed by the following moment conditions:

Figure 1.2: Spatial Distribution of Poverty in Metropolitan Region, Chile



Notes: This graph shows the spatial distribution of poverty in the Metropolitan Region, Chile.

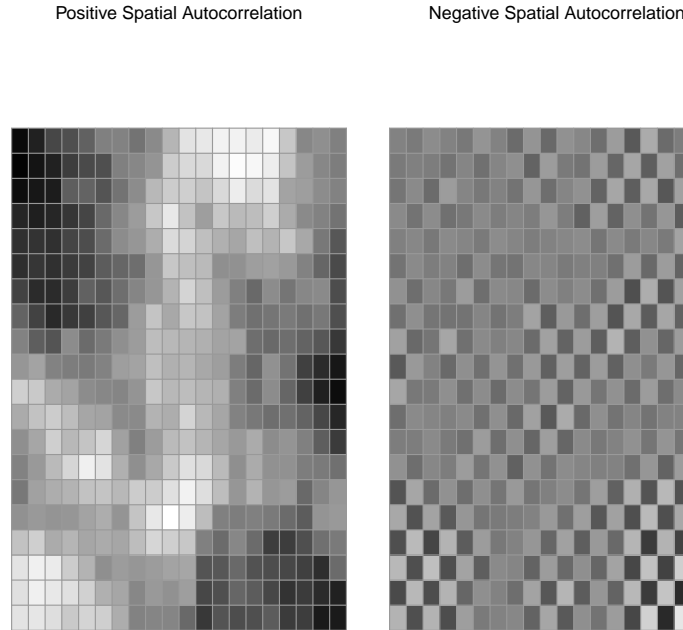
$$\text{Cov}(y_i, y_j) = \mathbb{E}(y_i y_j) - \mathbb{E}(y_i) \mathbb{E}(y_j) \neq 0 \text{ for } i \neq j,$$

where y_i and y_j are observations on a random variable at locations i and j in space, and i, j can be points or areal units. Therefore, a nonzero spatial autocorrelation exists between attributes of a feature defined at locations i and j if the covariance between feature attribute values at those points is nonzero. If this covariance is **positive** (i.e., if data with attribute values above the mean tend to be near other data with values above the mean), then we say there is **positive spatial autocorrelation**; if the converse is true, then we say there is **negative spatial autocorrelation**. Figure 1.3 show an example of positive and negative spatial autocorrelation.

Positive autocorrelation is much more common, but negative autocorrelation does exists, for example, in studies of welfare competition or federal grants competitions among local governments (Saavedra, 2000; Boarnet and Glazer, 2002), and studies of regional employment (Filiztekin, 2009; Pavlyuk, 2011), the cross-border lottery shopping (Garrett and Marsh, 2002), foreign direct investment in OECD countries (Garretsen and Peeters, 2009) and locations of Turkish manufacturing industry (Basdas, 2009). In short, we are interested in studying non-random spatial patterns and try to explain this non-randomness. Possible causes of non-randomness are (Gibbons et al., 2015):

- (a) Economic agents may be randomly allocated across space but some characteristics of locations varies across space and influences outcomes.
- (b) Location may have no causal effect on outcomes, but outcomes may be correlated across space because heterogeneous individuals or firms are non-randomly allocated across space.
- (c) Individual or firms may be randomly allocated across space but they interact so that decisions by one agent affects outcomes of other agents.

Figure 1.3: Spatial Autocorrelation



Notes: Spatial Autocorrelation among 400 spatial units arranged in an 20-by-20 regular square lattice grid. Different gray-tones refer to different values of the variable ranging from low values (white) to high values (black). The left plot shows positive spatial autocorrelation, whereas right plot shows negative spatial autocorrelation.

- (d) Individuals or firms may be non-randomly allocated across space and the characteristics of others nearby directly influences individual outcomes.

1.2 Spatial Weight Matrix

One of the crucial issues in spatial econometric is the problem of formally incorporating spatial dependence into the model. As we reviewed in Section 1.1.1, the main problem is that we have more parameter than observations. So, the question is: What would be a good criteria to define closeness in space? Or, in other words, how to determine which other units in the system influence the one under consideration?

The device typically used in spatial analysis to define the concept of closeness in space is the so-called “spatial weight matrix”, or more simply, \mathbf{W} matrix. If we assume that there are n spatial objects (regions, cities, countries), then \mathbf{W} will be a square matrix of dimension $n \times n$. This matrix imposes a structure in terms of what are the neighbors for each location. It assigns weights that measure the intensity of the relationship among pairs of spatial units. Thus, each element (i, j) of \mathbf{W} – which we denote by w_{ij} – expresses the degree of spatial proximity between the pair. This matrix can be represented in the form

$$\mathbf{W} = \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \dots & w_{nn} \end{pmatrix}$$

Generally, we assume that the diagonal elements of this “spatial neighbors” matrix are set to zero: “regions are not neighbors to themselves”.

A more formal definition of spatial weight matrix is the following:

Definition 1.2.1 — Spatial Weight Matrix. Let n be the number of spatial units. The spatial weight matrix, \mathbf{W} , a $n \times n$ **positive symmetric** and **non-stochastic** matrix with element w_{ij} at location i, j . The values of w_{ij} or the weights for each pair of locations are assigned by some preset rules which define the spatial relations among locations. By convention, $w_{ij} = 0$ for the diagonal elements.

Positive symmetric means that $w_{ij} \geq 0$ and $w_{ij} = w_{ji}$, for $i \neq j$. Thus, the interactions between spatial units cannot be negative. Non-stochastic means that the researcher takes \mathbf{W} as known *a priori*, and therefore, all results are conditional upon the specification of \mathbf{W} .

The definition of \mathbf{W} also requires a rule for w_{ij} . In other words, we need to figure out how to assign a real number to w_{ij} , for $i \neq j$, representing the strength of the spatial relationship between i and j . There are several ways of doing that. But, in general, there are two basic criteria. The first type establishes a relationship based on shared borders or vertices of lattice or irregular polygon data (contiguity). The second type establishes a relationship based on the distance between locations. Generally speaking, contiguity is most appropriate for geographic data expressed as polygons (so-called areal units), whereas distance is suited for point data, although in practice the distinction is not that absolute.

1.2.1 Weights Based on Boundaries

The availability of polygon or lattice data permits the construction of contiguity-based spatial weight matrices. A typical specification of the contiguity relationship in the spatial weight matrix is

$$w_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are contiguous,} \\ 0 & \text{if } i \text{ and } j \text{ are not contiguous.} \end{cases}$$

In a regular grid, neighbors (contiguity) can be defined in a number of ways. In analogy of the game of chess, rook contiguity, bishop contiguity and queen contiguity are distinguished.

Rook Contiguity

In this case, two locations are neighbors if they share at least part of a **common border or side**. In Figure 1.4 we have a regular grid with 9 regions: each square represents a region. If for example we want to define the neighbors of region 5 using the rook criteria, then its neighbors will be regions 2, 4, 6 and 8. Those represent the regions filled in red.

If we continue with this reasoning, then the 9×9 \mathbf{W} matrix will be:

Figure 1.4: Rook Contiguity

1	2	3
4	5	6
7	8	9

$$\mathbf{W} = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix} \quad (1.1)$$

Bishop Contiguity

In bishop contiguity (**which is seldom used in practice**), region i 's neighbors are located at its corners. Figure 1.5 shows the neighbors of region 5 under this scheme. The neighbors are regions 1, 3, 7 and 9. Note that regions in the interior will have more neighbors than those in the periphery.

Figure 1.5: Bishop Contiguity

1	2	3
4	5	6
7	8	9

The resulting \mathbf{W} matrix will be:

$$\mathbf{W} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

This criteria is seldom used in practice.

Queen Contiguity

In queen contiguity, any region that touches the boundary of region i , whether on a side or a single point, is considered neighbor. Under this criteria, the neighbors of 5 will be regions: 1, 2, 3, 4, 6, 7, 8 and 9.

Figure 1.6: Queen Contiguity

1	2	3
4	5	6
7	8	9

1.2.2 Weights Based on Distance

Weights may also be defined as a function of the distance between region i and j , d_{ij} . This distance is usually computed as the distance between their centroids, but it may of course be between other relevant points for each spatial units, such as the capital – or largest city– or each region. Unlike the weights based on contiguity, matrices based on distances only need the coordinates of the points.

There are several ways of computing the distance between two spatial units. Let x_i and x_j be the longitude; and y_i and y_j the latitude coordinates for region i and j , respectively. The most general concept of distance is the Minkowski metric:

$$d_{ij}^p = (|x_i - x_j|^p + |y_i - y_j|^p),$$

for two points i and j , with respective coordinates (x_i, y_i) and (x_j, y_j) , and with p as the parameter. The most familiar special case is the Euclidean or straight line distance with $p = 2$:

$$d_{ij}^e = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}.$$

Another employed metric is the Manhattan block distance. This measure only considers movement along the east-west and north-south directions, i.e., by straight angles. This yield a distance measure where $p = 1$:

$$d_{ij}^m = |x_i - x_j| + |y_i - y_j|.$$

All three measures presented above are useful if we consider the earth as a plane. For example, the Euclidean distance is the length of a straight line on a map, and is not necessarily the shortest distance if you take into account the curvature of the earth. The great circle distance take into account the curvature of Earth. Ships and aircraft usually follow the great circle geometry to minimize the distance and save time and money. In particular, the great circle distance is computed as:

$$d_{ij}^{cd} = r \times \arccos^{-1} [\cos |x_i - x_j| \cos y_i \cos y_j + \sin y_i \sin y_j]$$

where r is the Earth's radius. The arc distance is obtained in miles with $r = 3959$ and in kilometers with $r = 6371$.

Inverse Distance

Now we have to transform the information about the distances among spatial points into a weight scheme. The idea is that $w_{ijt} \rightarrow 0$ as $d_{ij} \rightarrow \infty$. In other words, the closer is j to i , the larger w_{ij} should be to conform to Tobler's first law.

In the inverse distance weighting scheme, the weights are inversely related to separation distance as shown below:

$$w_{ij} = \begin{cases} \frac{1}{d_{ij}^\alpha} & \text{if } i \neq j \\ 0 & \text{if } i = j, \end{cases}$$

where the exponent α is a parameter that is usually set by the researcher. In practice, the parameters are seldom estimated, but typically set to $\alpha = 1$ or $\alpha = 2$. Therefore, the weights are given by the reciprocal of the distance: the larger the distance between to spatial units, the lowest the spatial weight or the spatial connection. Finally, by convention, the diagonal elements of the spatial weights are set to zero and not computed. Plugging in a value of $d_{ii} = 0$ would yield division by zero for inverse distance weights.

Negative Exponential Model

Here the weights decline exponentially with separation distance

$$w_{ij} = \exp \left(-\frac{d_{ij}}{\alpha} \right),$$

where α is a parameter that is commonly chosen by researcher. Since the weights are given by the exponential of the negative distance, the greater the distance between i and j , the lower w_{ij} .

Both the inverse distance and the negative exponential distance depend not only on the parameter value and functional form, but also on the metric used for distance. Since the weights are inversely related to distance, larger values for the latter will yield small values for the former, and vice versa. This may be a problem in practice when the distances are so large

that the corresponding inverse distance weights become close to zero, possibly resulting in a zero spatial weight matrix. In addition, a potential problem may occur when the distance metric is such that distances take on values less than one, which is typically not a desired result (Anselin and Rey, 2014).

***k*-nearest Neighbors**

An alternative type of spatial weights that avoids the problem of isolates is to select the *k*-nearest neighbors. In contrast to the distance band, this is not a symmetric relation. However, a potential problem with this type of neighbors is the occurrence of ties, i.e., when more than one location *j* has the same distance from *i*. A number of solutions exist to break the tie, from randomly selecting one the *k*-th order neighbors, to including all of them.

Threshold Distance (Distance Band Weights)

In contrast to the *k*-nearest neighbors method, the threshold distance specifies that a region *i* is neighbor of *j* if the distance between them is less than a specified maximum distance:

$$w_{ij} = \begin{cases} 1 & \text{if } 0 \leq d_{ij} \leq d_{max} \\ 0 & \text{if } d_{ij} > d_{max}. \end{cases}$$

To avoid isolates that would result from too stringent a critical distance, the distance must be chosen such that each location has at least one neighbor. Such a distance conforms to a max-min criterion, i.e., it is the largest of the nearest neighbor distances.

Finally, it is important to note that a weights matrix obtained from a distance band is always symmetric, since distance is a symmetric relation.

1.2.3 Row-Standardized Weights Matrix

In practice, the spatial weights are seldom used in their binary (or distance) form, but subject to a transformation or standardization. In particular, we would like to compute weighted averages in which more weight is placed on nearby observations than on distant observations. To do so, we can define a row-standardized weight matrix \mathbf{W}^s , whose element w_{ij}^s is given by:

$$w_{ij}^s = \frac{w_{ij}}{\sum_j w_{ij}}.$$

This ensures that all weights are between 0 and 1 and facilitates the interpretation of operation with the weights matrix as an averaging of neighboring values as we will see below. The row-standardized weights matrix also ensures that the spatial parameter in many spatial stochastic processes are comparable between models (Anselin and Bera, 1998).

Another important feature is that, under row-standardization, the element of each row sum to unity and the sum of all weights, $S_0 = \sum_i \sum_j w_{ij} = n$, the total number of observations. This is a nice interpretation that we will explore later.

Another important issue is about symmetry. An important characteristic of symmetric matrix is that all its characteristic roots are real. However, **after the row standardization the matrices are no longer symmetric.**

The row-standardized matrix is also known in the literature as the row-stochastic matrix:

Definition 1.2.2 — Row-stochastic Matrix. A real $n \times n$ matrix \mathbf{A} is called **Markov matrix**, or **row-stochastic matrix** if

- (a) $a_{ij} \geq 0$ for $1 \leq i, j \leq n$;
- (b) $\sum_{j=1}^n a_{ij} = 1$ for $1 \leq i \leq n$

An important characteristic of the row-stochastic matrix is related to its eigen values:

Theorem 1.1 — Eigenvalues of row-stochastic Matrix. Every eigenvalue ω_i of a row-stochastic Matrix satisfies $|\omega| \leq 1$

Therefore, the eigenvalues of the row-stochastic (i.e., row-normalized, row standardized or Markov) neighborhood matrix $\mathbf{W}^s = (w_{ij}^s)$ are in the range $[-1, +1]$.

Finally, the behavior of \mathbf{W}^s is important for asymptotic properties of estimators and test statistics (Anselin and Bera, 1998, pp. 244). In particular, the \mathbf{W} matrix should be also exogenous, unless endogeneity is considered explicitly in the model specification.

1.2.4 Spatial Lagged Variables

Now that we have discussed the spatial weight matrix, we can create the so-called **spatially lagged variables** or **spatial lag operator**. The spatial lag operator takes the form $\mathbf{y}_L = \mathbf{W}\mathbf{y}$ with dimension $n \times 1$, where each element is given by $\mathbf{y}_{Li} = \sum_j w_{ij}y_j$, i.e., a weighted average of the \mathbf{y} values in the neighbor of i .

For example:

$$\mathbf{W}\mathbf{y} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 10 \\ 50 \\ 30 \end{pmatrix} = \begin{pmatrix} 50 \\ 10 + 30 \\ 50 \end{pmatrix}.$$

Using a row-standardized weight matrix:

$$\mathbf{W}\mathbf{y} = \begin{pmatrix} 0 & 1 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 10 \\ 50 \\ 30 \end{pmatrix} = \begin{pmatrix} 50 \\ 5 + 15 \\ 50 \end{pmatrix}.$$

As a result, for spatial unit i , the spatial lag of y_i , referred as \mathbf{y}_{Li} (the variable Wy observed for location i) is:

$$\mathbf{y}_{Li} = w_{i,1}y_1 + w_{i,2}y_2 + \dots + w_{i,n}y_n,$$

or,

$$\mathbf{y}_{Li} = \sum_{j=1}^n w_{i,j}y_j,$$

where the weights w_{ij} consists of the elements of the i th row of the matrix \mathbf{W} , matched up with the corresponding elements of the vector \mathbf{y} . In other words, this is a weighted sum of the values observed at neighboring locations, since the non-neighbors are not included.

- R** As stated by [Anselin \(1988, p. 23-24\)](#), standardization must be done with caution.¹ For example, when the weights are based on an inverse distance function (or similar concept of distance decay), which has a meaningful economic interpretation, scaling the rows so that the weights sum to one may result in a loss of that interpretation. Can you give an example?

1.2.5 Higher Order Spatial

So far we have learned how to define the geographical space by matrix \mathbf{W} . However, an interesting question is how to define higher-order neighbors. For example, we may be interested in defining the neighbors of the neighbors of a spatial unit. Or even we might be interested in the neighbors of neighbors of neighbors of spatial unit i . To discuss this interesting case we need to define **higher-order spatial weight matrices**.

We define the higher-order spatial weight matrix l as \mathbf{W}^l . So, for example the spatial weight of order $l = 2$ is given by $\mathbf{W}^2 = \mathbf{W}\mathbf{W}$, spatial weight matrix of order $l = 3$ is given by $\mathbf{W}^3 = \mathbf{W}\mathbf{W}\mathbf{W}$, and so on. What is the meaning of the element w_{ij} in this case? For spatial weights of order 2, the element w_{ij} of the weight matrix is 1 if polygon j is adjacent to the first order neighbors of polygon i and is 0 otherwise. Thus, for spatial neighbor weights of order n , the element w_{ij} of the weight matrix \mathbf{W} is 1 if polygon j is adjacent to the neighbors of order $n - 1$ of polygon i , and is 0 otherwise.

To illustrate these points, consider the following spatial structure for our example in Section 1.1:

$$\mathbf{W} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}. \quad (1.2)$$

Then $\mathbf{W}^2 = \mathbf{W}\mathbf{W}$ based on the 5×5 first-order contiguity matrix \mathbf{W} from (1.2) is:

$$\mathbf{W}^2 = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 2 & 0 & 1 & 0 \\ 1 & 0 & 2 & 0 & 1 \\ 0 & 1 & 0 & 2 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix} \quad (1.3)$$

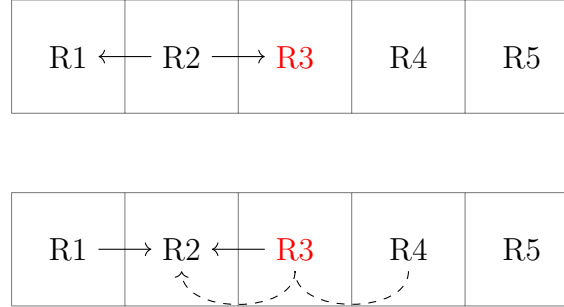
Note that for region $R1$, the second-order neighbors are regions $R1$ and $R3$. That is, region $R1$ is a second-order neighbor to itself as well as to region $R3$, which is a neighbor to the neighboring region $R2$.

Now consider $R2$. The first panel of Figure 1.7 shows the first-order neighbors of $R2$ given by the spatial weight matrix in (1.2): the first-order neighbors are $R1$ and $R3$. Panel B considers the second-order neighbors of $R2$: the second-order neighbors are $R2$ itself and $R4$. To understand this, note that there is a feedback effect from the first impact from $R2$ coming from $R1$ and $R3$ (first-order neighbors of $R2$). This explains why the element $w_{22}^2 = 2$. Moreover, there is an indirect effect coming from $R4$ through $R3$ that finally impacts $R2$. This represents the value of 1 for the element w_{24}^2 .

¹See also [Elhorst \(2014, p. 12\)](#) and references therein.

Similarly, for region $R3$, the second-order neighbors are regions $R1$ (which is a neighbor to the neighboring region $R2$), $R3$ (a second-order neighbor to itself), and $R5$ (which is a neighbor to the neighboring region $R4$).

Figure 1.7: Higher-Order Neighbors



Similarly, the third-order neighbors are:

$$\mathbf{W}^3 = \begin{pmatrix} 0 & 2 & 0 & 1 & 0 \\ 2 & 0 & 3 & 0 & 1 \\ 0 & 3 & 0 & 3 & 0 \\ 1 & 0 & 3 & 0 & 2 \\ 0 & 1 & 0 & 2 & 1 \end{pmatrix}$$

Could you explain the elements of this matrix?

1.3 Examples of Weight Matrices in R

Creating spatial weight matrices by hand is tedious (and in some cases almost impossible). However, there exists several statistical software that allow us to create them in a very simply fashion. First, we need the **shape file**, which has geographical information. The shapefile format is a digital vector storage for storing geometric location and associated attribute information. Nowadays it is possible to read and write geographical datasets using the shapefile format with a wide variety of software.

The shapefile format is simple because it can store the primate geometric data types of points, lines and polygons. Shapes (points/lines/polygons) together with data attributes can create infinitely many representations about geographic data. The three mandatory files have filename extensions **.shp**, **.shx**, and **.dbf**. The actual shapefile relates specifically to the **.shp** file, but alone is incomplete for distribution as the other supporting files are required. The characteristics of each file is the following:

- **.shp**: shape format; the feature geometry itself,
- **.shx**: shape index format; a positional index of the feature geometry to allow seeking forwards and backwards quickly,
- **.dbf**: attribute format; columnar attributes for each shape, in **dBase IV** format.

For simplicity in showing how to create neighbor objects in R, we work on the map consisting of the communes of the Metropolitan Region in Chile.

We first need to load the Metropolitan Region shape file in R. To do so, we will use the **sf** package, which allows us reading and handling spatial objects.

```
#Load package
library("sf")
```

If the shape file `mr_chile.shp` is in the same working directory, then we can load it into R using the command `read_sf`:

```
# Read shape file
mr <- read_sf("mr_chile.shp")
class(mr)

## [1] "sf"          "tbl_df"      "tbl"        "data.frame"
```

The function `read_sf` reads data from the shapefile into an object of class “**sf**”. The function `names` give us the name of the variables in the `.dbf` file associated with the shape file.

```
# Names of the variables in .dbf
names(mr)

## [1] "ID"          "NAME"        "NAME2"       "URB_POP"     "RUR_POP"
## [6] "MALE_POP"    "TOT_POP"     "FEM_POP"     "N_PARKS"     "N_PLAZA"
## [11] "CONS_HOUSE"  "M2_CONS_HA"  "GREEN_AREA"  "AREA"        "POVERTY"
## [16] "PER_CONTR_"  "PER_HON_SA"  "PER_PLANT_"  "NURSES"      "DOCTORS"
## [21] "CONSULT_RU"  "CONSULT_UR"  "POSTAS"      "ESTAB_MUN_"  "PSU_MUN_PR"
## [26] "PSU_PART_P"  "PSU_SUB_PR"  "STUDENT_SU"  "STUDENT_PA"  "STUDENT_MU"
## [31] "geometry"
```

We can plot the shapefile using the generic function `plot` in the following way

```
# Plot shapefile
plot(st_geometry(mr), main = "Metropolitan Region-Chile", axes = TRUE)
```

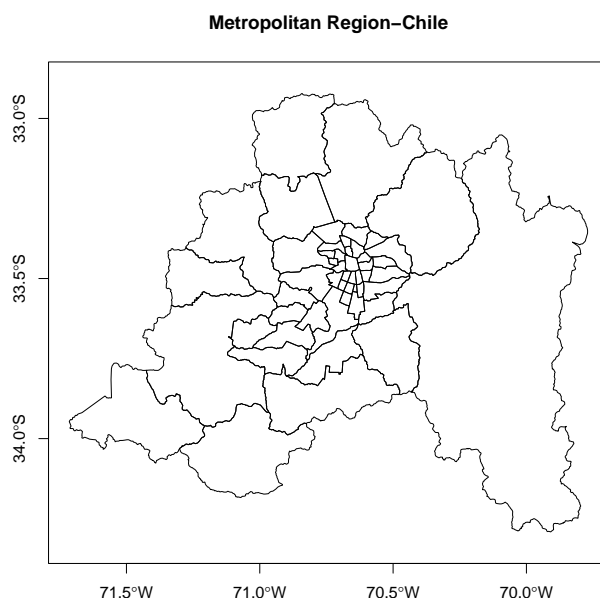
The metropolitan region with the 52 communes is shown in Figure 1.8.

1.3.1 Creating Contiguity Neighbors

To create spatial weight matrices we need the **spdep** package (Bivand et al., 2013). After installing it, we load the package

```
#Load package
library("spdep")
```

Figure 1.8: Plotting a Map in R



In the **spdep** package, neighbor relationships between n observations are represented by an object of class “**nb**”. This object is a list of length n with the index numbers of neighbors of each component recorded as an integer vector. If any observation has no neighbors, the component contains an integer zero.

The function **poly2nb** is used in order to construct weight matrices based on **contiguity**. Specifically, it creates a “neighbors list” based on regions with contiguous boundaries of class “**nb**”. Check out **help(poly2nb)** to see all the details and options.

First, we create a neighbor list based on the ‘queen’ criteria for the communes of the Metropolitan Region:

```
# Create queen W
sf_use_s2(FALSE)
queen.w <- poly2nb(as(mr, "Spatial"), queen = TRUE, row.names = mr$NAME)
```

Since we have an **nb** object to examine, we can present the standard methods for these objects. There are **print**, **summary**, **plot**, and other methods. The characteristics of the weights are obtained with the usual **summary** command:

```
# Summary of W
summary(queen.w)

## Neighbour list object:
## Number of regions: 52
## Number of nonzero links: 292
## Percentage nonzero weights: 10.79882
## Average number of links: 5.615385
```

```
## Link number distribution:
##
##  2  3  4  5  6  7  8  9 10 12
##  3  2  7 15 10 10  2  1  1  1
## 3 least connected regions:
## Tilttil San Pedro Maria Pinto with 2 links
## 1 most connected region:
## San Bernardo with 12 links
```

The output presents important information about the neighbors: it shows the number of regions, which corresponds to 52 communes in this example; the number of nonzero links; the percentage of nonzero weights; the average number of links, and so on.

The commune of San Bernardo is the most connected region with 12 neighbors under the queen scheme. The least connected regions are Tilttil, San Pedro, and Maria Pinto with 2 neighbors each of them. The output also shows the distribution of neighbors. For example, 7 out of 52 regions has 4 neighbors and only 2 communes has 8 neighbors.

We can also show the commune with the largest number of contiguities as follows:

```
# Plot communes with largest number of contiguities
cards <- card(queen.w)
maxconts <- which(cards == max(cards))
fg <- rep("grey", length(cards))
fg[maxconts] <- "red"
fg[queen.w[[maxconts]]] <- "blue"
plot(st_geometry(mr), col = fg)
```

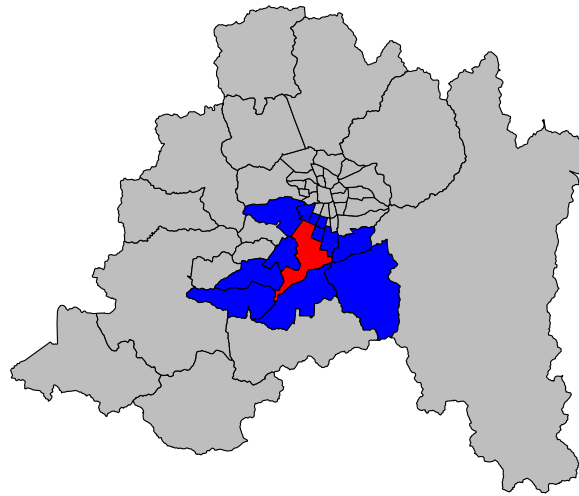
Figure 1.9 shows that the commune in red is the spatial unit (San Bernardo) with the largest number of neighbors based on the queen criteria, whereas the communes in blue are its neighbors.

To transform the `list` into an actual matrix \mathbf{W} , we can use the function `nb2listw`:

```
# From list to matrix
queen.wl <- nb2listw(queen.w, style = "W")
summary(queen.wl)

## Characteristics of weights list object:
## Neighbour list object:
## Number of regions: 52
## Number of nonzero links: 292
## Percentage nonzero weights: 10.79882
## Average number of links: 5.615385
## Link number distribution:
##
##  2  3  4  5  6  7  8  9 10 12
##  3  2  7 15 10 10  2  1  1  1
## 3 least connected regions:
## Tilttil San Pedro Maria Pinto with 2 links
```

Figure 1.9: Commune with largest number of contiguities



Notes: The commune in red is the spatial unit (San Bernardo) with the largest number of neighbors based on the queen criteria, whereas the communes in blue are its neighbors.

```
## 1 most connected region:
## San Bernardo with 12 links
##
## Weights style: W
## Weights constants summary:
##      n   nn S0      S1      S2
## W 52 2704 52 19.76751 216.466
```

An important argument of the function is **style**. This argument indicates what type of matrix to create. For example, **style = "W"** creates a row-standardize matrix so that $w_{ij}^s = w_{ij} / \sum_j w_{ij}$. After normalization, each row of \mathbf{W}^s will sum to 1. "B" is the basic binary coding; and "C" is globally standardize, that is, $w_{ij}^s = w_{ij} \cdot (n / \sum_i \sum_j w_{ij})$. If **style = "U"**, then $w_{ij}^s = w_{ij} / \sum_i \sum_j w_{ij}$. In a **minmax** matrix, the (i, j) th element of \mathbf{W}^s becomes $w_{ij}^s = w_{ij} / \min \{ \max_i(\tau_i), \max_i(c_i) \}$, with $\max_i(\tau_i)$ being the largest row sum of \mathbf{W} and $\max_i(c_i)$ being the largest column sum of \mathbf{W} (Kelejian and Prucha, 2010). Finally, "S" is the variance-stabilizing coding scheme where $w_{ij}^s = w_{ij} / \sqrt{\sum_j w_{ij}^2}$ (Tiefelsdorf et al., 1999).

Furthermore, the **summary** function reports constants used in the inference for global spatial autocorrelation statistics, which we will discuss later.

We can also see the attributes of the object using the function **attributes**:

```
# Attributes of wlist
attributes(queen.w)
```

```
## $class
## [1] "nb"
##
## $region.id
## [1] "Santiago"          "Cerro Navia"
## [4] "Conchali"          "Estacion Central"
## [7] "La Cisterna"       "La Granja"
## [10] "La Pintana"        "Lo Espejo"
## [13] "Lo Prado"          "Nunoa"
## [16] "Pedro Aguirre Cerda" "Providencia"
## [19] "Quinta Normal"     "Renca"
## [22] "San Joaquin"       "San Ramon"
## [25] "Independencia"     "Las Condes"
## [28] "Vitacura"          "Huechuraba"
## [31] "Maipu"             "San Bernardo"
## [34] "Tiltil"            "Colina"
## [37] "Lo Barnechea"      "Paine"
## [40] "Buin"              "Melipilla"
## [43] "San Pedro"         "Curacavi"
## [46] "Penaflor"          "Padre Hurtado"
## [49] "El Monte"          "Isla de Maipo"
## [52] "San Jose de Maipo"
##
## $call
## poly2nb(pl = as(mr, "Spatial"), row.names = mr$NAME, queen = TRUE)
##
## $type
## [1] "queen"
##
## $sym
## [1] TRUE
```

We may ask whether the matrix is symmetric using:

```
# Symmetric W
is.symmetric.nb(queen.w)

## [1] TRUE
```

As we previously discussed, generally weight matrix based on boundaries are symmetric. Now, we construct a binary matrix using the Rook criteria:

```
# Rook W
rook.w <- poly2nb(as(mr, "Spatial"), row.names = mr$NAME, queen = FALSE)
summary(rook.w)

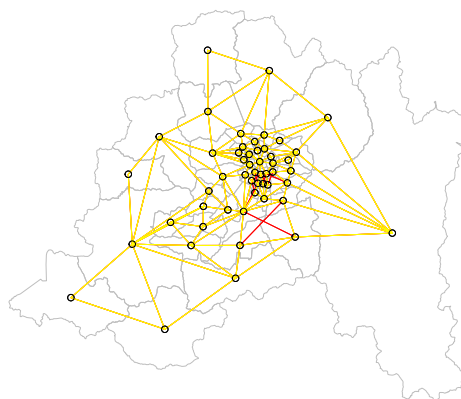
## Neighbour list object:
```

```
## Number of regions: 52
## Number of nonzero links: 272
## Percentage nonzero weights: 10.05917
## Average number of links: 5.230769
## Link number distribution:
##
##  2  3  4  5  6  7  8  9 10
##  3  3 12 16  7  6  2  1  2
## 3 least connected regions:
## Tiltit San Pedro Maria Pinto with 2 links
## 2 most connected regions:
## Santiago San Bernardo with 10 links
```

Finally, we can plot the weight matrices using the following set of commands (see Figure 1.10).

```
# Plot Queen and Rook W Matrices
plot(st_geometry(mr), border = "grey")
coords <- st_coordinates(st_centroid(st_geometry(mr)))
plot(queen.w, coords, add = TRUE, col = "red")
plot(rook.w, coords, add = TRUE, col = "yellow")
```

Figure 1.10: Queen and Rook Criteria for MR



1.3.2 Creating Distance-based Neighbors

We now construct spatial weight matrices using the k -nearest neighbors criteria.

```
# K-neighbors
head(coords, 5) # show coordinates

##           X           Y
## 1 -70.65599 -33.45406
## 2 -70.71742 -33.50027
## 3 -70.74504 -33.42278
## 4 -70.67735 -33.38372
## 5 -70.67640 -33.56294

k1neigh <- knearneigh(coords, k = 1, longlat = TRUE) # 1-nearest neighbor
k2neigh <- knearneigh(coords, k = 2, longlat = TRUE) # 2-nearest neighbor
```

The function `coords` extract the spatial coordinates from the shape file, whereas the function `knearneigh` returns a matrix with the indices of points belonging to the set of the k -nearest neighbors of each other. The argument `k` indicates the number of nearest neighbors to be returned. If point coordinates are longitude-latitude decimal degrees, then distances are measured in kilometers if `longlat = TRUE`. Furthermore, if `longlat = TRUE`, great circle distances are used. Note that the objects `k1neigh` and `k2neigh` are of class `knn`.

Weight matrices based on inverse distance can be computed in the following way (see Section 1.2.2):

```
# Inverse weight matrix
dist.mat <- as.matrix(dist(coords, method = "euclidean"))
dist.mat[1:5, 1:5]

##           1           2           3           4           5
## 1 0.00000000 0.07687010 0.09438408 0.07350782 0.11078109
## 2 0.07687010 0.00000000 0.08226867 0.12324109 0.07489489
## 3 0.09438408 0.08226867 0.00000000 0.07814455 0.15606360
## 4 0.07350782 0.12324109 0.07814455 0.00000000 0.17922003
## 5 0.11078109 0.07489489 0.15606360 0.17922003 0.00000000

dist.mat.inv <- 1 / dist.mat # 1 / d_{ij}
diag(dist.mat.inv) <- 0 # 0 in the diagonal
dist.mat.inv[1:5, 1:5]

##           1           2           3           4           5
## 1 0.000000 13.008960 10.595007 13.603994 9.026811
## 2 13.008960 0.000000 12.155295 8.114177 13.352046
## 3 10.595007 12.155295 0.000000 12.796797 6.407644
## 4 13.603994 8.114177 12.796797 0.000000 5.579733
## 5 9.026811 13.352046 6.407644 5.579733 0.000000

# Standardized inverse weight matrix
dist.mat.inve <- mat2listw(dist.mat.inv, style = "W", row.names = mr$NAME)
summary(dist.mat.inve)
```

```
## Characteristics of weights list object:
## Neighbour list object:
## Number of regions: 52
## Number of nonzero links: 2652
## Percentage nonzero weights: 98.07692
## Average number of links: 51
## Link number distribution:
##
## 51
## 52
## 52 least connected regions:
## Santiago Cerillos Cerro Navia Conchali El Bosque Estacion Central La Cisterna La Flor
## 52 most connected regions:
## Santiago Cerillos Cerro Navia Conchali El Bosque Estacion Central La Cisterna La Flor
##
## Weights style: W
## Weights constants summary:
##      n   nn S0      S1      S2
## W 52 2704 52 2.902384 214.3332
```

The function `dist` from **stats** package computes and returns the distance matrix computed by using the specified distance measure—euclidean distance in this example—to compute the distance between the rows of a data matrix. The other methods that can be used are `maximum`, `manhattan`, `canberra`, `binary` or `minkowski`. Finally, the `mat2listw` function converts a square spatial weight matrix as a sequence of number `1:nrow(x)`.²

The following code plot the different weight matrices:

```
# Plot Weights
par(mfrow = c(3, 2))
plot(st_geometry(mr), border = "grey", main = "Queen")
plot(queen.w, coords, add = TRUE, col = "red")
plot(st_geometry(mr), border = "grey", main = "1-Neigh")
plot(knn2nb(k1neigh), coords, add = TRUE, col = "red")
plot(st_geometry(mr), border = "grey", main = "2-Neigh")
plot(knn2nb(k2neigh), coords, add = TRUE, col = "red")
plot(st_geometry(mr), border = "grey", main = "Inverse Distance")
plot(dist.mat.inve, coords, add = TRUE, col = "red")
```

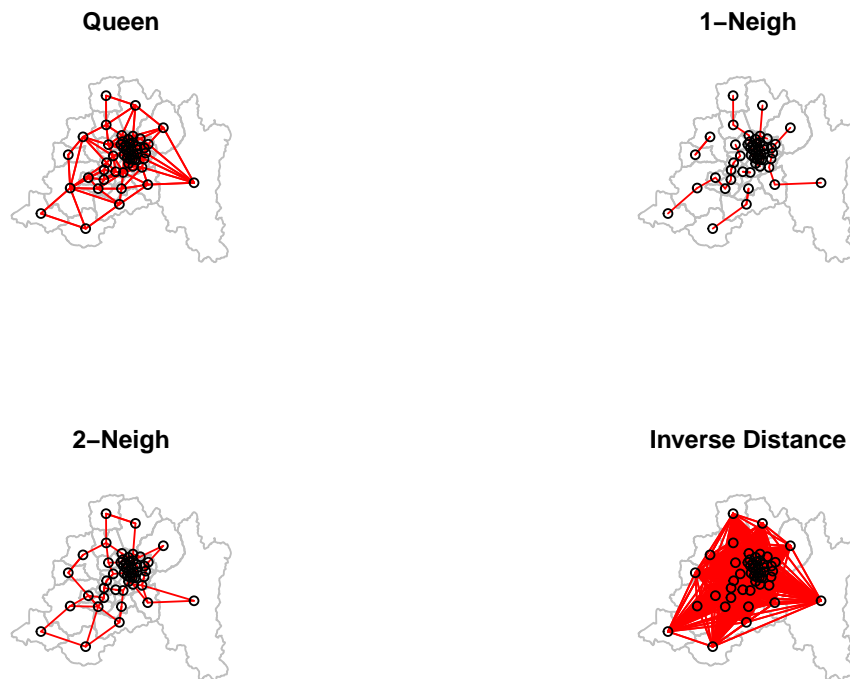
1.3.3 Constructing a Spatially Lagged Variable

Spatially lagged variables are important elements of many spatial test and spatial regression specifications. In **spdep**, they are constructed by means of the `lag.listw` function.

First, we will combine the variables `POVERTY` and `URB_POP` into a matrix and check the contents with `head`

²For more about spatial weight matrices see (Stewart and Zhukov, 2010).

Figure 1.11: Different Spatial Weight Schemes for MR



```
# X matrix
X <- cbind(mr$POVERTY, mr$URB_POP)
head(X, 5)

##      [,1]  [,2]
## [1,]    8 159919
## [2,]    9  65262
## [3,]   18 131850
## [4,]   12 104634
## [5,]   14 166514
```

Now, we can construct a spatially lagged version of this matrix, using the `queen.w` weights:

```
# Create WX
WX <- lag.listw(nb2listw(queen.w), X)
head(WX)

##           [,1]      [,2]
## [1,]  9.10000 100138.9
## [2,] 12.40000 299498.4
## [3,] 14.00000 144756.5
## [4,] 14.60000 121974.2
## [5,] 18.25000 170266.5
## [6,] 10.42857 236231.1
```

1.4 Testing for Spatial Autocorrelation

As we stated in Section 1.1.2, spatial autocorrelation refers to the correlation of a variable with itself in space. It can be positive (when high values correlate with high neighboring values or when low values correlate with low neighboring values) or negative (spatial outlier for high-low or low-high values). So the next question is how to test whether the spatial pattern we observe truly follows a spatial autocorrelated process or is completely random. In other words, we need a test of spatial autocorrelation to formally examine whether the observed value of a variable at one location is independent of values of that variable at neighboring locations.

1.4.1 Global Spatial Autocorrelation: Moran's I

Global spatial autocorrelation is a measure of overall clustering. So the main goal of these indices is to summarize the degree to which similar observations tend to occur near each other. Those indices calculate the similarity of values at location i and j then ‘weight’ the similarity by the proximity of locations i and j . High similarities with high weight indicate similar values that are close together, whereas low similarities with high weight indicate dissimilar values that are close together.

The most general measure used is the Moran's I.³ This statistic is a measure of overall clustering that exists in a dataset. It is assessed by means of a test of a null hypothesis of random location. Therefore, rejection of this null hypothesis suggests a spatial pattern or spatial structure.

Moran's I is given by:

$$I = \frac{\sum_{i=1}^n \sum_{j=1, j \neq i}^n w_{ij} (x_i - \bar{x}) (x_j - \bar{x})}{S_0 \sum_{i=1}^n (x_i - \bar{x})^2 / n} = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{ij} (x_i - \bar{x}) (x_j - \bar{x})}{S_0 \sum_{i=1}^n (x_i - \bar{x})^2}, \quad (1.4)$$

where $S_0 = \sum_{i=1}^n \sum_{j=1}^n w_{ij}$ and w_{ij} is an element of the spatial weight matrix that measures spatial distance or connectivity between regions i and j . In matrix form:

³There exists several other measures of global spatial autocorrelation as the Geary's C test. But, in this notes we will focus only in the Moran's I.

$$I = \frac{n}{S_0} \frac{\mathbf{z}^\top \mathbf{W} \mathbf{z}}{\mathbf{z}^\top \mathbf{z}},$$

where:

$$\mathbf{z} = \begin{pmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{pmatrix},$$

If the \mathbf{W} matrix is row standardized, then:

$$I = \frac{\mathbf{z}^\top \mathbf{W}^s \mathbf{z}}{\mathbf{z}^\top \mathbf{z}},$$

because $S_0 = n$. Values range from -1 (perfect dispersion) to +1 (perfect correlation). A zero value indicates a random spatial pattern.

A very useful tool for understanding the Moran's I test is the Moran Scatterplot. The idea of the Moran scatterplot is to display the variable for each region (on the horizontal axis) against the standardized spatial weighted average (average of the neighbors' x , also called spatial lag) on the vertical axis (See Figure 1.12). As pointed out by Anselin (1996), expressing the variables in standardized form (i.e. with mean zero and standard deviation equal to one) allows assessment of both the global spatial association, since the slope of the line is the Moran's I coefficient, and local spatial association (the quadrant in the scatterplot). The Moran scatterplot is therefore divided into four different quadrants corresponding to the four types of local spatial association between a region and its neighbors:

- Quadrant I displays the region with high x (above the average) surrounded by regions with high x (above the average). This quadrant is usually denoted High-High.
- Quadrant II show the regions with low value surrounded by region with high values. This quadrant is usually denoted Low-High.
- Quadrant III display the regions with low value surrounded by regions with low values, and is denoted Low-Low.
- Quadrant IV shows the regions with high value surrounded by regions with low values. It is noted High-Low.

Regions located in quadrant I and III refer to positive spatial autocorrelation, the spatial clustering of similar values, whereas quadrant II and IV represent negative spatial autocorrelation, the spatial clustering of dissimilar values.

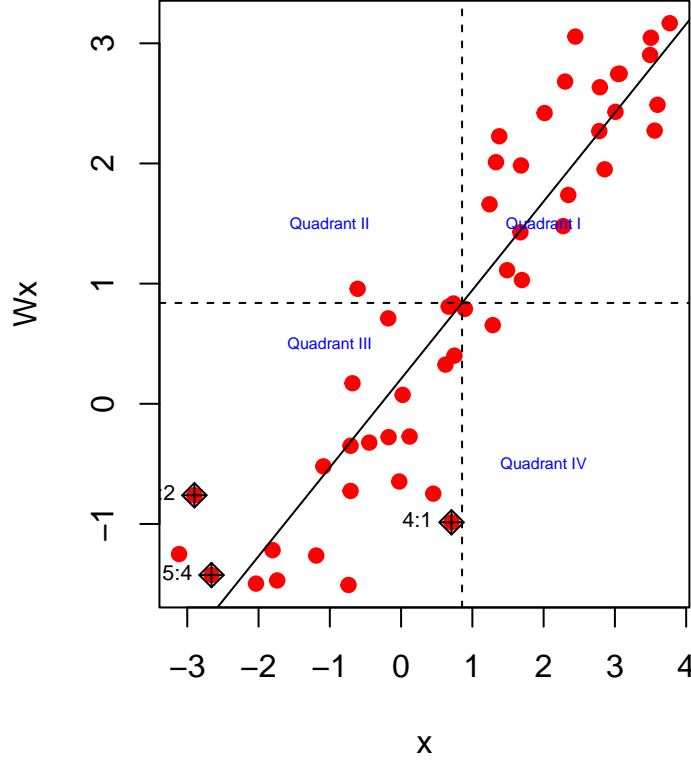
To understand Moran's I, it is important to note the similarity of the Moran's I with the OLS coefficient. Recall that

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (1.5)$$

Then looking at (1.4), Moran's I is equivalent to the slope coefficient of a linear regression of the spatial lag $\mathbf{W}\mathbf{x}$ on the observation vector \mathbf{x} measured in deviation from their mean. It is, however, not equivalent to the slope of \mathbf{x} on $\mathbf{W}\mathbf{x}$ which would be a more natural way.

The hypothesis tested by the Moran's I is the following:

Figure 1.12: Moran Scatterplot



- H_0 : \mathbf{x} is spatially independent; the observed \mathbf{x} is assigned at random among locations. In this case I is close to zero.
- H_1 : \mathbf{x} is not spatially independent. In this case I is statistically different from zero.

What is the distribution of the Moran's I ? We are interested in the distribution of:

$$\frac{I - \mathbb{E}[I]}{\sqrt{\mathbb{V}(I)}}$$

There are two ways to compute the mean and variance of Moran's I . The first one is under the normal assumption of x_i and the second one is under randomization of x_i . Under the normal assumption, it is assumed that the random variable x_i are the result of n independently drawings from a normal population. Under the randomization assumption, no matter what the underlying distribution of the populations, we consider the observed values of x_i were repeatedly randomly permuted.

Moments Under Normality Assumption

Theorem 1.2 gives the moments of Moran's I under normality.

Theorem 1.2 — Moran's I Under Normality. Assume that $\{\mathbf{x}_i\} = \{x_1, x_2, \dots, x_n\}$ are independent and distributed as $N(\mu, \sigma^2)$, but μ and σ^2 are unknown. Then:

$$\mathbb{E}(I) = -\frac{1}{n-1} \quad (1.6)$$

and

$$\mathbb{E}(I^2) = \frac{n^2 S_1 - n S_2 + 3 S_0^2}{S_0^2 (n^2 - 1)} \quad (1.7)$$

where $S_0 = \sum_{i=1}^n \sum_{j=1}^n w_{ij}$, $S_1 = \sum_{i=1}^n \sum_{j=1}^n (w_{ij} + w_{ji})^2 / 2$, $S_2 = \sum_{i=1}^n (w_{i.} + w_{.i})^2$, where $w_{i.} = \sum_{j=1}^n w_{ij}$ and $w_{.i} = \sum_{j=1}^n w_{ji}$. Then:

$$\mathbb{V}(I) = \mathbb{E}(I^2) - \mathbb{E}(I)^2 \quad (1.8)$$

Proof. Let $z_i = x_i - \bar{x}$. The following moments are true for z_i :

$$\begin{aligned} \mathbb{E}[z_i] &= 0 \\ \mathbb{E}[z_i^2] &= \sigma^2 - \frac{\sigma^2}{n} \\ \mathbb{E}[z_i z_j] &= -\frac{\sigma^2}{n} \\ \mathbb{E}[z_i^2 z_j^2] &= \frac{(n^2 - 2n + 3)\sigma^2}{n^2} \\ \mathbb{E}[z_i^2 z_j z_k] &= -\frac{(n-3)\sigma^4}{n} \\ \mathbb{E}[z_i z_j z_k z_l] &= \frac{3\sigma^4}{n^2} \end{aligned}$$

Then:

$$\begin{aligned} \mathbb{E}[I] &= \frac{n}{S_0} \frac{\mathbb{E}[\sum_{i=1}^n \sum_{j=1}^n w_{ij} z_i z_j]}{\mathbb{E}[\sum_{i=1}^n z_i^2]} = \frac{n}{S_0} \sum_{i=1}^n \sum_{j=1}^n w_{ij} \frac{\mathbb{E}[z_i z_j]}{\sum_{i=1}^n \mathbb{E}[z_i^2]} \\ &= \frac{-n S_0 \frac{\sigma^2}{n}}{S_0 n (1 - 1/n) \sigma^2} \\ &= -\frac{\frac{\sigma^2}{n}}{(1 - 1/n) \sigma^2} \\ &= -\frac{1}{n-1} \end{aligned} \quad (1.9)$$

and

$$\begin{aligned}
\mathbb{E}[I^2] &= \mathbb{E}\left[\frac{n^2}{S_0^2} \frac{\left[\sum_{i=1}^n \sum_{j=1}^n w_{ij} z_i z_j\right]^2}{\left[\sum_{i=1}^n z_i^2\right]^2}\right] \\
&= \frac{n^2}{S_0^2} \mathbb{E}\left[\frac{1/2 \sum_{(2)} (w_{ij} + w_{ji})^2 z_i^2 z_j^2 + \sum_{(3)} (w_{ij} + w_{ji})(w_{ik} + w_{ki}) z_i^2 z_j z_k + \sum_{(4)} w_{ij} w_{kl} z_i z_j z_k z_l}{s}\right]
\end{aligned} \tag{1.10}$$

■

Moran's I under Randomization

Theorem 1.3 gives the moments of Moran's I under randomization.

Theorem 1.3 — Moran's I Under Randomization. Under permutation, we have:

$$\mathbb{E}(I) = -\frac{1}{n-1} \tag{1.11}$$

and

$$\mathbb{E}(I^2) = \frac{n[(n^2 - 3n + 3)S_1 - nS_2 + 3S_0^2] - b_2[(n^2 - n)S_1 - 2nS_2 + 6S_0^2]}{(n-1)(n-2)(n-3)S_0^2} \tag{1.12}$$

where $S_0 = \sum_{i=1}^n \sum_{j=1}^n w_{ij}$, $S_1 = \sum_{i=1}^n \sum_{j=1}^n (w_{ij} + w_{ji})^2/2$, $S_2 = \sum_{i=1}^n (w_{i.} + w_{.i})^2$, where $w_{i.} = \sum_{j=1}^n w_{ij}$ and $w_{.i} = \sum_{j=1}^n w_{ji}$. Then:

$$\mathbb{V}(I) = \mathbb{E}(I^2) - \mathbb{E}(I)^2 \tag{1.13}$$

It is important to note that the expected value of Moran's I under normality and randomization is the same.

Monte Carlo Moran's I

The normality assumption is a very strong assumption. However we can use the Moran's I test based on Monte Carlo simulation.

The idea for any Monte Carlo test is the following:

- To test a null hypothesis H_0 (no spatial autocorrelation in our case), we specify a test statistic T such that large values of T are evidence against H_0 .
- Let T have observed value t_{obs} . We generally want to compute

$$p\text{-value} = \Pr(T \geq t_{obs} | H_0) \tag{1.14}$$

Therefore we need the distribution of T when H_0 is true to evaluate this probability.

The algorithm for the Morans' I Monte Carlo test is the following:

Algorithm 1.4 — Moran's' I Monte Carlo Test. The procedure is the following:

- (a) Rearrange the spatial data by shuffling their location and compute the Moran's I S times. This will create the distribution under H_0 . This operationalizes spatial randomness.
- (b) Let $I_1^*, I_2^*, \dots, I_S^*$ be the Moran's I for each time. A consistent Monte Carlo p-value is then:

$$\hat{p} = \frac{1 + \sum_{s=1}^S 1(I_s^* \geq I_{obs})}{S + 1} \quad (1.15)$$

- (c) For tests at the α level or at $100(1 - \alpha)\%$ confidence intervals, there are reasons for choosing S so that $\alpha(S + 1)$ is an integer. For example, use $S = 999$ for confidence intervals and hypothesis tests when $\alpha = 0.05$.

1.5 Application: Poverty in Santiago, Chile

In this section we undertake an exploratory spatial data analysis (ESDA) for poverty in Metropolitan Region, Chile.

1.5.1 Cloropleth Graphs

If we are interested in the geographical variation in poverty, we should start by plotting the spatial distribution of poverty. This can be useful in a variety of ways. Usually, aggregate or national level indicators hide important differences between different spatial units. Thus, poverty mapping helps to highlight geographical variations. In addition to this, another advantage of poverty maps is their legibility- maps are powerful tools for representing complex information in a visual format that is easy to understand.

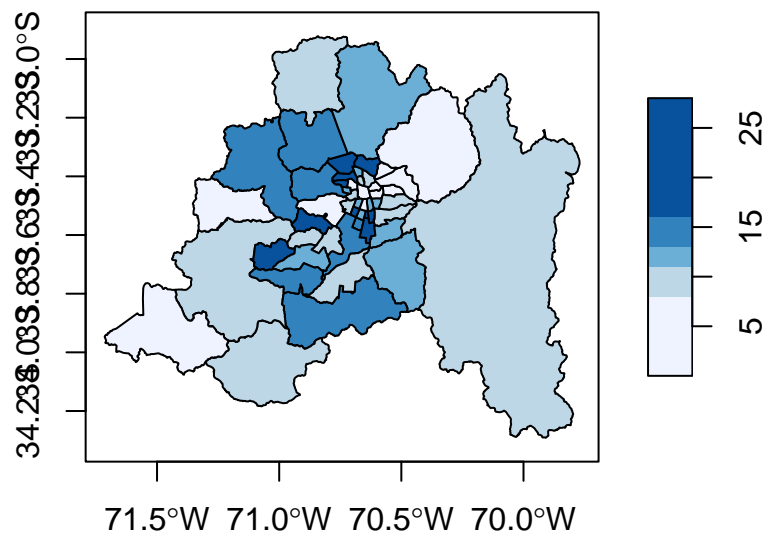
So, we start by plotting the geographical variation of poverty among communes by using the `plot` function. In particular, we use a cloropleth⁴ map using the quantile classification. In a quantile graph, the variable is sorted and grouped in categories with equal number of observations, or quantiles.

```
# Cloropleth graphs ----
library("RColorBrewer")
plot(mr["POVERTY"],
     breaks = "quantile",
     nbreaks = 5,
     pal = brewer.pal(5, "Blues"),
     main = "",
     axes = TRUE)
```

Figure 1.13 provides some useful insights. First, it clearly shows that the spatial pattern of poverty in the MR is not spatially homogeneous, but rather the intensity of poverty varies across space. Second, it provides an example of how disaggregated poverty indicators can

⁴The name of this technique is derived from the Greek words *choros* - space, and *pleth* - value

Figure 1.13: Choropleth map: Poverty in the Metropolitan Region



reveal additional information to aggregate indicators. It shows that poverty intensity is lower peripheral communes than central communes.

How to interpret quantile maps? A quantile classification scheme is an ordinal ranking of the data values, dividing the distribution into intervals that have an equal number of data values. Quantile classification ensures maps are easily comparable and can be ‘easy to read’.

However, regarding the possible spatial association that seems to be derived from the above figure for the poverty variable, it is necessary to note that the results are sensitive to the number of defined intervals (among other things). Therefore, it is necessary to conduct a comprehensive and formal analysis about the potential presence of spatial dependence to ascertain whether there exists a pattern of statistically significant spatial autocorrelation in the spatial distribution of poverty. That is why now we calculate the Moran’s I test.

1.5.2 Moran’s I Test

First, we create two spatial weight matrices (queen and rook) to assess the robustness of the test under different spatial schemes.

```
# Generate W matrices
queen.w <- poly2nb(as(mr, "Spatial"), row.names = mr$NAME, queen = TRUE)
rook.w   <- poly2nb(as(mr, "Spatial"), row.names = mr$NAME, queen = FALSE)
```

Moran’s I test statistic for spatial autocorrelation is implemented in **spdep** (Bivand and Piras, 2015). There are mainly two function for computing this test: `moran.test`, where the inference is based on a normal or randomization assumption, and `moran.mc`, for a permutation-based test.

```
# Moran's I test
moran.test(mr$POVERTY, listw = nb2listw(queen.w), randomisation = FALSE,
           alternative = 'two.sided')
```




```
##
## Moran I test under normality
##
## data:  mr$POVERTY
## weights: nb2listw(queen.w)
##
## Moran I statistic standard deviate = 4.0453, p-value = 5.225e-05
## alternative hypothesis: two.sided
## sample estimates:
## Moran I statistic      Expectation      Variance
##      0.306497992      -0.019607843      0.006498517

moran.test(mr$POVERTY, listw = nb2listw(rook.w), randomisation = FALSE,
           alternative = 'two.sided')

##
## Moran I test under normality
##
## data:  mr$POVERTY
## weights: nb2listw(rook.w)
##
## Moran I statistic standard deviate = 4.3309, p-value = 1.485e-05
## alternative hypothesis: two.sided
## sample estimates:
## Moran I statistic      Expectation      Variance
##      0.342282943      -0.019607843      0.006982432
```

The `randomisation` option is set to `TRUE` by default, which implies that in order to get inference based on a normal approximation, it must be explicitly set to `FALSE`, as in our case. Similarly, the default is a one-sided test, so that in order to obtain the results for the more commonly used two-sided test, the option `alternative` must be explicitly to `'two.sided'`. Note also that the `zero.policy` option is set to `FALSE` by default, which means that islands result in a missing value code `NA`. Setting this option to `TRUE` will set the spatial lag for island to the customary zero value.

The results show that the Moran's I statistic are ≈ 0.30 and 0.34 , respectively, and highly significant. This implies that there is evidence of robust **positive spatial autocorrelation** in the poverty variable (since we are rejecting the null hypothesis of random spatial distribution).

-  If you compute the Moran's I test for two different variables, but using the same spatial weight matrix, the expectation and variance of the Moran's I test statistic will be the same under the normal approximation. Why?

The test under randomization gives the following results:

```
# Moran test under randomization
moran.test(mr$POVERTY, listw = nb2listw(queen.w),
           alternative = 'two.sided')

##
##  Moran I test under randomisation
##
## data:  mr$POVERTY
## weights: nb2listw(queen.w)
##
## Moran I statistic standard deviate = 4.0689, p-value = 4.723e-05
## alternative hypothesis: two.sided
## sample estimates:
## Moran I statistic      Expectation      Variance
##      0.306497992      -0.019607843      0.006423226
```

Note how the value of the statistic and its expectation do not change relative to the normal case, only the variance is different.

We can carry out a Moran's I test based on random permutation the function `moran.mc`. Unlike previous test, it needs the number of permutations `nsim`. Since the rank of the observed statistic is computed relative to the reference distribution of statistics for the permuted data sets, it is good practice to set this number to something ending on 9 (such as 99 or 999). This will lead to rounded pseudo p-values like 0.01 or 0.001.

```
# Moran's Test
set.seed(1234)
moran.mc(mr$POVERTY, listw = nb2listw(queen.w),
         nsim = 99)

##
##  Monte-Carlo simulation of Moran I
##
## data:  mr$POVERTY
## weights: nb2listw(queen.w)
## number of simulations + 1: 100
##
## statistic = 0.3065, observed rank = 100, p-value = 0.01
## alternative hypothesis: greater
```

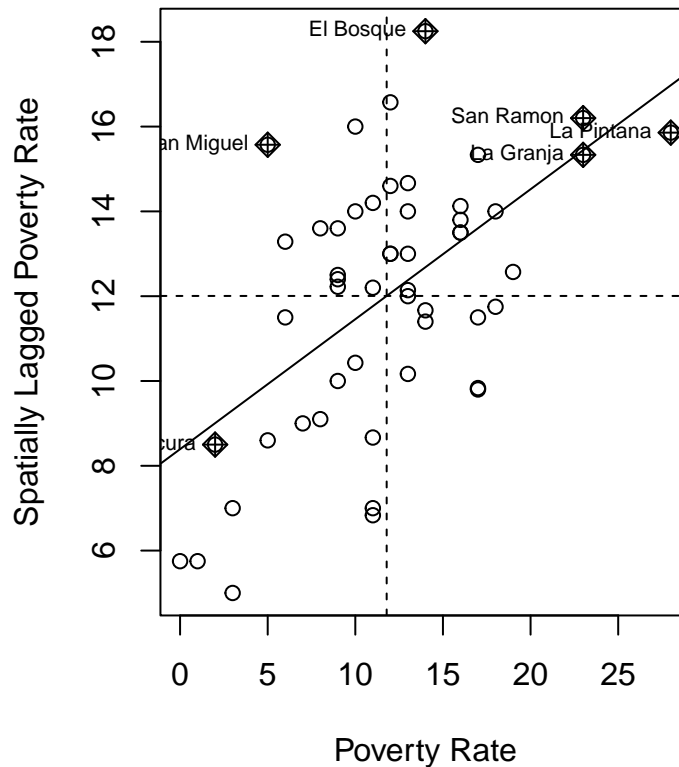
Note that none of the permuted data sets yielded a Moran's I greater than the observed value of 0.3065, hence a pseudo p-value of $(0 + 1)/(99 + 1) = 0.01$.

The Moran scatter plot can also be obtained using the function `moran.plot` of **spdep**:

```
# Moran's plot
moran.plot(mr$POVERTY, listw = nb2listw(queen.w))
```

Figure 1.14 displays the Moran scatterplot of poverty with the queen weight matrix. Positive spatial autocorrelation, detected by the value of the Moran's I , is reflected by the

Figure 1.14: Moran Plot for Poverty



fact that most of the communes are located in quadrant I and III. However, there are some exceptions such as the communes located in quadrant II and IV. For example, San Miguel is a commune with low poverty rate, but surrounded by communes with high poverty.

A major limitation of Moran's I is that it cannot provide information on the specific locations of spatial patterns; it only indicates the presence of spatial autocorrelation globally. A single overall indication is given of whether spatial autocorrelation exists in the dataset, but no indication is given of whether local variations exist in spatial autocorrelation (e.g., concentrations, outliers) across the spatial extent of the data.

1.6 Exercises

Exercise 1.1 Another method used for creating spatial weight matrices in Monte Carlo studies is the “ k -ahead and k -behind” criterion in a circular world. (This was introduced by [Kelejian and Prucha \(1999\)](#)). In this approach, each spatial unit is assumed to have k neighbors which are ahead of it in the order of sample, and k units which are behind it. The number k is typically chosen to be small relative to the sample size. Thus, each spatial unit has $2k$ neighbors. Weighting matrices which are built on this framework are typically row normalized, and all of the nonzero elements in the matrix are $1/(2k)$. Suppose $n = 10$ and

$k = 2$. Specify the third row of the 10×10 weighting matrix.

Exercise 1.2 For a general sample size, say n , which corresponds to a checkerboard of squares, what is the minimum number of neighbors a unit can have if the weighting matrix is based on a queen pattern?

Exercise 1.3 Let INC_r the income per capita in cross-sectional unit $r = 1, \dots, n$. Consider the following specification for w_{ij} :

$$w_{ij} = \alpha \left[1 - \frac{|INC_i - INC_j|}{INC_i + INC_j} \right],$$

where α is some pre-selected positive constant. Show that α will cancel if the weight matrix is row-normalized.

Exercise 1.4 Create in R your own function to plot a Moran Scatterplot. Show that your function works well using a simulated example.

Spatial Models

Previously, we have reviewed some preliminary concepts in spatial econometric, such as spatial dependency and spatial autocorrelation. In this chapter we will analyze the formulation of spatial models. In particular, in Section 2.1 we will derive a complete taxonomy of spatial models including the Spatial Lag Model, Spatial Durbin Model, Spatial Error Model and the Spatial Autocorrelation Model. We will give a brief motivation of each of them and some examples. In Section 2.3 we show how to understand the ‘spillover’ effects and how to interpret marginal effects in the spatial model framework.

2.1 Taxonomy of Models

As we showed in previous chapter (in particular in Section 1.1.1), it is not possible to model spatial dependencies using traditional econometrics. The main problem is that we have more parameters than observations. However, using the spatial weight matrix we can solve that problem by reducing the number of parameters to just one parameter using the weighted average of the y values in the neighborhood of i .

2.1.1 Spatial Lag Model

So, given the problem of insufficient degree of freedom, how to model a situation where the dependent variable depends also of the spatially lagged variable? Instead of using a full system of equation, we can model the spatial dependency as:

$$y_i = \alpha + \rho \sum_{j=1}^n w_{ij} y_j + \epsilon_i, \quad i = 1, \dots, n, \quad (2.1)$$

where w_{ij} is the (i, j) th element of the spatial weight matrix \mathbf{W} matrix (see Definition 1.2.1); y_i is the dependent variable for spatial unit i , so that $\sum_{j=1}^n w_{ij} y_j$ is the weighted average of the dependent variable for the neighbors of i (or spatial lag); ϵ_i is the error term such that $\mathbb{E}(\epsilon_i) = 0$; and ρ is the spatial autoregressive parameter which measures the intensity of the spatial interdependence: $\rho > 0$ indicates a positive spatial dependence, whereas $\rho < 0$ indicates a negative spatial dependence. It should be clear that if $\rho = 0$, then we have the traditional linear regression model. By including a spatially lagged variable we are making explicit the existence of spatial spillovers effects due to, for example,

geographical proximity. This data generating process is known as a *Spatial Autoregressive Process* or also labeled as SAR or *Spatial Lag Model* SLM. Since the model (2.1) does not include explanatory variables the model is known as the pure SLM or SAR model.

Figure 2.1: The SLM for two tegions

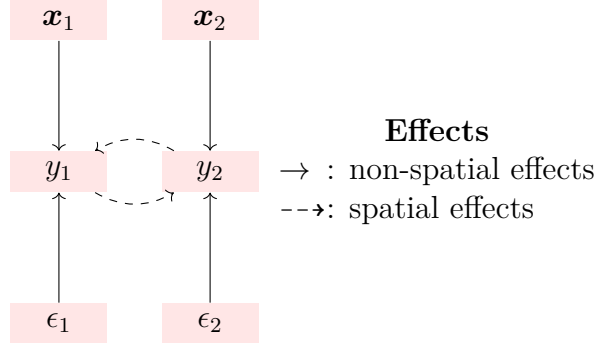


Figure 2.1 represents the spatial autoregressive model in (2.1) for two regions. The variables (x_1, x_2) and unobserved terms (ϵ_1, ϵ_2) have a direct effect on y for both regions. Note that the model incorporates spatial spillover effects by the effect of y_1 on y_2 and vice versa. That is, the model reflects the ‘*simultaneity*’ inherent in spatial autocorrelation.

The model can also be written in vector form as

$$y_i = \alpha + \rho \underset{(1 \times n)}{\mathbf{w}_i}^\top \underset{(n \times 1)}{\mathbf{y}} + \epsilon_i, \quad i = 1, \dots, n,$$

where \mathbf{w}_i is the i th row of \mathbf{W} . A full SLM specification with covariates in matrix form can be written as:

$$\mathbf{y} = \alpha \mathbf{1}_n + \rho \mathbf{W} \mathbf{y} + \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad (2.2)$$

where \mathbf{y} is a $n \times 1$ vector of observations on the dependent variable; \mathbf{X} is an $n \times k$ matrix of observations on the explanatory variables; $\boldsymbol{\beta}$ is the $k \times 1$ vector of parameters and α is the constant; and $\mathbf{1}_n$ is a $n \times 1$ vector of ones.

Reduced Form and Parameter Space

An important concept in the context of spatial models is the difference between structural and reduced form model. Roughly, the reduced form of a model is the one which the endogenous variables are expressed as functions of the exogenous variables. The structural form is the ‘behavioral model’ that relates the variables. For example, Equation (2.2) is the structural model for the SLM, which relates the all the (exogenous and endogenous) variables with the dependent variable \mathbf{y} . However, a better way of think about the model is how the dependent variable is generated. This is the so-called data generating process (DGP). If we solve the system (2.2) for the endogenous variables, \mathbf{y} , we will obtain the reduced-form model. Thus, the implied DGP or “reduced form equation” for the SLM given in Equation (2.2) is:

$$\mathbf{y} = (\mathbf{I}_n - \rho \mathbf{W})^{-1} (\alpha \mathbf{1}_n + \mathbf{X} \boldsymbol{\beta}) + (\mathbf{I}_n - \rho \mathbf{W})^{-1} \boldsymbol{\epsilon}, \quad (2.3)$$

which no longer contains any spatially lagged dependent variable on the right-hand size. The Equation (2.3) expresses the simultaneous nature of the spatial autoregressive process.

- R** The **reduced form** of a system of equations is the result of solving the system for the **endogenous variables**. This gives the latter as functions of the exogenous variables, if any. For example, the general expression of a structural form is $f(\mathbf{y}, \mathbf{X}, \boldsymbol{\varepsilon}) = \mathbf{0}$, whereas the reduced form of this model is given by $\mathbf{y} = g(\mathbf{X}, \boldsymbol{\varepsilon})$, with g as function.

Without restrictions on $(\mathbf{I}_n - \rho \mathbf{W})$ —and $(\alpha \mathbf{z}_n + \mathbf{X} \boldsymbol{\beta})$ —the coefficients cannot be identified from data. In other words, in order to obtain the reduced form we need $(\mathbf{I}_n - \rho \mathbf{W})$ to be invertible. From standard algebra theory a matrix \mathbf{A} is invertible if $\det(\mathbf{A}) \neq 0$. Thus, we require that $\det(\mathbf{I}_n - \rho \mathbf{W}) \neq 0$. The question is: which values of ρ lead to non-singular $(\mathbf{I}_n - \rho \mathbf{W})$? For **symmetric** matrices, the compact open interval for $\rho \in (\omega_{min}^{-1}, \omega_{max}^{-1})$ will lead to a symmetric positive definite $(\mathbf{I}_n - \rho \mathbf{W})$, where ω_{min} and ω_{max} are the minimum and maximum eigen value of \mathbf{W} , respectively. This gives rise to the following Theorem:

Theorem 2.1 — Invertibility. Let \mathbf{W} be a weighting matrix, such that $w_{ii} = 0$ for all $i = 1, \dots, n$, and assume that all of the roots of \mathbf{W} are real. Assume also that \mathbf{W} is not row normalized. Let ω_{min} and ω_{max} be the minimum and maximum eigen value of \mathbf{W} . Assume also that $\omega_{max} > 0$ and $\omega_{min} < 0$. Then $(\mathbf{I}_n - \rho \mathbf{W})$ is nonsingular for all:

$$\omega_{min}^{-1} < \rho < \omega_{max}^{-1}$$

- R** The roots of a non-symmetric matrix will typically not all be real, e.g., some will be complex

Recall that for ease of interpretation, it is common practice to normalize \mathbf{W} such that the elements of each row sum to unity. Since \mathbf{W} is nonnegative, this ensures that all weights are between 0 and 1, and has the effect that the weighting operation can be interpreted as an averaging of neighboring values.

According to our Theorem 1.1 (Eigenvalues of row-stochastic Matrix) the eigenvalues of the row-stochastic (i.e., row-normalized) neighborhood matrix \mathbf{W} are in the range $[-1, +1]$. In this case $\rho \in (-1, 1)$, however it is misleading to consider ρ as a conventional correlation coefficient vector between \mathbf{y} and its spatial lag $\mathbf{W}\mathbf{y}$. This is only the result of considering the standard row-standardized matrix. Other standardization methods will lead to other potential parameter space of ρ .

Theorem 2.2 — Invertibility of Row-Normalized \mathbf{W} matrix. If \mathbf{W} is row-normalized, then $(\mathbf{I}_n - \rho \mathbf{W})^{-1}$ exists for all $|\rho| < 1$

In spite of its popularity, row-normalized weighting has its drawbacks. As we suggested in the remark in Section 1.2.4, row normalization alters the internal weighting structure of \mathbf{W} so that comparisons between rows become somewhat problematic. In view of this limitation, it is natural to consider simple scalar normalization which multiply \mathbf{W} by a single number, say $a \cdot \mathbf{W}$, which removes any measure-unit effect but preserves relations between all rows of \mathbf{W} .

In particular, let

$$\begin{aligned}
a &= \min \{r, c\} \\
r &= \max_i \sum_j |w_{ij}| \quad \text{maximal row sum of the absolute values} \\
c &= \max_j \sum_i |w_{ij}| \quad \text{maximal column sum of the absolute values.}
\end{aligned}$$

Then, assuming that the elements of \mathbf{W} are nonnegative, $(\mathbf{I}_n - \rho \mathbf{W})$ will be nonsingular for all $|\rho| < 1/a$. Note that this normalization has the advantage of ensuring that the resulting spatial weights, w_{ij} , are all between 0 and 1, and hence can still be interpreted as relative influence intensities. This could be taken as the parameter space.

This is an important result because a model which has a weighting matrix which is not row normalized can always be normalized in such a way that the inverse needed to solve the model will exist in an easily established region.

R For further details on normalizing \mathbf{W} and the parameter space of ρ see [Elhorst \(2014, section 2.4\)](#) and [Kelejian and Prucha \(2010, section 2.2\)](#)

When $|\rho| < 1$, $(\mathbf{I}_n - \rho \mathbf{W})^{-1}$ implies an infinite series also called the Leontief expansion. An approximation for this series is given in the following Lemma.

Lemma 2.3 — Leontief Expansion. If $|\rho| < 1$, then

$$(\mathbf{I} - \rho \mathbf{W})^{-1} = \sum_{i=0}^{\infty} (\rho \mathbf{W})^i$$

Then, using Lemma 2.3 (Leontief Expansion), the reduced model in Equation (2.3) can be written as:

$$\begin{aligned}
\mathbf{y} &= (\mathbf{I}_n + \rho \mathbf{W} + \rho^2 \mathbf{W}^2 + \dots) (\alpha \mathbf{z}_n + \mathbf{X} \beta) + (\mathbf{I}_n + \rho \mathbf{W} + \rho^2 \mathbf{W}^2 + \dots) \boldsymbol{\varepsilon}, \\
&= \alpha \mathbf{z}_n + \rho \mathbf{W} \mathbf{z}_n \alpha + \rho^2 \mathbf{W}^2 \mathbf{z}_n \alpha + \dots + \mathbf{X} \beta + \rho \mathbf{W} \mathbf{X} \beta + \rho^2 \mathbf{W}^2 \mathbf{X} \beta + \dots \\
&\quad + \boldsymbol{\varepsilon} + \rho \mathbf{W} \boldsymbol{\varepsilon} + \rho^2 \mathbf{W}^2 \boldsymbol{\varepsilon}.
\end{aligned} \tag{2.4}$$

Expression (2.4) can be simplified since the infinite series:

$$\alpha \mathbf{z}_n + \rho \mathbf{W} \mathbf{z}_n \alpha + \rho^2 \mathbf{W}^2 \mathbf{z}_n \alpha + \dots \rightarrow \frac{\mathbf{z}_n \alpha}{(1 - \rho)},$$

since α is a scalar, the parameter $|\rho| < 1$, and \mathbf{W} is row-stochastic. By definition $\mathbf{W} \mathbf{z}_n = \mathbf{z}_n$ and therefore $\mathbf{W} \mathbf{W} \mathbf{z}_n = \mathbf{W} \mathbf{z}_n = \mathbf{z}_n$. Consequently, $\mathbf{W}^l \mathbf{z}_n = \mathbf{z}_n$ for $l \geq 0$ (recall that $\mathbf{W}^0 = \mathbf{I}_n$). This allows us to write:

$$\mathbf{y} = \frac{1}{(1 - \rho)} \mathbf{z}_n \alpha + \mathbf{X} \beta + \rho \mathbf{W} \mathbf{X} \beta + \rho^2 \mathbf{W}^2 \mathbf{X} \beta + \dots + \boldsymbol{\varepsilon} + \rho \mathbf{W} \boldsymbol{\varepsilon} + \rho^2 \mathbf{W}^2 \boldsymbol{\varepsilon} + \dots$$

This expression allows defining two effects: a multiplier effect affecting the explanatory variables and a spatial diffusion effect affecting the error terms. With respect to the explanatory variables, this expression means that, on average, the value of \mathbf{y} at one location i is not only explained by the values of the explanatory variables associated to this location

but also by those associated to all other locations (neighbors or not) via the inverse spatial transformation $(\mathbf{I}_n - \rho \mathbf{W})^{-1}$. This spatial multiplier effect decreases with distance. This can be seen if we consider the powers of \mathbf{W} in the series expansion of $(\mathbf{I}_n - \rho \mathbf{W})^{-1}$.

With respect to the error process, this expression means that a random (unobserved) shock in a location i not only affects the value of y in this location but also has an impact on the values of y in all other locations via the same spatial inverse transformation. To see this, recall that \mathbf{W}^2 will reflect second-order contiguous neighbors, those that are neighbors to the first-order neighbors (review Section 1.2.5). Since the neighbor of the neighbor (second-order neighbor) to an observation i includes observation i itself, \mathbf{W}^2 has positive elements on the diagonal when each observations has at least one neighbor. That is, higher-order spatial lags can lead to a connectivity relation for an observations i such that $\mathbf{W}\boldsymbol{\varepsilon}$ will extract observations from the vector $\boldsymbol{\varepsilon}$ that point back to the observation i itself. This implies that there exists a simultaneous feedback. This diffusion effect also declines with distance. We will explore this mechanism more deeply in Section 2.3.

Considering the reduced form Equation (2.3), we might be able to find the mean and variance-covariance matrix of the complete system as function of exogenous variables. The expectation is given by:

$$\begin{aligned}\mathbb{E}(\mathbf{y}|\mathbf{X}, \mathbf{W}) &= \mathbb{E}\left[(\mathbf{I}_n - \rho \mathbf{W})^{-1}(\alpha \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}) + (\mathbf{I}_n - \rho \mathbf{W})^{-1}\boldsymbol{\varepsilon} \mid \mathbf{X}, \mathbf{W}\right] \\ &= (\mathbf{I}_n - \rho \mathbf{W})^{-1}(\alpha \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}).\end{aligned}\tag{2.5}$$

From Equation (2.3), we derive the variance-covariance matrix of \mathbf{y} :

$$\begin{aligned}\mathbb{V}(\mathbf{y}|\mathbf{W}, \mathbf{X}) &= \mathbb{E}(\mathbf{y}\mathbf{y}^\top | \mathbf{W}, \mathbf{X}) \\ &= (\mathbf{I}_n - \rho \mathbf{W})^{-1} \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top | \mathbf{W}, \mathbf{X}) (\mathbf{I}_n - \rho \mathbf{W}^\top)\end{aligned}\tag{2.6}$$

This $n \times n$ variance-covariance matrix is full, which implies that each location is correlated with every other location in the system. However, this correlation decreases with distance. Since we have not assumed anything about the error variance, we can say that $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top | \mathbf{W}, \mathbf{X})$ is a full matrix, say $\boldsymbol{\Omega}_\varepsilon$. This covers the possibility of heteroskedasticity, spatial autocorrelation, or both. In absence of either of these complications, the variance matrix simplifies to the usual $\sigma^2 \mathbf{I}_n$.

■ **Example 2.1 — County homicide rates in US.** In the criminology literature there has been a great emphasis of spatial diffusion of crime. The idea is that criminal violence may spread geographically via a diffusion process. For example, some researchers suggests that certain social processes such as illegal drug markets and gang rivalries may be important for explaining the pattern and mechanisms of the spread of homicides (Cohen and Tita, 1999).

In particular, empirical literature has focused on homicide rates and their determinants using the following OLS specification:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i \quad i = 1, \dots, n,$$

where y_i is the homicide rate in spatial unit i and \mathbf{x}_i is a $k \times 1$ set of covariates that explain homicide rates across spatial units. However, this model does not allow capturing the idea of spatial diffusion and spatial effects of homicide rates. For example, Baller et al. (2001), after rejecting the null hypothesis of spatial randomness on homicide rates, propose (among other spatial models) the following SLM process for modeling homicide rates using a county-level data for the decennial years in the 1960 to 1990 time period:

$$\mathbf{y} = \alpha \mathbf{r}_n + \rho \mathbf{W} \mathbf{y} + \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where \mathbf{y} is the homicide rates for the US counties, \mathbf{X} includes a deprivation, population density, median age, the unemployment rate, percent divorced, and a Southern dummy variable based on census definitions. As explained by [Baller et al. \(2001\)](#), if homicides rates are determined solely by the structural factors included in the \mathbf{X} matrix, there should be no spatial patterning of homicide beyond that created by socio-demographic similarities of geographically proximate counties. If this is the case, once all x_k are included in the model, the spatial relationship between y_i and y_j will become nonsignificant. This implies that $\rho = 0$.

This is the model most compatible with common notions of diffusion processes because it implies an influence of neighbors' homicide rates that is not simply an artifact of measured or unmeasured independent variables. Rather, homicide events in one place actually increase the likelihood of homicides in nearby locales. ■

2.1.2 Spatial Durbin Model

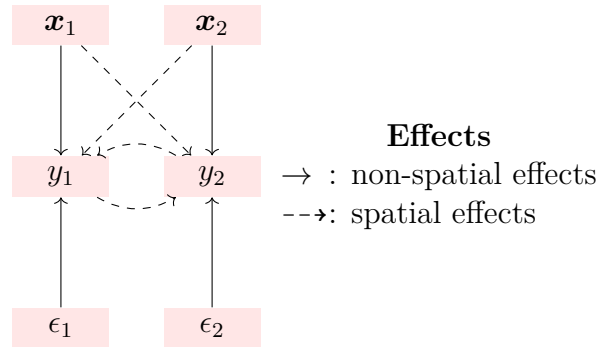
The Spatial Durbin Model (SDM) model is shown in Equation (2.7) along with its associated data generating process in Equation (2.8):

$$\mathbf{y} = \rho \mathbf{W} \mathbf{y} + \alpha \mathbf{r}_n + \mathbf{X} \boldsymbol{\beta} + \mathbf{W} \mathbf{X} \boldsymbol{\gamma} + \boldsymbol{\epsilon} \quad (2.7)$$

$$\mathbf{y} = (\mathbf{I}_n - \rho \mathbf{W})^{-1} (\alpha \mathbf{r}_n + \mathbf{X} \boldsymbol{\beta} + \mathbf{W} \mathbf{X} \boldsymbol{\gamma} + \boldsymbol{\epsilon}) \quad (2.8)$$

The SDM results in a spatial autoregressive model of a special form, including not only the spatially lagged dependent variable and the explanatory variables, but also the spatially lagged explanatory variables, $\mathbf{W} \mathbf{X}$: \mathbf{y} depends on own-regional factors from matrix \mathbf{X} , plus the same factors averaged over the n neighboring regions. This idea is shown in Figure 2.2. Note that Region 1 not only exerts an impact on Region 2 (an vice versa) via y , but also via the the independent variable x .

Figure 2.2: The SDM for Two Regions



As an example, consider that y is some measure of air pollution in each region. Thus, $\mathbf{W} \mathbf{y}$ states that air pollution in region 1 might affect pollution in region 2, and vice versa. If \mathbf{X} contains a measure of population density, the variable $\mathbf{W} \mathbf{X}$ would indicate that density and region 1 (2) would affect air pollution in region 2 (1).

This model has also very good properties in terms of calculation of marginal effects that will explore later.

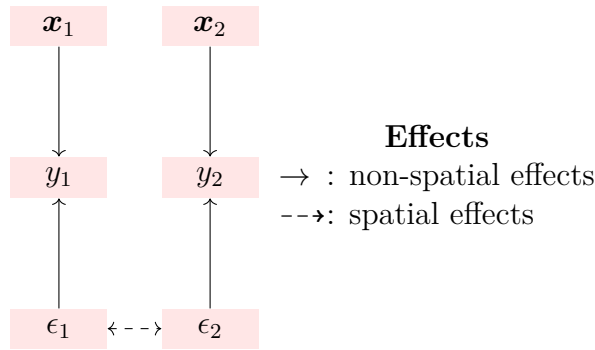
2.1.3 Spatial Error Model

Another form of spatial dependence occurs when the dependence works through the error process, in that the errors from different areas may display spatial autocorrelation. The Spatial Error Model (SEM) is formulated as:

$$\begin{aligned} \mathbf{y} &= \alpha \mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \\ \mathbf{u} &= \lambda \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}. \end{aligned} \quad (2.9)$$

where λ is the autoregressive parameter for the error lag $\mathbf{W}\mathbf{u}$ (to distinguish the notation from the spatial autoregressive coefficient ρ in a spatial lag model), and $\boldsymbol{\varepsilon}$ is a generally a i.i.d noise. Figure 2.3 visualizes the SEM for two regions. Note that the error term of both regions are related, and the only spatial effect goes from ϵ_1 to ϵ_2 and vice versa.

Figure 2.3: The SEM for two regions



As stated by [Anselin and Bera \(1998\)](#), spatial error dependence may be interpreted as a nuisance (and the parameter λ as a nuisance parameter) in the sense that it reflects spatial autocorrelation in measurement errors or in variables that are otherwise not crucial to the model (i.e., the “ignored” variables spillovers across the spatial units of observations).

Unlike previous models, interactions effects among the error terms do not require a theoretical model for a spatial or social interaction process, but instead, are consistent with a situation where determinants of the dependent variable omitted from the model are spatially autocorrelated, or with a situation where unobserved shocks follows a spatial pattern.

The spatial diffusion of this model can be analyzed if we consider the reduced form equation. If the matrix $(\mathbf{I}_n - \lambda \mathbf{W})$ is not singular, then (2.9) can be written under the following reduced form:

$$\mathbf{y} = \alpha \mathbf{1} + \mathbf{X}\boldsymbol{\beta} + (\mathbf{I}_n - \lambda \mathbf{W})^{-1} \boldsymbol{\varepsilon}. \quad (2.10)$$

This expression leads to a global spatial diffusion effect, but there is not spatial multiplier effect. The variance-covariance matrix is given by:

$$\begin{aligned} \mathbb{E}(\mathbf{y}\mathbf{y}^\top | \mathbf{W}, \mathbf{X}) &= \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top | \mathbf{W}, \mathbf{X}) \\ &= (\mathbf{I}_n - \lambda \mathbf{W})^{-1} \mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top | \mathbf{W}, \mathbf{X}) (\mathbf{I}_n - \lambda \mathbf{W}^\top)^{-1} \end{aligned} \quad (2.11)$$

R Interaction effects among the unobserved terms may also be interpreted to reflect a mechanism to correct rent-seeking politicians for unanticipated fiscal policy changes. See for example [Allers and Elhorst \(2005\)](#).

2.1.4 Spatial Autocorrelation Model

A more general model is the one that includes the key modeling insights from both spatial lag and spatial error model describe above. This model is called the Spatial Autocorrelation Model (SAC) and the its structural representation is the following:

$$y_i = \alpha + \rho \sum_{j=1}^n w_{ij} y_j + \sum_{k=1}^K x_{ik} \beta_k + u_i$$

$$u_i = \lambda \sum_{j=1}^n m_{ij} u_j + \epsilon_i$$

or more compactly,

$$\begin{aligned} \mathbf{y} &= \alpha \mathbf{1}_n + \rho \mathbf{W} \mathbf{y} + \mathbf{X} \boldsymbol{\beta} + \mathbf{u} \\ \mathbf{u} &= \lambda \mathbf{M} \mathbf{u} + \boldsymbol{\varepsilon} \end{aligned} \quad (2.12)$$

where the matrix \mathbf{W} and \mathbf{M} are $n \times n$ spatial-weighting matrices.¹ In this model, spatial interactions in the dependent variable and the disturbances are considered. As standard, the spatial weight matrices \mathbf{W} and \mathbf{M} are taken to be known and nonstochastic. These matrices are part of the model definition, and in many applications, $\mathbf{M} = \mathbf{W}$. When $\rho = 0$, the model reduces to the SEM. When $\lambda = 0$ the model reduces to the SLM (SAR) specification. Setting $\rho = 0$ and $\lambda = 0$ causes the model to reduce to a linear regression model with exogenous variables.

The reduced form is given by:

$$\mathbf{y} = (\mathbf{I}_n - \rho \mathbf{W})^{-1} (\mathbf{X} \boldsymbol{\beta} + \alpha \mathbf{1}_n) + (\mathbf{I}_n - \rho \mathbf{W})^{-1} (\mathbf{I}_n - \rho \mathbf{M})^{-1} \boldsymbol{\varepsilon}. \quad (2.13)$$

Figure 2.4 gives a more complete taxonomy for different spatial models. The more complete model is the General Nesting Spatial Model (GNS or Manski's Model), which includes spatial dependence in the dependent variable, exogenous variables and the error term. Different restrictions give rise to different spatial models.

Starting with the GNS model:

- Imposing the restriction $\boldsymbol{\gamma} = \mathbf{0}$ leads to the SAC model that includes both a spatial lag for the dependent variable and spatial lag for the error term, but excludes the influence of the spatially lagged explanatory variables.
- Imposing the restriction $\lambda = 0$ leads to the SDM.
- Imposing the restriction $\rho = 0$ leads to the Spatial Durbin Error Model (SDEM).

Starting with the SDM:

- The so-called common factor parameter restrictions ($\boldsymbol{\gamma} = -\rho \boldsymbol{\beta}$) yields the spatial error regression model (SEM) specification that assumes that externalities across spatial unites are mostly a nuisance spatial dependence problem caused by the regional transmission of random shocks.

¹This model is also known as SARAR(1, 1) model or Cliff-Ord models because of the impact that [Cliff and Ord \(1973\)](#) had on the subsequent literature. Note that SARAR(1, 1) is a special case of the more general SARAR(p, q) model.

$$\mathbf{y}_{t-1} = \rho \mathbf{W} \mathbf{y}_{t-2} + \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}_{t-1},$$

producing:

$$\begin{aligned} \mathbf{y}_t &= \mathbf{X} \boldsymbol{\beta} + \rho \mathbf{W} (\mathbf{X} \boldsymbol{\beta} + \rho \mathbf{W} \mathbf{y}_{t-2} + \boldsymbol{\varepsilon}_{t-1}) + \boldsymbol{\varepsilon}_t \\ &= \mathbf{X} \boldsymbol{\beta} + \rho \mathbf{W} \mathbf{X} \boldsymbol{\beta} + \rho^2 \mathbf{W}^2 \mathbf{y}_{t-2} + \boldsymbol{\varepsilon}_t + \rho \mathbf{W} \boldsymbol{\varepsilon}_{t-1}. \end{aligned} \quad (2.15)$$

Recursive substitution for past values of the vector \mathbf{y}_{t-r} on the right-hand side of (2.15) over q periods leads to:

$$\begin{aligned} \mathbf{y}_t &= (\mathbf{I}_n + \rho \mathbf{W} + \rho^2 \mathbf{W}^2 + \dots + \rho^{q-1} \mathbf{W}^{q-1}) \mathbf{X} \boldsymbol{\beta} + \rho^q \mathbf{W}^q \mathbf{y}_{t-q} + \mathbf{u}, \\ \mathbf{u} &= \boldsymbol{\varepsilon}_t + \rho \mathbf{W} \boldsymbol{\varepsilon}_{t-1} + \rho^2 \mathbf{W}^2 \boldsymbol{\varepsilon}_{t-2} + \dots + \rho^{q-1} \mathbf{W}^{q-1} \boldsymbol{\varepsilon}_{t-(q-1)}. \end{aligned}$$

The expected value of this spatial process is:

$$\mathbb{E}(\mathbf{y}_t) = (\mathbf{I}_n + \rho \mathbf{W} + \rho^2 \mathbf{W}^2 + \dots + \rho^{q-1} \mathbf{W}^{q-1}) \mathbf{X} \boldsymbol{\beta} + \rho^q \mathbf{W}^q \mathbf{y}_{t-q}, \quad (2.16)$$

where we use the fact that $\mathbb{E}(\boldsymbol{\varepsilon}_{t-r}) = 0, r = 0, \dots, q-1$, which also implies that $\mathbb{E}(\mathbf{u}) = \mathbf{0}$. Finally, taking the limit of (2.16),

$$\lim_{q \rightarrow \infty} \mathbb{E}(\mathbf{y}_t) = (\mathbf{I}_n - \rho \mathbf{W})^{-1} \mathbf{X} \boldsymbol{\beta}. \quad (2.17)$$

Note that we use the fact that the magnitude of $\rho^q \mathbf{W}^q \mathbf{y}_{t-q}$ tends to zero for large q , under the assumption that $|\rho| < 1$ and assuming that \mathbf{W} is row-stochastic, so the matrix \mathbf{W} has a principal eigenvalue of 1.

Equation (2.17) states that we can interpret the observed cross-sectional relation as the outcome or expectation of a long-run equilibrium or steady state. Note that this provides a dynamic motivation for the data generating process of the cross-sectional SLM that serves as a **workhorse** of spatial regression modeling. That is, a cross-sectional SLM relation can arise from time-dependence of decisions by economic agents located at various point in space when decisions depend on those neighbors.

2.2.2 SEM and Omitted Variables Motivation

Consider the following process:

$$\mathbf{y} = \mathbf{x} \boldsymbol{\beta} + \mathbf{z} \theta,$$

where \mathbf{x} and \mathbf{z} are **uncorrelated** vectors of dimension $n \times 1$, and the vector \mathbf{z} follows the following spatial autoregressive process:

$$\begin{aligned} \mathbf{z} &= \rho \mathbf{W} \mathbf{z} + \mathbf{r} \\ \mathbf{z} &= (\mathbf{I}_n - \rho \mathbf{W})^{-1} \mathbf{r} \end{aligned}$$

where $\mathbf{r} \sim N(0, \sigma_\epsilon^2 \mathbf{I}_n)$. Examples of \mathbf{z} are culture, social capital, or neighborhood prestige.

If \mathbf{z} is not observed, then:

$$\begin{aligned} \mathbf{y} &= \mathbf{x} \boldsymbol{\beta} + \mathbf{u} \\ \mathbf{u} &= (\mathbf{I}_n - \rho \mathbf{W})^{-1} \boldsymbol{\varepsilon} \end{aligned} \quad (2.18)$$

where $\boldsymbol{\varepsilon} = \theta \mathbf{r}$. Then, we have the DGP for the SEM.

2.2.3 SDM and Omitted Variables Motivation

Now suppose that \mathbf{X} and $\boldsymbol{\varepsilon}$ from (2.18) are correlated, given by the following process:

$$\begin{aligned}\boldsymbol{\varepsilon} &= \mathbf{x}\gamma + \mathbf{v} \\ \mathbf{v} &\sim N(0, \sigma^2 \mathbf{I}_n)\end{aligned}\tag{2.19}$$

where the scalar parameters γ and σ^2 govern the strength of the relationship between \mathbf{X} and $\mathbf{z} = (\mathbf{I}_n - \rho\mathbf{W})^{-1}\mathbf{r}$. Inserting (2.19) into (2.18), we obtain:

$$\begin{aligned}\mathbf{y} &= \mathbf{x}\beta + (\mathbf{I}_n - \rho\mathbf{W})^{-1}\boldsymbol{\varepsilon} \\ &= \mathbf{x}\beta + (\mathbf{I}_n - \rho\mathbf{W})^{-1}(\mathbf{x}\gamma + \mathbf{v}) \\ &= \mathbf{x}\beta + (\mathbf{I}_n - \rho\mathbf{W})^{-1}\mathbf{x}\gamma + (\mathbf{I}_n - \rho\mathbf{W})^{-1}\mathbf{v} \\ (\mathbf{I}_n - \rho\mathbf{W})\mathbf{y} &= (\mathbf{I}_n - \rho\mathbf{W})\mathbf{x}\beta + \mathbf{v} \\ \mathbf{y} &= \rho\mathbf{W}\mathbf{y} + \mathbf{x}(\beta + \gamma) + \mathbf{W}\mathbf{x}(-\rho\beta) + \mathbf{v}\end{aligned}\tag{2.20}$$

This is the Spatial Durbin Model (SDM), which includes a spatial lag of the dependent variable \mathbf{y} , as well as the explanatory variables \mathbf{x} .

2.3 Interpreting Spatial Models

2.3.1 Measuring Spillovers

A major focus of regional science is measuring spatial spillover. A basic definition of spillovers in a spatial context would be that changes occurring in one region exert impacts on other regions (LeSage and Pace, 2014). Some examples are:

- Changes in tax rate by one spatial unit might exert an impact on tax rate setting decisions of nearby regions, a phenomenon that has been labeled tax mimicking and yardstick competition between local government.
- Situations where home improvements made by one homeowner exert a beneficial impact on selling prices of neighboring homes.
- Innovation by university researchers diffuses to nearby firms.
- Air or water pollution generated in one region spills over to nearby regions.

The models reviewed in the previous section can be used to formally define the concept of a spatial spillover and, more importantly, to provide estimates of the quantitative magnitude of spillovers and test their statistical significance. There is however a distinction between global and local spillovers, which is discussed in Anselin (2003) and LeSage and Pace (2014).

We start our discussion about spillovers by formally defining global spillovers.

Definition 2.3.1 — Global Spillovers. Global spillovers arise when changes in a characteristic of one region impact all regions' outcomes. This applies even to the region itself since impacts can pass to the neighbors and back to the own region (feedback). Specifically, global spillovers impact the neighbors, neighbors to the neighbors, neighbors to the neighbors to the neighbors and so on.

The endogenous interactions produced by global spillovers lead to a scenario where changes in one region set in motion a sequence of adjustments in (potentially) all regions in the sample such that a new long-run steady state equilibrium arises (LeSage, 2014).

As explained by LeSage (2014), global spillovers might arise when considering local policies interactions. For example: *“it seems plausible that changes in levels of public assistance (cigarette taxes) in state A would lead to a reaction by neighboring states B to change their levels of assistances (taxes), which in turn produces a game-theoretic (feedback) response of state A, and also responses of states C who are neighbors to neighboring states B, and so on.”*

The following definition corresponds to local spillovers.

Definition 2.3.2 — Local Spillovers. Local spillovers represent a situation where the impact fall only on nearby or immediate neighbors, dying out before they impact regions that are neighbors to the neighbors.

As it can be noted from the previous definitions, the main difference is that feedback or endogenous interaction is only possible for global spillovers.

2.3.2 Marginal Effects

Mathematically, the notion of spillover can be thought as the derivative $\partial y_i / \partial x_j$. This means that changes to explanatory variables in region i impact the dependent variable in region $j \neq i$.

As an illustration, consider the SDM model, which can be re-written as:

$$\begin{aligned}
 (\mathbf{I}_n - \rho \mathbf{W}) \mathbf{y} &= \mathbf{X} \boldsymbol{\beta} + \mathbf{W} \mathbf{X} \boldsymbol{\theta} + \boldsymbol{\varepsilon}, \\
 \mathbf{y} &= (\mathbf{I}_n - \rho \mathbf{W})^{-1} \mathbf{X} \boldsymbol{\beta} + (\mathbf{I}_n - \rho \mathbf{W})^{-1} \mathbf{W} \mathbf{X} \boldsymbol{\theta} + (\mathbf{I}_n - \rho \mathbf{W})^{-1} \boldsymbol{\varepsilon}, \\
 \mathbf{y} &= \mathbf{A}(\mathbf{W})^{-1} \mathbf{X} \boldsymbol{\beta} + \mathbf{A}(\mathbf{W})^{-1} \mathbf{W} \mathbf{X} \boldsymbol{\theta} + \mathbf{A}(\mathbf{W})^{-1} \boldsymbol{\varepsilon}, \quad \text{since } \mathbf{A}(\mathbf{W}) = (\mathbf{I}_n - \rho \mathbf{W})^{-1} \\
 \mathbf{y} &= \mathbf{A}(\mathbf{W})^{-1} (\mathbf{X} \boldsymbol{\beta} + \mathbf{W} \mathbf{X} \boldsymbol{\theta}) + \mathbf{A}(\mathbf{W})^{-1} \boldsymbol{\varepsilon}, \\
 \mathbf{y} &= \sum_{r=1}^K \mathbf{A}(\mathbf{W})^{-1} (\mathbf{I}_n \beta_r + \mathbf{W} \theta_r) \mathbf{x}_r + \mathbf{A}(\mathbf{W})^{-1} \boldsymbol{\varepsilon}, \\
 \underbrace{\mathbf{y}}_{(n \times 1)} &= \sum_{r=1}^K \underbrace{\mathbf{S}_r(\mathbf{W})}_{(n \times n)} \underbrace{\mathbf{x}_r}_{n \times 1} + \underbrace{\mathbf{A}(\mathbf{W})^{-1}}_{(n \times n)} \underbrace{\boldsymbol{\varepsilon}}_{n \times 1}
 \end{aligned}$$

where $\mathbf{S}_r = \mathbf{A}(\mathbf{W})^{-1} (\mathbf{I}_n \beta_r + \mathbf{W} \theta_r)$, and

$$\mathbf{x}_r = \begin{pmatrix} x_{r1} \\ x_{r2} \\ \vdots \\ x_{rn} \end{pmatrix}.$$

Assuming that $\mathbb{E}(\epsilon_i) = 0$, then the expansion of the expected value yields:

$$\begin{pmatrix} \mathbb{E}(y_1) \\ \mathbb{E}(y_2) \\ \vdots \\ \mathbb{E}(y_n) \end{pmatrix} = \sum_{r=1}^K \begin{pmatrix} \mathbf{S}_r(\mathbf{W})_{11} & \mathbf{S}_r(\mathbf{W})_{12} & \dots & \mathbf{S}_r(\mathbf{W})_{1n} \\ \mathbf{S}_r(\mathbf{W})_{21} & \mathbf{S}_r(\mathbf{W})_{22} & \dots & \mathbf{S}_r(\mathbf{W})_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_r(\mathbf{W})_{n1} & \mathbf{S}_r(\mathbf{W})_{n2} & \dots & \mathbf{S}_r(\mathbf{W})_{nn} \end{pmatrix} \begin{pmatrix} x_{1r} \\ x_{2r} \\ \vdots \\ x_{nr} \end{pmatrix}. \quad (2.21)$$

For the dependent variable for spatial unit i , Equation (2.21) would be:

$$\mathbb{E}(y_i) = \sum_{r=1}^k [\mathbf{S}_r(\mathbf{W})_{i1}x_{1r} + \mathbf{S}_r(\mathbf{W})_{i2}x_{2r} + \dots + \mathbf{S}_r(\mathbf{W})_{in}x_{nr}]. \quad (2.22)$$

So, the impact on the expected value of location i given a change in the explanatory variable x_r in location j is

$$\frac{\partial \mathbb{E}(y_i)}{\partial x_{jr}} = \mathbf{S}_r(\mathbf{W})_{ij} \quad (2.23)$$

where $\mathbf{S}_r(\mathbf{W})_{ij}$ is this equation represents the i, j th element of the matrix $\mathbf{S}_r(\mathbf{W})$. This result implies that, unlike the OLS model, a change in some variable in certain region will potentially affect the expected value of the dependent variable in all other regions. Given this characteristic, this type of effect is known as **indirect effect**.

The impact of the expected value of region i , given a change in certain variable for the same region is given by

$$\frac{\partial \mathbb{E}(y_i)}{\partial x_{ir}} = \mathbf{S}_r(\mathbf{W})_{ii}. \quad (2.24)$$

This impact includes the **effect of feedback loops** where observation i affects observation j and observation j also affects observation i : a change in x_{ir} will affect the expected value of dependent variable in i , then will pass through the neighbors of i and back to the region itself. To shed more light on this, let us write the all the marginal effects in matrix notation as follows:

$$\begin{aligned} \begin{pmatrix} \frac{\partial \mathbb{E}(\mathbf{y})}{\partial x_{1r}} & \frac{\partial \mathbb{E}(\mathbf{y})}{\partial x_{2r}} & \dots & \frac{\partial \mathbb{E}(\mathbf{y})}{\partial x_{nr}} \end{pmatrix}_{(n \times n)} &= \begin{pmatrix} \frac{\partial \mathbb{E}(y_1)}{\partial x_{1r}} & \frac{\partial \mathbb{E}(y_1)}{\partial x_{2r}} & \dots & \frac{\partial \mathbb{E}(y_1)}{\partial x_{nr}} \\ \frac{\partial \mathbb{E}(y_2)}{\partial x_{1r}} & \frac{\partial \mathbb{E}(y_2)}{\partial x_{2r}} & \dots & \frac{\partial \mathbb{E}(y_2)}{\partial x_{nr}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mathbb{E}(y_n)}{\partial x_{1r}} & \frac{\partial \mathbb{E}(y_n)}{\partial x_{2r}} & \dots & \frac{\partial \mathbb{E}(y_n)}{\partial x_{nr}} \end{pmatrix} \\ &= \mathbf{A}(\mathbf{W})^{-1} (\mathbf{I}_n \beta_r + \mathbf{W} \theta_r) = \mathbf{S}_r(\mathbf{W}) \\ &= (\mathbf{I}_n - \rho \mathbf{W})^{-1} \begin{pmatrix} \beta_r & w_{12} \theta_r & \dots & w_{1n} \theta_r \\ w_{21} \theta_r & \beta_r & \dots & w_{2n} \theta_r \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} \theta_r & w_{n2} \theta_r & \dots & \beta_r \end{pmatrix} \end{aligned} \quad (2.25)$$

This expression is somewhat difficult to understand. To provide a better understanding we follow [Elhorst \(2010\)](#) and consider a model with 3 regions arranged linearly² with the following matrices:

$$\mathbf{W} = \begin{pmatrix} 0 & 1 & 0 \\ w_{21} & 0 & w_{23} \\ 0 & 1 & 0 \end{pmatrix} \quad (2.26)$$

and

²Unit 1 is neighbor of unit 2, unit 2 is a neighbor of both units 1 and 3, and unit 3 is a neighbor of unit 2.

$$\mathbf{A}(\mathbf{W})^{-1} = \frac{1}{1 - \rho^2} \begin{pmatrix} 1 - w_{23}\rho^2 & \rho & \rho^2 w_{23} \\ \rho w_{21} & 1 & \rho w_{23} \\ \rho^2 w_{21} & \rho & 1 - w_{21}\rho^2 \end{pmatrix} \quad (2.27)$$

where $w_{12} = w_{31} = 1$ since units 1 and 3 have only one neighbor, and $w_{21} + w_{23} = 1$, so we explicitly consider a row-standardized matrix. Substituting Equations (2.26) and (2.27) into Equation (2.25) we get:

$$\begin{pmatrix} \frac{\partial \mathbb{E}(\mathbf{y})}{\partial x_{1r}} & \frac{\partial \mathbb{E}(\mathbf{y})}{\partial x_{2r}} & \frac{\partial \mathbb{E}(\mathbf{y})}{\partial x_{3r}} \end{pmatrix} = \frac{1}{1 - \rho^2} \begin{pmatrix} (1 - w_{23}\rho^2)\beta_r + (w_{21}\rho)\theta_r & \rho\beta_r + \theta_r & (w_{23}\rho^2)\beta_r + (\rho w_{23})\theta_r \\ (w_{21}\rho)\beta_r + w_{21}\theta_r & \beta_r + \rho\theta_r & (w_{23}\rho)\beta_r + w_{23}\theta_r \\ (w_{21}\rho^2)\beta_r + (w_{21}\rho)\theta_r & \rho\beta_r + \theta_r & (1 - w_{21}\rho^2)\beta_r + (w_{23}\rho)\theta_r \end{pmatrix}$$

Every diagonal element of this matrix represents a direct effect. Consequently, indirect effect do not occur if both $\rho = 0$ and $\theta_k = 0$, since all non-diagonal elements will then be zero. Another important insight is that direct and indirect effects are different for different spatial units in the sample. Direct effects are different because the diagonal elements of the matrix $(\mathbf{I}_n - \rho\mathbf{W})^{-1}$ are different for different units, provided that $\rho \neq 0$. Indirect effects are different because both the non-diagonal elements of the matrix $(\mathbf{I}_n - \rho\mathbf{W})^{-1}$ and of the matrix \mathbf{W} are different for different units, provided that $\rho \neq 0$ and/or $\theta_k \neq 0$. Finally, note that indirect effects that occur if $\theta_k \neq 0$ are **local effects**, whereas indirect effects that occur if $\rho \neq 0$ are **global effects**.

Summary Measures

In general, the change of each variable in each region implies n^2 potential marginal effects. If we have K variables in our model, this implies $K \times n^2$ potential measures. Even for small values of n and K , it may already be rather difficult to report these results compactly. To overcome this problem, LeSage and Pace (2010, p. 36-37) propose the following scalar summary measures:

Definition 2.3.3 — Average Direct Impact. Let $\mathbf{S}_r = \mathbf{A}(\mathbf{W})^{-1}(\mathbf{I}_n\beta_r + \mathbf{W}\theta_r)$ for variable r . The impact of changes in the i th observation of x_r , which is denoted x_{ir} , on y_i could be summarized by measuring the average $S_r(\mathbf{W})_{ii}$, which equals

$$\text{ADI} = \frac{1}{n} \text{tr}(\mathbf{S}_r(\mathbf{W})) \quad (2.28)$$

Averaging over the direct impact associated with all observations i is similar in spirit to typical regression coefficient interpretations that represent average response of the dependent to independent variables over the sample of observations.

Definition 2.3.4 — Average Total Impact to an Observation. Let $\mathbf{S}_r = \mathbf{A}(\mathbf{W})^{-1}(\mathbf{I}_n\beta_r + \mathbf{W}\theta_r)$ for variable r . The sum across the i th row of $\mathbf{S}_r(\mathbf{W})$ would be represent the total impact on individual observation y_i resulting from changing the r th explanatory variable by the same amount across all n observations. There are n of these sums given by the column vector $\mathbf{c}_r = \mathbf{S}_r(\mathbf{W})\mathbf{1}_n$, so an average of these total impacts is:

$$\text{ATIT} = \frac{1}{n} \mathbf{1}'_n \mathbf{c}_r \quad (2.29)$$

Definition 2.3.5 — Average Total Impact from an Observation. Let $\mathbf{S}_r = \mathbf{A}(\mathbf{W})^{-1}(\mathbf{I}_n\beta_r + \mathbf{W}\theta_r)$ for variable r . The sum down the j th column of $\mathbf{S}_r(\mathbf{W})$ would yield the total impact over all y_i from changing the r th explanatory variable by an amount in the j th observation. There are n of these sums given by the row vector $\mathbf{r}_r = \mathbf{z}'_n \mathbf{S}_r(\mathbf{W})$, so an average of these total impacts is:

$$\text{ATIF} = \frac{1}{n} \mathbf{r}_r \mathbf{z}_n \quad (2.30)$$

The definition 2.3.5 relates how changes in a single observation j influences all observations. In contrast, definition 2.3.4 considers how changes in all observations influences a single observation i . In both cases, averaging over all n observations, leads to the same numerical result. The implication of this interesting result is that the **average total impact** is the average of all derivatives of y_i with respect to x_{jr} for any i, j .

Therefore:

$$\bar{M}(r)_{\text{direct}} = n^{-1} \text{tr}(\mathbf{S}_r(\mathbf{W})) \quad (2.31)$$

$$\bar{M}(r)_{\text{total}} = n^{-1} \mathbf{z}'_n \mathbf{S}_r(\mathbf{W}) \mathbf{z}_n \quad (2.32)$$

$$\bar{M}(r)_{\text{indirect}} = \bar{M}(r)_{\text{total}} - \bar{M}(r)_{\text{direct}} \quad (2.33)$$

Given our example above, we obtain a direct effect of:

$$\frac{(3 - \rho^2)}{3(1 - \rho^2)} \beta_k + \frac{2\rho}{3(1 - \rho^2)} \theta_k,$$

and an indirect effect of

$$\frac{3\rho + \rho^2}{3(1 - \rho^2)} \beta_k + \frac{3 + \rho}{3(1 - \rho^2)} \theta_k.$$

Unfortunately, since every application will have its own unique number of observations n and spatial weight matrix (\mathbf{W}), these formulae cannot be generalized.

■ **Example 2.2 — The effect of number of workers on commuting times.** Kirby and LeSage (2009) use an SDM specification to consider changes in the (logged) number of workers in the US census tracts with commuting times exceeding 45 minutes one way, between 1990 and 2000 (See also the example in Section 2.5). The motivation of this investigation is the fact that the percentage of the US workers with these long commute times in 1990 was 12.5% compared to 15.4% in 2000, an increase of more than 10%. When deciding which model to estimate, they note that spillover impacts from an increase in commuters traveling long distances to work would seem **global** in nature, since the congestion effects of more travelers on one segment of a metropolitan area roadway network impact travel times of other travelers on the entire network. Furthermore, they state that **feedback** effects seem likely since congestion arising from commuting decisions by workers in one tract will spillover to neighboring tracts, which in turn create congestion feedback to the own tract. These two observations led the authors to specify the following SDM:

$$\mathbf{y} = \rho \mathbf{W} \mathbf{y} + \alpha \mathbf{z}_n + \mathbf{X} \beta + \mathbf{W} \mathbf{X} \theta + \varepsilon,$$

where \mathbf{X} includes the (logged) number of workers with long commute times (\mathbf{y}), variables related to location decision of households; age, gender and income distribution of resident

population, and geographical characteristics of the tract, and $\mathbf{W}\mathbf{X}$ includes these same characteristics of neighboring census tracts. Based on a comparison of **direct**, **indirect** and **total effects** estimates from the 1990 and 2000 models, they conclude that the suite of variables reflecting the age and gender distribution of population in the tracts represents the primary explanation for changes in the number of workers with long commute times between 1990 and 2000. The spillover impacts of the number of employed females in the 1990 model was positive suggesting that more employed females in a tract produced an increase in long commute times for neighboring tract commuters. In contrast, for the 2000 model, spillovers associated with employed females were negative, so that more employed females in a tract reduced long commute times for workers located in neighboring tracts. ■

■ **Example 2.3 — Effect of pollution on housing price.** Kim et al. (2003) use a spatial-lag hedonic model in order to assess the direct and indirect effect of quality air on housing price. The main model is the following:

$$\mathbf{p} = \rho \mathbf{W}\mathbf{p} + \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{X}_3\boldsymbol{\beta}_3 + \boldsymbol{\varepsilon},$$

where \mathbf{p} is the vector of housing prices, ρ is a spatial autocorrelation parameter, \mathbf{W} is the $n \times n$ spatial weight matrix, \mathbf{X}_1 is a matrix with observations on structural characteristics, \mathbf{X}_2 is a matrix with observations on neighborhood characteristics, and \mathbf{X}_3 is a matrix with observations on environmental quality (SO_2 and NO_x).

The marginal implicit price (marginal benefit) of the hedonic equation is derived as

$$\left(\frac{\partial \mathbb{E}(\mathbf{p})}{\partial x_{1r}} \quad \frac{\partial \mathbb{E}(\mathbf{p})}{\partial x_{2r}} \quad \dots \quad \frac{\partial \mathbb{E}(\mathbf{p})}{\partial x_{nr}} \right) = \mathbf{A}(\mathbf{W})^{-1} \mathbf{I}_n \boldsymbol{\beta}_r \quad \text{where} \quad \mathbf{A}(\mathbf{W})^{-1} = (\mathbf{I}_n - \rho \mathbf{W})^{-1}$$

Focusing on the first row the interpretation is the following: the housing price of location i is not only affected by a marginal change air quality of location i but also is affected by marginal changes of air quality in other locations. That is, the total impact of a change in air quality on housing price at location i is the sum of the direct impacts $\partial p_1 / \partial x_{1k}$ plus induced impacts $\sum_{i=2}^n \partial p_1 / \partial x_{ik}$ (See our Definition 2.3.4).

An important point evidenced by Kim et al. (2003) is that, if the row-sums of \mathbf{W} is less than or equal to one and ρ in the proper parameter space, i.e., $\rho < 1$, then the total average effect can be computed as $\beta_r / (1 - \rho)$. To see this note that

$$\begin{aligned}
n^{-1} \mathbf{z}^\top \mathbf{S}_r(\mathbf{W}) \mathbf{z} &= n^{-1} \mathbf{z}^\top \left[\mathbf{A}(\mathbf{W})^{-1} (\mathbf{I} \beta_r) \right] \mathbf{z} \\
&= n^{-1} \mathbf{z}^\top \left[(\mathbf{I}_n - \rho \mathbf{W})^{-1} \right] (\mathbf{I} \beta_r) \mathbf{z} \\
&= n^{-1} \mathbf{z}^\top \left[\mathbf{I}_n + \rho \mathbf{W} + \rho^2 \mathbf{W}^2 + \dots \right] (\mathbf{I} \beta_r) \mathbf{z} \quad \text{using Lemma 2.3} \\
&= n^{-1} \mathbf{z}^\top \left[\mathbf{I}_n \beta_r + \rho \mathbf{W} \beta_r + \rho^2 \mathbf{W}^2 \beta_r + \dots \right] \mathbf{z} \\
&= n^{-1} \mathbf{z}^\top \left[\mathbf{I}_n \mathbf{z} \beta_r + \rho \mathbf{W} \mathbf{z} \beta_r + \rho^2 \mathbf{W} (\mathbf{W} \mathbf{z}) \beta_r + \rho^3 \mathbf{W} \mathbf{W} (\mathbf{W} \mathbf{z}) \right] \\
&= n^{-1} \mathbf{z}^\top \left[\beta_r \mathbf{z} + \rho \beta_r \mathbf{z} + \rho^2 \beta_r \mathbf{z} + \rho^3 \beta_r \mathbf{z} + \dots \right] \quad \because \mathbf{W}^l \mathbf{z} = \mathbf{z} \\
&= n^{-1} \mathbf{z}^\top \left[\beta_r + \rho \beta_r + \rho^2 \beta_r + \rho^3 \beta_r + \dots \right] \mathbf{z} \\
&= n^{-1} \left[\beta_r + \rho \beta_r + \rho^2 \beta_r + \rho^3 + \dots \right] \mathbf{z}^\top \mathbf{z} \\
&= n^{-1} \left[\beta_r + \rho \beta_r + \rho^2 \beta_r + \rho^3 \beta_r + \dots \right] n \\
&= \frac{\beta_r}{(1 - \rho)}
\end{aligned} \tag{2.34}$$

The model is estimated in a semi-log functional form, therefore the estimated coefficients can be interpreted as semi-elasticities. In particular, note that the elasticity for SO_2 is given by:

$$\begin{aligned}
\epsilon_{\text{SO}_2} &= \left(\frac{\text{SO}_2}{p} \right) \left(\frac{dp}{d\text{SO}_2} \right) \\
&= \left(\frac{\text{SO}_2}{p} \right) \left(\frac{\beta_r}{(1 - \rho)} \cdot p \right) \quad \text{since the model is log-lin} \\
&= \frac{\beta_r}{(1 - \rho)} \cdot \text{SO}_2
\end{aligned} \tag{2.35}$$

Using the estimated $\hat{\rho} = 0.549$ and replacing SO_2 by its mean value they obtain that the elasticity of housing price from a given small change in air quality is about $0.348 \approx 4\%$. The marginal benefits per household of a permanent 4% improvement in air quality using $\beta_{\text{SO}_2} (\mathbf{I}_n - \rho \mathbf{W})^{-1} \mathbf{p}$ is about \$2333 (1.43% of mean house value) for owners. ■

■ **Example 2.4 — Human capital and labor productivity.** Fischer et al. (2009) analyze the role of human capital in explaining labor productivity variation among European region. In particular they estimate the following model:

$$\mathbf{y} = \rho \mathbf{W} \mathbf{y} + \mathbf{X} \boldsymbol{\beta} + \mathbf{W} \mathbf{X} \boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

where \mathbf{y} is the vector of observations on the (log of) labor productivity level at the end of the sample period (2004) and \mathbf{X} contains (the log of) labor productivity and human capital at the beginning of the sample period (1995). The parameter ρ is expected to be positive indicating that regional productivity levels are positively related to a linear combination of neighboring regions' productivity. The parameter vector $\boldsymbol{\gamma}$ captures two types of spatial externalities: spatial effects working through the level of labor productivity and spatial effects working through the level of human capital, both at the beginning of the sample period.

The estimated parameter of the spatial autoregressive parameter is $\hat{\rho} = 0.664$ providing evidence for the existence of significant spatial effects working through the dependent variable.

The mean direct impact for the human capital is 0.1317, whereas the indirect impact is -0.1968. They interpret the indirect impact in two ways. First, they argue that the indirect impact reflects how a change in the human capital level of all regions by some constant would impact the labor productivity of a typical region (observation). The sign of the estimated mean indirect impact implies that an increase in the initial level of human capital of all other regions would decrease the productivity level of a typical region. This indirect impact takes into account the fact that the change in initial human capital level negatively impacts other regions' labor productivity, which in turn negatively influences our typical region's labor productivity due to the presence of positive spatial dependence on neighboring regions' labor productivity levels.

Second [Fischer et al. \(2009\)](#) measure the cumulative impact of a change in region's i initial level of human capital averaged over all other regions. The impact from changing a single region's initial level of human capital on each of the other region's labor productivity is small, but cumulatively the impact measures -0.1968. ■

R A very good paper for those interesting in making the connection between global/local spillovers and different spatial model specifications is [LeSage \(2014\)](#). This is a must-read paper.

2.3.3 Partitioning Global Effects Estimates Over Space

It should bear in mind that these scalar summary measures of impact reflect how these changes would work through the simultaneous dependence system over time to culminate in a new steady state equilibrium. Therefore, they should be considered as those impacts that would take place once all regions reach their equilibrium after the initial change in the variable of interest (See our discussion in Section 2.2.1). However one could track the cumulative effects as the impacts pass through neighbors, neighbors of neighbors and so on.

R Cross-sectional observations could be viewed as reflecting a (comparative static) slice at one point in time of a long-run steady-state equilibrium relationship, and the partial derivatives viewed as reflecting a comparative static analysis of changes that represent new steady-state relationship that would arise ([LeSage, 2014](#)).

Intuition tell us that impacts arising from a change in the explanatory variables will influence low-order neighbors more than higher-order neighbors. Therefore, we would expect a decline in the impacts' magnitude as we move from lower- to higher-order neighbors. To get a better idea of this process is necessary to consider the matrix $\mathbf{S}_r(\mathbf{W})$ and recognize, by Lemma 2.3, that this matrix can be expressed as a linear combination of power of the weight matrix \mathbf{W} . In particular, recall that if \mathbf{W} is a row standardized matrix such that $\rho \in (-1, 1)$, then by Lemma 2.3:

$$\left(\frac{\partial \mathbb{E}(\mathbf{y})}{\partial x_{1r}} \quad \frac{\partial \mathbb{E}(\mathbf{y})}{\partial x_{2r}} \quad \dots \quad \frac{\partial \mathbb{E}(\mathbf{y})}{\partial x_{nr}} \right) \approx \left(\mathbf{I}_n + \rho \mathbf{W} + \rho^2 \mathbf{W}^2 + \rho^3 \mathbf{W}^3 + \dots + \rho^l \mathbf{W}^l \right) \mathbf{I}_n \beta_r \quad (2.36)$$

This expression allow us to observe the impact associated with each power of \mathbf{W} , where these powers corresponds to the observation themselves (zero-order), immediate neighbors (first-order), neighbors of neighbors (second-order), and so on. Using this expansion we could account for both the cumulative effects as marginal and total direct, indirect associated with different order of neighbors.

2.4 Predictors for Spatial Models

Kelejian and Prucha (2007) consider different information sets and define predictors as conditional means based on these information sets. Consider the SAC (SARAR) model:

$$\begin{aligned} \mathbf{y} &= \rho \mathbf{W} \mathbf{y} + \mathbf{X} \boldsymbol{\beta} + \mathbf{u}, \\ \mathbf{u} &= \lambda \mathbf{W} \mathbf{u} + \boldsymbol{\varepsilon}. \end{aligned} \quad (2.37)$$

Thus, the reduced form equation is:

$$\mathbf{y} = (\mathbf{I}_n - \rho \mathbf{W})^{-1} \mathbf{X} \boldsymbol{\beta} + (\mathbf{I}_n - \rho \mathbf{W})^{-1} (\mathbf{I}_n - \lambda \mathbf{W})^{-1} \boldsymbol{\varepsilon}. \quad (2.38)$$

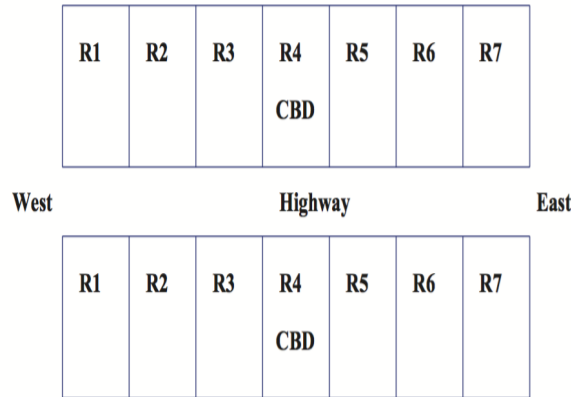
Assume that $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_n)$

2.5 Lesage's Book Example

2.5.1 Commuting Times and Congestion

In this section we use LeSage and Pace (2010)'s example as an illustration of spatial spillovers.³ For this purpose consider a set of seven regions show in Figure 2.5, which represent three regions to the west and three to the east of a central business district (CBD). In particular, consider region $R4$ as being the central business district. Since the entire region contains only a single roadway, all commuters share this route to and from the CBD.

Figure 2.5: Regions east and west of the CBD



We observe the following set of the sample data for these regions that relates travel times to the CBD (in minutes) contained in the dependent variable vector \mathbf{y} to distance (in miles) and population density (population per square block) of the regions in the two columns of the matrix \mathbf{X} .

³This example is further explore in Kirby and LeSage (2009) with a real application.

$$y = \begin{pmatrix} \text{Travel times} \\ 42 \\ 37 \\ 30 \\ 26 \\ 30 \\ 37 \\ 42 \end{pmatrix} \quad X = \begin{pmatrix} \text{Density} & \text{Distance} \\ 10 & 30 \\ 20 & 20 \\ 30 & 10 \\ 50 & 0 \\ 30 & 10 \\ 20 & 20 \\ 10 & 30 \end{pmatrix} \begin{matrix} \text{ex-urban areas } R1 \\ \text{far suburbs } R2 \\ \text{near suburbs } R3 \\ \text{CBD } R4 \\ \text{near suburbs } R5 \\ \text{far suburbs } R6 \\ \text{ex-urban areas } R7 \end{matrix}$$

According to [LeSage and Pace \(2010\)](#), the pattern of longer travel times for more distant regions R1 and R7 versus nearer R3 and R5 found in vector \mathbf{y} seems to clearly violate independence, since travel times appear similar for neighboring regions (see also Example 2.2). However one can argue that the observed pattern is not due to spatial dependence, but rather it is explained by the variables Distance and Density associated with each region, since these also appear similar for neighboring regions. Note that even for individual residing in the CBD, it takes time to go somewhere else in the CBD. Therefore, the travel time for intra-CBD travel is 26 minutes despite having a distance of 0 miles.

If we assume that the observed data was collected in a given day and averaged over a 24-hour period, it can be hypothesized that congestion effects that arise from the shared highway can explain the observed pattern of travel times. It is reasonable to claim that longer travel times in one region should lead to longer travel times in neighboring regions on any given day. This is because commuters pass from one region to another as they travel along the highway to the CBD.

Congestion effects represent one type of spatial spillover, which do not occur simultaneously, but require some time for the traffic delay to arise. From a modeling point of view, this effect cannot be captured by OLS model with distance and density as independent variables. These are dynamic feedback effects from travel time on a particular day that impact travel times of neighboring regions in the short time interval required for the traffic delay to occur. Since the explanatory variable distance would not change from day to day, and population density would change very slowly on a daily time scale, these variables would not be capable of explaining daily delay phenomena.

A better way of explaining congestion is by the following DGP:

$$\mathbf{y} = \rho_0 \mathbf{W} \mathbf{y} + \mathbf{X} \boldsymbol{\beta}_0 + \boldsymbol{\varepsilon},$$

such that:

$$\hat{\mathbf{y}} = (\mathbf{I}_n - \hat{\rho} \mathbf{W})^{-1} \mathbf{X} \hat{\boldsymbol{\beta}},$$

where the estimated parameters are $\hat{\boldsymbol{\beta}} = (0.135, 0.561)'$ and $\hat{\rho} = 0.640$ (assume that somehow we have estimated these parameters). Note that the estimated spatial autoregressive parameters indicates positive spatial dependence in the commuting times.

2.5.2 Computing Effects in R

Now think about the following question: What would be the estimated spillovers if region *R2* doubles its population density? To answer this question we first obtain the predicted values of travel times before the change.⁴ That is, we first obtain:

$$\hat{\mathbf{y}}^{(1)} = (\mathbf{I}_n - \hat{\rho}\mathbf{W})^{-1} \mathbf{X}\hat{\boldsymbol{\beta}}.$$

```
# Estimated coefficients
b <- c(0.135, 0.561)
rho <- 0.642

# W and X
X <- cbind(c(10, 20, 30, 50, 30, 20, 10),
           c(30, 20, 10, 0, 10, 20, 30))
W <- cbind(c(0, 1, 0, 0, 0, 0, 0),
           c(1, 0, 1, 0, 0, 0, 0),
           c(0, 1, 0, 1, 0, 0, 0),
           c(0, 0, 1, 0, 1, 0, 0),
           c(0, 0, 0, 1, 0, 1, 0),
           c(0, 0, 0, 0, 1, 0, 1),
           c(0, 0, 0, 0, 0, 1, 0))
Ws <- W / rowSums(W)

# Prediction
yhat_1 <- solve(diag(nrow(W)) - rho * Ws) %*% crossprod(t(X), b)
```

Now we estimate the predicted values of travel times after the change in population density in *R2* using:

$$\hat{\mathbf{y}}^{(2)} = (\mathbf{I}_n - \hat{\rho}\mathbf{W})^{-1} \widetilde{\mathbf{X}}\hat{\boldsymbol{\beta}} \quad (2.39)$$

where $\widetilde{\mathbf{X}}$ is the new matrix reflecting a doubling of the population density of region *R2*.⁵ A comparison of predictions $\hat{\mathbf{y}}^{(1)}$ and $\hat{\mathbf{y}}^{(2)}$ are going to be used to illustrate how the model generates spatial spillovers.

```
# Now we double the population density of a single region
X_d <- cbind(c(10, 40, 30, 50, 30, 20, 10),
            c(30, 20, 10, 0, 10, 20, 30))

# Compute predicted value after the change
yhat_2 <- solve(diag(nrow(W)) - rho * Ws) %*% crossprod(t(X_d), b)

# Results
```

⁴Note that there is a typo in LeSage and Pace (2010), because in their equation (1.19) they double distance, not density.

⁵For more about prediction in the spatial context see Kelejian and Prucha (2007).

```

result <- cbind(yhat_1, yhat_2, yhat_2 - yhat_1)
colnames(result) <- c("y1", "y2", "y2 - y1")
round(result, 2)

##           y1      y2 y2 - y1
## [1,] 41.90 44.46    2.56
## [2,] 36.95 40.93    3.99
## [3,] 29.84 31.28    1.45
## [4,] 25.90 26.43    0.53
## [5,] 29.84 30.03    0.19
## [6,] 36.95 37.03    0.08
## [7,] 41.90 41.95    0.05

sum(yhat_2 - yhat_1)

## [1] 8.846915

```

The two set of predictions show that the change in region $R2$ population density has a direct effect that increases the commuting times for residents of region $R2$ by ≈ 4 minutes. It also has an indirect or spillover effect that produces an increase in commuting times for the other six regions. Furthermore, it can be noticed that the increase in commuting times for neighboring regions $R1$ and $R3$ are the greatest and these spillovers decline as we move to regions in the sample that are located farther away from region $R2$ where the change in population density occurred.

What is the cumulative indirect impacts? Adding up the increased commuting times across all other regions (excluding the own-region change in commuting time), we find that equals $\approx 4.86(2.56 + 1.45 + 0.53 + 0.19 + 0.08 + 0.05)$ minutes, which is larger than the direct (own-region) impact of 4 minutes. Finally, the total impact of all residents of the seven regions from the change in population density of region $R2$ is the sum of the direct and indirect effects, or 8.85 minutes increase in travel times to the CBD.

Now assume that the OLS estimates for the example above are: $\hat{\beta}_{OLS} = [0.55, 1.25]$. Using these estimates we compute the OLS predictions based on the matrices \mathbf{X} and $\tilde{\mathbf{X}}$ as shown above.

```

# Ols prediction
b_ols <- c(0.55, 1.25)
yhat_1 <- crossprod(t(X), b_ols)
yhat_2 <- crossprod(t(X_d), b_ols)
result <- cbind(yhat_1, yhat_2, yhat_2 - yhat_1)
colnames(result) <- c("y1", "y2", "y2 - y1")
round(result, 2)

##           y1      y2 y2 - y1
## [1,] 43.0 43.0      0
## [2,] 36.0 47.0     11
## [3,] 29.0 29.0      0
## [4,] 27.5 27.5      0

```

```
## [5,] 29.0 29.0      0
## [6,] 36.0 36.0      0
## [7,] 43.0 43.0      0
```

The results show no spatial spillovers. Only the travel time of $R2$ is affected by the change in population density of region $R2$. It can be also observed that OLS prediction is upward bias. This is the main message here. An OLS model does not allow for spatial spillover impacts and generates biased marginal effects.

Now we further explore our formulas and definition from previous Section. As we showed in Equation (2.24), the impact of changes in the i th observation of x_r on y_i is $S_r(W)_{ii}$. Given the SLM structure of our example, this is equivalent to

$$\frac{\partial \mathbb{E}(\text{CT}_i)}{\partial \text{density}_i} = S_{\text{density}}(\mathbf{W})_{ii}, \quad \text{where } S_{\text{density}} = (\mathbf{I} - \rho \mathbf{W})^{-1} \mathbf{I} \beta_{\text{density}}.$$

We can compute our S_{density} in the following way.

```
# Compute S(W) matrix for density
b_dens <- 0.135
S <- solve(diag(nrow(W)) - rho * Ws) %*% diag(nrow(W)) * b_dens
colnames(S) <- rownames(S) <- c("R1", "R2", "R3", "R4", "R5", "R6", "R7")
```

Then, the direct impact of doubling population density of $R2$ on the expected value of commuting time for $R2$ is given by

$$\Delta \mathbb{E}(\text{CT}_2) = S_{\text{density}}(\mathbf{W})_{22} \Delta \text{density}_2 = S_{\text{density}}(\mathbf{W})_{22} \cdot 20$$

In R, this equals :

```
# Direct impact of R2 on R2
round(S[2,2] * 20, 2)

## [1] 3.99
```

Note that this value is the same as that found using the predicted value procedure: by doubling population density in $R2$ increases the commuting times for residents of region $R2$ by ≈ 4 minutes.

Finding the indirect impact on region $R1$ is similar given Equation 2.23. The indirect impact on region $R1$ is given by:

$$\Delta \mathbb{E}(\text{CT}_1) = S_{\text{density}}(\mathbf{W})_{12} \Delta \text{density}_2 = S_{\text{density}}(\mathbf{W})_{12} \cdot 20$$

That is:

```
# Indirect impact of R2 on R1
round(S[1,2] * 20, 2)

## [1] 2.56
```

Again, note that is the same value computed before: An increase of 100% of population density in $R2$ implies an increase of travel time of region $R1$ to CBD of about 2.56 minutes, after considering all feedback effects.

An interesting question would be the following: What would be the impact on commuting time on $R1$ if population density increases by 20 in all the Regions? To answer this question, we should recall our definition 2.3.4 states that the sum across the i th row of $\mathbf{S}_r(\mathbf{W})$ would be represent the total impact on individual observation y_i resulting from changing the r th explanatory variable by the same amount across n observations.

```
# ATIT
round(sum(S[1, ]) * 20, 2)

## [1] 7.54
```

This number implies that the total impact to $R1$ will be an increase of commuting time of ≈ 7.5 minutes. Using the formula for ATIT gives the same result:

```
# ATIT
n <- nrow(W)
vones <- rep(1, n)
round(((t(vones) %*% S %*% vones) / n ) * 20, 2)

##      [,1]
## [1,] 7.54
```

Similarly, we could ask: What would be the impact of increasing density by 20 in $R1$ on all the other regions? This is equivalent to our definition 2.3.5 which state that the sum down the j th column of $\mathbf{S}_r(\mathbf{W})$ would yield the total impact over all y_i from changing the r th explanatory variable by an amount in the j th observation.

```
# ATIF
round(sum(S[, 1]) * 20, 2)

## [1] 5.54
```

In words, increasing density by 20 in $R1$ would imply a total effect in all the regions of about 7.54 minutes.

Imagine that you are a policy maker and you are considering in implementing a policy to reduce population density and hence reduce commuting time in the regions. However, given that resources are scarce, you must select which region to implement this policy. In order to produce a greater effect of policy you could use the estimated spatial model and look for the region that will have the greatest overall impact (considering feedback effects). Basically, this involves calculating the column sum of $\mathbf{S}_r(\mathbf{W})$ for each region in the following way:

```
# Computing colsums of S(W)
round(colSums(S), 2)

##   R1   R2   R3   R4   R5   R6   R7
## 0.28 0.44 0.40 0.39 0.40 0.44 0.28
```

Note that the impact of decreasing population density by 1 will have a greater reduction in commuting time if applied in regions $R2$ and $R6$ (why?)

Finally, the average direct, indirect and total effects of an increase in 1 in population density in all the regions can be computed as follows.

```
# Average Direct Impact
ADI <- sum(diag(S)) / nrow(W)
round(ADI, 4)

## [1] 0.1837

# Average Total Impact
Total <- crossprod(rep(1, nrow(W)), S) %*% rep(1, nrow(W)) / nrow(W)
round(Total, 4)

##          [,1]
## [1,] 0.3771

# Average Indirect Impact
round(Total - ADI, 4)

##          [,1]
## [1,] 0.1934
```

Equation (2.34) of Example 2.3, we show that the total effect can be also be computed as $\beta_r/(1 - \rho)$. We know show that this proposition is true for our example

```
#Check total effect
b_dens / (1 - rho )

## [1] 0.377095
```

2.5.3 Cumulative Effects

The main idea of this exercise is to show how the change in some explanatory variable produces changes in the independent variable in all the spatial units by decomposing them into cumulative and marginal impacts for different order of neighbors as explained in Section 2.3.3.

First, we load the package **expm** which will allow us to compute power of matrices in a loop. Then we create the estimated coefficients along with the \mathbf{W} matrix:

```
# Package to compute power of a matrix
library("expm")
```

In order to create the decomposition for the ADI, AII and ATI, we create the following loop from $q = 0$ to $q = 10$:

```
## Loop for decomposition
out <- matrix(NA, nrow = 11, ncol = 3) # Matrix for the results
colnames(out) <- c("Total", "Direct", "Indirect") # colnames
rownames(out) <- paste("q", sep = "=", seq(0, 10)) # rownames

for (q in 0:10) {
  if (q == 0) { # If q=0, then Sr = I * beta
    S <- diag(n) * b_dens
  } else {
    S <- (rho ^ q * Ws %^% q) * b_dens
  }
  q <- q + 1 # the row = 0 doesn't exist!
  out[q, 2] <- sum(diag(S)) / n
  out[q, 1] <- crossprod(rep(1, n), S) %*% rep(1, n) / n
  out[q, 3] <- out[q, 1] - out[q, 2]
}
```

The results are the following

```
# Print results
round(out, 4)

##      Total Direct Indirect
## q=0  0.1350 0.1350  0.0000
## q=1  0.0867 0.0000  0.0867
## q=2  0.0556 0.0318  0.0238
## q=3  0.0357 0.0000  0.0357
## q=4  0.0229 0.0106  0.0123
## q=5  0.0147 0.0000  0.0147
## q=6  0.0095 0.0039  0.0056
## q=7  0.0061 0.0000  0.0061
## q=8  0.0039 0.0015  0.0024
## q=9  0.0025 0.0000  0.0025
## q=10 0.0016 0.0006  0.0010

round(colSums(out), 4)

##      Total      Direct Indirect
## 0.3742 0.1834 0.1909
```

This table shows both the cumulative and partitioned direct, indirect and total impacts associated with orders 0 to 10 for the SLM. The cumulative direct impact from previous section equal to 0.1837, which given the coefficient 0.1350 indicates that *there is a feedback equal to $(0.1837 - 0.1350) = 0.0487$ arising from each region impacting neighbors that in turn impacts neighbors to neighbors and so on.*

The column sum of the matrix `out` shows that by the time we reach 10th-order neighbors we have accounted for 0.1834 of the 0.1837 cumulative direct effect. It is important noting

that for \mathbf{W}^0 there is no indirect effect, only direct effects, and for \mathbf{W}^1 there is no direct effect, only indirect. To see this, note that when $q = 0$ we obtain $\mathbf{W}^0 = \mathbf{I}_n$:

```
Ws %~% 0

##      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
## [1,]    1    0    0    0    0    0    0
## [2,]    0    1    0    0    0    0    0
## [3,]    0    0    1    0    0    0    0
## [4,]    0    0    0    1    0    0    0
## [5,]    0    0    0    0    1    0    0
## [6,]    0    0    0    0    0    1    0
## [7,]    0    0    0    0    0    0    1
```

Thus, we have $\mathbf{S}_r(\mathbf{W}) = \mathbf{I}_n \beta_r = 0.1350 \mathbf{I}_n$. When $q = 1$ we have only indirect effect since there are zero elements on the diagonal of the matrix \mathbf{W} . This also occurs for $q = 3, 5, 7, 9$:

```
Ws %~% 1

##      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
## [1,] 0.0  1.0  0.0  0.0  0.0  0.0  0.0
## [2,] 0.5  0.0  0.5  0.0  0.0  0.0  0.0
## [3,] 0.0  0.5  0.0  0.5  0.0  0.0  0.0
## [4,] 0.0  0.0  0.5  0.0  0.5  0.0  0.0
## [5,] 0.0  0.0  0.0  0.5  0.0  0.5  0.0
## [6,] 0.0  0.0  0.0  0.0  0.5  0.0  0.5
## [7,] 0.0  0.0  0.0  0.0  0.0  1.0  0.0
```

```
Ws %~% 3

##      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
## [1,] 0.000 0.750 0.000 0.250 0.000 0.000 0.000
## [2,] 0.375 0.000 0.500 0.000 0.125 0.000 0.000
## [3,] 0.000 0.500 0.000 0.375 0.000 0.125 0.000
## [4,] 0.125 0.000 0.375 0.000 0.375 0.000 0.125
## [5,] 0.000 0.125 0.000 0.375 0.000 0.500 0.000
## [6,] 0.000 0.000 0.125 0.000 0.500 0.000 0.375
## [7,] 0.000 0.000 0.000 0.250 0.000 0.750 0.000
```

Also, the row-stochastic nature of \mathbf{W} leads to an average of the sum of the rows that takes the form $\beta_r \times \rho = 0.135 \times 0.642 = 0.0867$, when $q = 1$.

The matrix `out` also shows that both direct and indirect effects fall out as the order of neighbors increases, however the indirect or spatial spillovers effects decay more slowly as we move to higher-order neighbors.

2.6 Exercises

Exercise 2.1 Assume three regions with row-normalized spatial weight matrix given in Equation (2.26). Derive the total, direct and indirect effects for the following models:

(a) Spatial Durbin Model given by:

$$\mathbf{y} = \rho \mathbf{W} \mathbf{y} + \alpha \mathbf{1}_n + \mathbf{X} \boldsymbol{\beta} + \mathbf{W} \mathbf{X} \boldsymbol{\theta} + \boldsymbol{\varepsilon} \quad (2.40)$$

(b) Spatial Lag Model given by:

$$\mathbf{y} = \rho \mathbf{W} \mathbf{y} + \alpha \mathbf{1}_n + \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2.41)$$

(c) Spatial Durbin Error Model given by:

$$\mathbf{y} = \alpha \mathbf{1}_n + \mathbf{X} \boldsymbol{\beta} + \mathbf{W} \mathbf{X} \boldsymbol{\theta} + \mathbf{u} \quad (2.42)$$

$$\mathbf{u} = \lambda \mathbf{W} \mathbf{u} + \boldsymbol{\varepsilon} \quad (2.43)$$

(d) OLS given by:

$$\mathbf{y} = \alpha \mathbf{1}_n + \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2.44)$$

(e) Spatial Error model given by:

$$\begin{aligned} \mathbf{y} &= \alpha \mathbf{1}_n + \mathbf{X} \boldsymbol{\beta} + \mathbf{u} \\ \mathbf{u} &= \lambda \mathbf{W} \mathbf{u} + \boldsymbol{\varepsilon} \end{aligned} \quad (2.45)$$

Exercise 2.2 Consider your results for the SLM and SDM models from Exercise 2.1. Show that for the SLM model the ratio between the indirect and the direct effect of a particular explanatory variable is independent of β_k . Show that this is not the case for the SDM model. What do you conclude?

Exercise 2.3 Recall that if the row-sums of \mathbf{W} is less than or equal to one and ρ is in the proper parameter space, i.e., $\rho < 1$, the total average effect for variable r can be computed as $\beta_r / (1 - \rho)$. What is the sign of the parameter that matters the most when calculating the sign of the total effect? Does the ρ or β_r ?

Part II

Estimation Methods

Review of Asymptotic Theory

This chapter provides some basic definitions and concepts for asymptotic theory.

3.1 Convergence of Deterministic Sequences

In order to understand the asymptotic behavior of stochastic sequences we need first to refresh some concepts about deterministic (non-random) sequences. Recall that a sequence of nonstochastic real numbers $\{a_n\}$ converges to a if for any $\epsilon > 0$, there exists $n^* = n^*(\epsilon)$ such that for all $n > n^*$,

$$|a_n - a| < \epsilon,$$

e.g., if $a_n = 2 + 3/n$, then the limit is 2 since $|a_n - a| = |2 + 3/n - 2| = |3/n| < \epsilon$ for all $n > n^* = 3/\epsilon$.

Definition 3.1.1 give us a formal statement regarding nonstochastic sequence of numbers.

Definition 3.1.1 — Deterministic convergence. The sequence $\{b_n : n = 1, 2, \dots\}$ of real numbers converges to the limit b if for every $\epsilon > 0$ there exists and $n^*(\epsilon)$ such that if $n > n^*(\epsilon)$ then $|b_n - b| < \epsilon$. This is also indicated as follows:

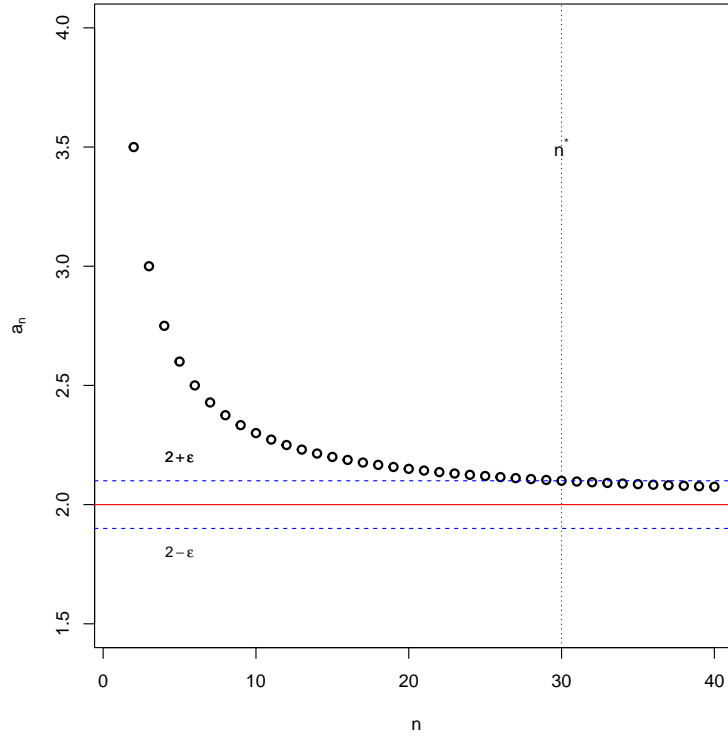
$$\lim_{n \rightarrow \infty} b_n = b$$

In Definition 3.1.1 by choosing a very small ϵ , we ensure that b_n gets arbitrarily close to its limit b for all n that is sufficiently large. In fact, the smaller ϵ is, the larger $n(\epsilon)$ will be. So, ϵ can be interpreted as a prespecified tolerance level for the discrepancy between b_n and b . When a limit exists, we say that the sequence $\{b_n\}$ **converges** to b as n tends to infinity, written $b_n \rightarrow b$ as $n \rightarrow \infty$.

Figure 3.1 shows that the sequence $2 + 3/n$ converges to 2. Note that if $\epsilon = 0.1$ then it is always true that a_n will be always between $2 + \epsilon$ and $2 - \epsilon$ if and only if $n \geq n^* = 30$.

In econometric (and specially in spatial econometrics) we talk a lot about sequences of matrices. Probably you are asking yourself, what is a sequence of matrices? Hopefully, the following example will give you some intuition.

■ **Example 3.1 — A sequence of Matrices.** Let \mathbf{X}_n be an $n \times 2$ matrix whose i th row is defined by the 1×2 vector $[1, i]$ so that

Figure 3.1: Convergence of sequence $2 + 3/n$ 

Notes: This graphs shows the convergence of the sequence $2 + 3/n$ where $\epsilon = 0.1$ and $a = 2$.

$$\mathbf{X}_n = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & n \end{pmatrix}$$

Then

$$\left\{ \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 3/2 \\ 3/2 & 5/2 \end{pmatrix}, \begin{pmatrix} 1 & 2 \\ 2 & 14/3 \end{pmatrix}, \dots \right\}$$

is a sequence of matrices $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3, \dots$ defined by the function $\mathbf{Y}_n = \frac{1}{n} \mathbf{X}_n^\top \mathbf{X}_n$, where the n th element of the sequence is defined as

$$\mathbf{Y}_n = \begin{pmatrix} 1 & \frac{\sum_{i=1}^n i}{n} \\ \frac{\sum_{i=1}^n i}{n} & \frac{\sum_{i=1}^n i^2}{n} \end{pmatrix} = \begin{pmatrix} 1 & \frac{(n+1)}{2} \\ \frac{(n+1)}{2} & \frac{(n+1)(2n+1)}{6} \end{pmatrix}$$

■

Now, we formally state the concept of convergence for matrices.

Definition 3.1.2 — Limit of a Real-Valued Matrix Sequence. Let $\{\mathbf{X}_n\}$ be a sequence whose elements are $q \times k$ real-valued matrices. Suppose there exists a $q \times k$ matrix of real numbers \mathbf{X} such that $\mathbf{X}_n[i, j] \rightarrow \mathbf{X}[i, j]$ for $i = 1, \dots, q$ and $j = 1, \dots, k$. Then the matrix \mathbf{X} is the limit of the matrix sequence $\{\mathbf{X}_n\}$ as $n \rightarrow \infty$. If the limit does not exist, the sequence is said to be divergent

The definition of the limit implies that for a sufficiently large choice of n , the matrix \mathbf{X}_n becomes arbitrarily close to the matrix \mathbf{X} , **element by element**.

Often we wish to consider the limit of a continuous function of a sequence.

Definition 3.1.3 — Limit of a continuous function of a sequence. Given $\mathbf{g} : \mathbb{R}^k \rightarrow \mathbb{R}^l$ ($k, l \in \mathbb{N}$) and $\mathbf{b} \in \mathbb{R}^k$,

- (a) the function \mathbf{g} is continuous at \mathbf{b} if for any sequence $\{\mathbf{b}_n\}$ such that $\mathbf{b}_n \rightarrow \mathbf{b}$, $\mathbf{g}(\mathbf{b}_n) \rightarrow \mathbf{g}(\mathbf{b})$;
- (b) or equivalently, the function \mathbf{g} is continuous at \mathbf{b} if for every $\epsilon > 0$ there exists $\delta(\epsilon) > 0$ such that if $\mathbf{a} \in \mathbb{R}^k$ and $|a_i - b_i| < \delta(\epsilon)$, $i = 1, \dots, k$, then $|g_j(\mathbf{a}) - g_j(\mathbf{b})| < \epsilon$, $j = 1, \dots, l$. Further, if $B \subset \mathbb{R}^k$, then \mathbf{g} is continuous on B if it is continuous at every point of B .

■ **Example 3.2** If $\mathbf{a}_n \rightarrow \mathbf{a}$ and $\mathbf{b}_n \rightarrow \mathbf{b}$, then $\mathbf{a}_n + \mathbf{b}_n \rightarrow \mathbf{a} + \mathbf{b}$ and $\mathbf{a}_n \mathbf{b}_n^\top \rightarrow \mathbf{a} \mathbf{b}^\top$. ■

■ **Example 3.3** The matrix inverse function is continuous at every point that represents a nonsingular matrix, so that if $\mathbf{X}^\top \mathbf{X}/n \rightarrow \mathbf{M}$, a finite nonsingular matrix, then $(\mathbf{X}^\top \mathbf{X}/n)^{-1} \rightarrow \mathbf{M}^{-1}$. ■

Sometimes, some sequences does not have a limit, but we can say whether they are **bounded**:

Definition 3.1.4 — Bounded sequence. A sequence $\{b_n : n = 1, 2, \dots\}$ is *bounded* if and only if there is some $a < \infty$ such that $|b_n| \leq a$ for all $n = 1, 2, \dots$. Otherwise, we say that $\{b_n\}$ is *unbounded*.

Thus, for a sequence of real numbers to be bounded, there must exist a positive number that is larger than the absolute value of each and every number in the sequence. For a sequence that has no limit and is also unbounded, we write $b_n \rightarrow \infty$, denoting that the sequence diverges to infinity.

■ **Example 3.4 — Bounded Sequences.** Consider $a_n = (-1)^n$, then a_n does not have a limit, but it is bounded since $-1 \leq a_n \leq 1$. The sequence $a_n = 1/n$ is bounded, since $0 \leq a_n \leq 1$ for all $n = 1, 2, \dots$. ■

■ **Example 3.5 — Boundedness and Limit of Matrices.** Consider the following examples:

- (a) Recall the sequence of matrices in Example 3.1. In this case, only the sequence $\{\mathbf{Y}_n[1, 1]\}$ is bounded. All other sequences of matrix elements are unbounded and, in fact, diverge to infinity. Since all the sequences of matrix elements must be bounded for the matrix sequence to converge, the matrix does not have a limit.
- (b) Let $\{\mathbf{X}_n\}$ be a sequence of matrices such that

$$\mathbf{X}_n = \begin{pmatrix} 3n^{-1} & n^{-1} \\ 3 & 1 + n^{-1} \end{pmatrix}.$$

All four sequences of the matrix elements are bounded, since $|3n^{-1}| \leq 3$, $|n^{-1}| \leq 1$, $|3| \leq 3$, and $|1 + n^{-1}| \leq 2$, for all n . Furthermore, limits exists for all four sequences of matrix elements, since $3n^{-1} \rightarrow 0$, $n^{-1} \rightarrow 0$, $3 \rightarrow 3$, and $1 + n^{-1} \rightarrow 1$. Thus

$$\mathbf{X}_n \rightarrow \mathbf{X} = \begin{pmatrix} 0 & 0 \\ 3 & 1 \end{pmatrix}$$

■

Often it is useful to have a measure of the *order of magnitude* of a particular sequence without particularly worrying about its convergence.

Definition 3.1.5 — Big and little O. Consider the following definitions:

- (a) A sequence $\{x_n\}$ is $O(n^\lambda)$ (at most of order n^λ) if $n^{-\lambda}x_n$ is bounded. When $\lambda = 0$, $\{x_n\}$ is bounded, and we also write $x_n = O(1)$.
- (b) $\{x_n\}$ is $o(n^\lambda)$ if $n^{-\lambda}x_n \rightarrow 0$. When $\lambda = 0$, x_n converges to zero, and we also write $a_n = o(1)$.
- (c) If $\{X_n[i, j]\}$ is $O(n^\lambda)$ or $o(n^\lambda)$ for all i and j , then the matrix sequence $\{\mathbf{X}_n\}$ is said to be $O(n^\lambda)$ or $o(n^\lambda)$.

The big O notation describes the asymptotic behavior of functions. Basically, it tells you how fast a function grows or declines.



From the definitions we can say that if $X_n = o(n^\lambda)$, then $X_n = O(n^\lambda)$. In other words, **any convergent sequence is bounded**. The opposite is not true. Recall the example $a_n = (-1)^n$.

■ **Example 3.6 — Order of Magnitude of a Sequence.** Consider the following examples:

- (a) Let $\{x_n\}$ be defined by $x_n = 3n^3 - n^2 + 2$. Then $\{x_n\}$ is $O(n^3)$, since $n^{-3}x_n = 3 - n^{-1} + 2n^{-3}$ is bounded. Also $\{x_n\}$ is $o(n^{3+\epsilon})$ for any $\epsilon > 0$ since $n^{-3-\epsilon}x_n = 3n^{-\epsilon} - n^{-1-\epsilon} + 2n^{-3-\epsilon} \rightarrow 0$. For example, Figure 3.2 plots $n^{-3}x_n$, which is bounded between 4 ($n = 1$) and 2.75 ($n = 2$). Note also that if we choose $\epsilon = 0.1$, then $n^{3.1}x_n$ clearly converges to 0.
- (b) Let $\{x_n\}$ be defined by $x_n = 3 + n^{-1}$. Then $\{x_n\}$ is $O(1)$, since x_n is bounded, and $\{x_n\}$ is $o(n^\epsilon)$; $\forall \epsilon > 0$, since $n^{-3}x_n = 3n^{-\epsilon} + n^{-1-\epsilon} \rightarrow 0$.
- (c) Let the vector sequence $\{\mathbf{x}_n\}$ be defined by

$$\begin{pmatrix} \mathbf{x}_n[1] \\ \mathbf{x}_n[2] \end{pmatrix} = \begin{pmatrix} 3n^{-1} \\ n^{-1} \end{pmatrix}.$$

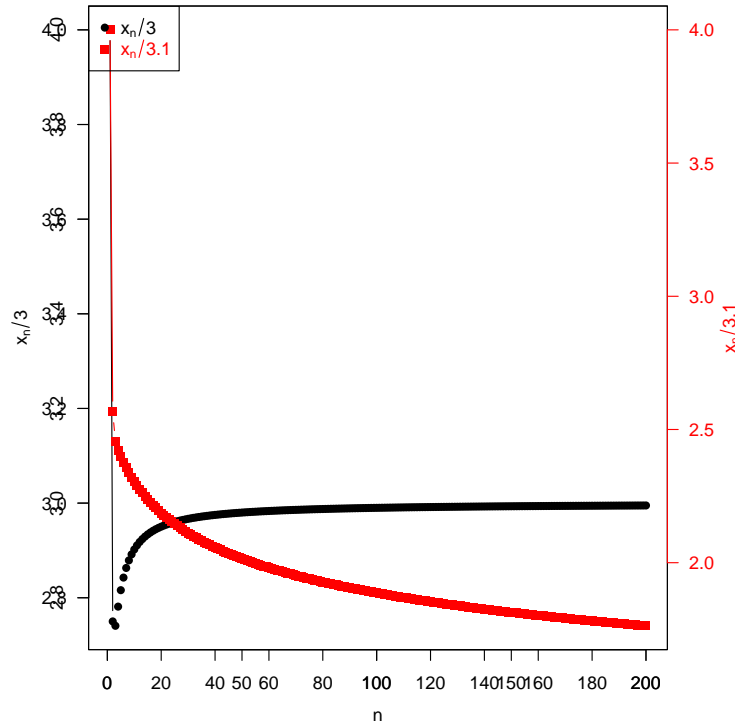
Then the vector sequence $\{\mathbf{x}_n\}$ is $o(1)$ and $O(1)$, since

$$\mathbf{x}_n \rightarrow \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

■

The following proposition gives some elementary facts about the orders of magnitude of sums and products of sequences.

Figure 3.2: Bounded sequence



Notes: This graphs shows that the sequence $\{x_n\}$ defined by $x_n = 3n^3 - n^2 + 2$ is $O(n^3)$ and $o(n^{3+\epsilon})$. For plotting $\epsilon = 0.1$ was selected.

Proposition 3.1 — Properties of big and little O. Let a_n and b_n be scalars.

- (a) If $a_n = O(n^\lambda)$ and $b_n = O(n^\mu)$, then $a_n b_n = O(n^{\lambda+\mu})$ and $a_n + b_n = O(n^k)$, where $k = \max[\lambda, \mu]$.
- (b) If $a_n = o(n^\lambda)$ and $b_n = o(n^\mu)$, then $a_n b_n = o(n^{\lambda+\mu})$ and $a_n + b_n = o(n^k)$, where $k = \max[\lambda, \mu]$.
- (c) If $a_n = O(n^\lambda)$ and $b_n = o(n^\mu)$, then $a_n b_n = o(n^{\lambda+\mu})$ and $a_n + b_n = O(n^k)$, where $k = \max[\lambda, \mu]$.

3.2 Convergence in Probability

In the previous section we reviewed how a sequence of real number converges to a real number. What about the sequence of random variables such as econometric estimators? When considering a sequence of *random variables* we cannot be certain that $|a_n - a| < \epsilon$, even for large n , due to the **randomness**. Instead, we require that **the probability of being within ϵ is arbitrarily close to one** as $n \rightarrow \infty$. The next definition is more appropriate for convergence in random variables.

Definition 3.2.1 — Convergence in Probability. A sequence of random variables $\{X_n\}$ **convergence in probability** to a constant (non-random) α if, for any $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} \Pr(|X_n - \alpha| > \epsilon) = 0$$

The constant α is called the **probability limit** of X_n and is written as $\text{plim } X_n = \alpha$ or $X_n \xrightarrow{p} \alpha$. Evidently,

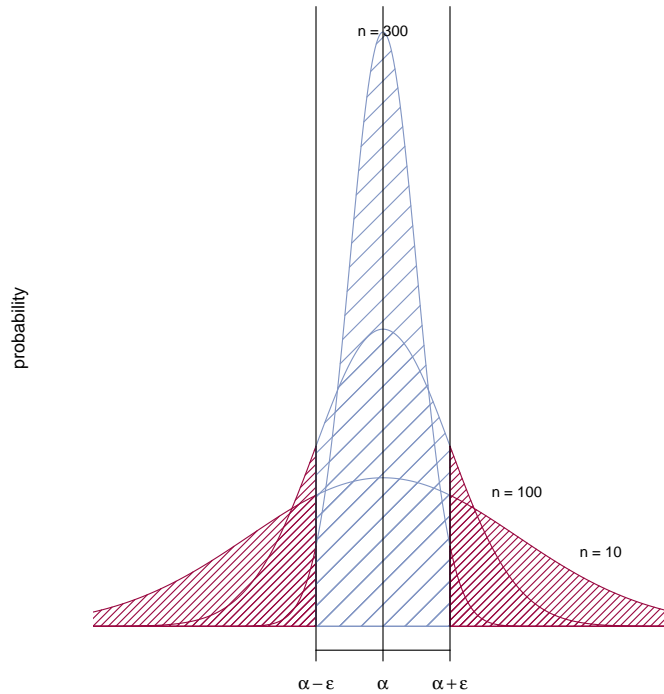
$$X_n \xrightarrow{p} \alpha \quad \text{is the same as} \quad X_n - \alpha \xrightarrow{p} 0$$

Thus, roughly, convergence in probability states that for large n , the probability is high that X_n will be close α .

This definition can be understood if we look at Figure 3.3. Note that the expression $|X_n - \alpha| > \epsilon$ can be true or false. The probability that it is true is given by the distribution $F_n(\cdot)$ of X_n . Figure 3.3 shows that the probability that $|X_n - \alpha| > \epsilon$, denoted by the red-dashed area outside the interval $\alpha \pm \epsilon$, becomes smaller as n increases. Conversely, the probability of $|X_n - \alpha| < \epsilon$, given by the blue-dashed area, will become higher and higher as $n \rightarrow \infty$. In the limit, this probability should be equal to 1. That is:

$$\lim_{n \rightarrow \infty} \Pr(|X_n - \alpha| < \epsilon) = 1$$

Figure 3.3: Illustration of convergence in probability to a constant



Notes: This graphs shows that the probability of $|X_n - \alpha| > \epsilon$, which is denoted by the red-dashed areas, becomes smaller as n increases.

Definition (3.2.1) can be easily extended to a sequence of random vectors or random matrices (by viewing a matrix as a vector whose elements have been rearranged) by requiring element-by-element convergence in probability. That is, a sequence of k -dimensional random vectors $\{\mathbf{x}_n\}$ converges in probability to a k -dimensional vector of constants $\boldsymbol{\alpha}$ if, for any $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} \Pr(\|\mathbf{x}_n - \boldsymbol{\alpha}\| > \epsilon)$$

Note that $\|\mathbf{x}_n - \boldsymbol{\alpha}\|$ is the Euclidean distance

$$\left[(\mathbf{x}_n - \boldsymbol{\alpha})' (\mathbf{x}_n - \boldsymbol{\alpha}) \right]^{1/2} = \sqrt{(x_{1n} - \alpha_1)^2 + \dots + (x_{Kn} - \alpha_K)^2} = \|\mathbf{x}_n - \boldsymbol{\alpha}\|$$

Therefore,

$$\mathbf{x}_n \xrightarrow{p} \boldsymbol{\alpha} \quad \text{iff} \quad \Pr \left[\sqrt{\sum_{j=1}^k (x_{j,n} - \alpha_j)^2} > \epsilon \right] \xrightarrow{p} 0$$

as $n \rightarrow \infty$ for $\epsilon > 0$ and $\forall j = 1, \dots, k$, where:

$$\mathbf{x}_n = \begin{pmatrix} x_{1n} \\ \vdots \\ x_{kn} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_k \end{pmatrix}$$

R $\mathbf{x}_n \xrightarrow{p} \boldsymbol{\alpha}$ if and only if $x_{jn} \xrightarrow{p} \alpha_j$ for $j = 1, \dots, k$. That is, vector convergence in probability is equivalent to component convergence in probability for each component. See our previous discussion of vector sequence.

Definition 3.2.2 — Probability Limits of Matrices (and Vectors for $k = 1$). Let $\{\mathbf{Y}_n\}$ be a sequence of $m \times k$ random matrices. Then

$$\text{plim} \begin{pmatrix} Y_n[1, 1] & \dots & Y_n[1, k] \\ \vdots & \ddots & \vdots \\ Y_n[m, 1] & \dots & Y_n[m, k] \end{pmatrix} = \begin{pmatrix} \text{plim } Y_n[1, 1] & \dots & \text{plim } Y_n[1, k] \\ \vdots & \ddots & \vdots \\ \text{plim } Y_n[m, 1] & \dots & \text{plim } Y_n[m, k] \end{pmatrix}$$

The expectation $\mathbb{E}(\cdot)$ is a linear operator, that is, **we cannot** state that $\mathbb{E}[\exp(\hat{\theta})] = \exp[\mathbb{E}(\hat{\theta})]$. Thus, we would like to know if the plim has the same property. Fortunately, the continuous mapping theorem tell us that we can interchange them.

Theorem 3.2 — Continuous Mapping Theorem. Given a continuous function $g(X)$, if $X_n \xrightarrow{p} X$ then $g(X_n) \xrightarrow{p} g(X)$ as $n \rightarrow \infty$, or equivalently, $\text{plim}[g(X_n)] = g[\text{plim}(X_n)]$.

The Continuous Mapping Theorem is a very useful theorem. Unlike the expectation operator, it shows that the plim operator passes through nonlinear functions, provided they are continuous. The lack of this property for the \mathbb{E} operator makes finite sample analysis difficult for many estimators.

It is useful to know the vector form of this Theorem. Let $\mathbf{g} : \mathbb{R}^K \rightarrow \mathbb{R}^J$ be a function continuous at some point $\boldsymbol{\alpha} \in \mathbb{R}^K$. Then,

$$\mathbf{x}_n \xrightarrow{p} \boldsymbol{\alpha} \implies \mathbf{g}(\mathbf{x}_n) \xrightarrow{p} \mathbf{g}(\boldsymbol{\alpha}),$$

if $\mathbf{g}(\cdot)$ is continuous at $\text{plim } \mathbf{x}_n$.

Now that we have presented the meaning of convergence in probability, it is time to define what we understand for “consistency” in econometrics.

Definition 3.2.3 — Consistent Estimator. An estimator $\hat{\theta}_n$ of a parameter θ is a consistent estimator θ if and only if

$$\text{plim } \hat{\theta}_n = \theta,$$

which can also be written as:

$$\hat{\theta}_n \xrightarrow{p} \theta.$$

In words, a **consistent estimator** is an estimator—a rule for computing estimates of a parameter θ —having the property that as the number of data used increases without bound, the resulting sequence of estimates converges in probability to θ . This means that the distributions of the estimates become more and more concentrated near the true value of the parameters being estimated, so that the probability of the estimator being arbitrary close to θ converges to one.

R Convergence in probability is also referred to as weak consistency, and since this has been the most familiar stochastic convergence concept in econometric, the word “weak” if often simply dropped.

■ **Example 3.7** In this example, we will show that the OLS estimator is consistent under the following assumptions:

- (a) $y_i = \mathbf{x}_i^\top \beta_0 + \epsilon_i$, $i = 1, \dots, n$; $\beta_0 \in \mathbb{R}^k$;
- (b) $\mathbf{X}^\top \boldsymbol{\epsilon} / n \xrightarrow{p} \mathbf{0}$;
- (c) $\mathbf{X}^\top \mathbf{X} / n \xrightarrow{p} \mathbf{M}$, finite and positive definite.

The sampling error is:

$$\hat{\beta}_n = \beta_0 + \left(\frac{\mathbf{X}^\top \mathbf{X}}{n} \right)^{-1} \frac{\mathbf{X}^\top \boldsymbol{\epsilon}}{n}. \quad (3.1)$$

Since $\mathbf{X}^\top \mathbf{X} / n \xrightarrow{p} \mathbf{M}$, it follows from Theorem 3.2 that

$$\det \left(\frac{\mathbf{X}^\top \mathbf{X}}{n} \right) \xrightarrow{p} \det(\mathbf{M}). \quad (3.2)$$

Because \mathbf{M} is positive definite, $\det \det(\mathbf{M}) > 0$. It follows that for all n sufficiently large $\det \det \left(\frac{\mathbf{X}^\top \mathbf{X}}{n} \right) > 0$, so $\left(\frac{\mathbf{X}^\top \mathbf{X}}{n} \right)^{-1}$ exists for all n sufficiently large. Hence $\hat{\beta}_n$ in Equation (3.1) exists for all n sufficiently large. It follows from Theorem 3.2 that

$$\hat{\beta}_n \xrightarrow{p} \beta_0 + \mathbf{M}^{-1} \mathbf{0} = \beta_0, \quad (3.3)$$

given (b) and (c) ■

Definition 3.2.4 — Strong Convergence in Probability. A sequence of random variables $\{X_n\}$ **convergence in probability strongly, or, almost surely** to a constant (non-random) α if, for any $\epsilon > 0$,

$$\Pr \left(\lim_{n \rightarrow \infty} X_n = \alpha \right) = 1$$

This is written $X_n \xrightarrow{a.s.} \alpha$, as $n \rightarrow \infty$. An equivalent condition for almost sure convergence is

$$\lim_{n \rightarrow \infty} \Pr (|X_m - X| < \epsilon, \forall m \geq n) = 1$$

The extension to random vector is analogous to that for convergence in probability. Note also that this concept is stronger than convergence in probability; that is, if a sequence converges almost surely, then it converges in probability.

$$\textcircled{R} \quad \xrightarrow{a.s.} \implies \xrightarrow{p}$$

3.2.1 Convergence in Quadratic Mean

We will make frequent use of a special case of convergence in probability, **convergence in mean square** or **convergence in quadratic mean**

Theorem 3.3 — Convergence in Quadratic Mean. If X_n has mean μ_n and variance σ_n^2 such that the ordinary limits of μ_n and σ_n^2 are c and 0, respectively, then X_n converges in mean square to c ,

$$X_n \xrightarrow{q.m.} c$$

and

$$\text{plim } X_n = c.$$

This theorem implies that $X_n \xrightarrow{q.m.} c \implies X_n \xrightarrow{p} c$. The conditions for convergence in mean square are usually easier to verify than those for the more general form.

The vector form of this type of convergence is the following. We say that the sequence of random vectors $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ converges in quadratic mean to the random vector \mathbf{z} if $\mathbb{E}(\mathbf{x}_n \mathbf{x}_n')$ and $\mathbb{E}(\mathbf{z} \mathbf{z}')$ exists for all n if

$$\lim_{n \rightarrow \infty} \mathbb{E} [(\mathbf{x}_n - \mathbf{z})' (\mathbf{x}_n - \mathbf{z})] = \mathbf{0}$$

A special case of convergence in quadratic mean occurs when \mathbf{z} , instead of being a random vector, is a vector of unknown parameters, say $\boldsymbol{\theta}$, and \mathbf{x}_n is an estimator for $\boldsymbol{\theta}$. Under these circumstances we can write:

$$\begin{aligned} \mathbb{E} [(\mathbf{x}_n - \mathbf{z})' (\mathbf{x}_n - \mathbf{z})] &= (\mathbb{E} [\mathbf{x}_n] - \boldsymbol{\theta})' (\mathbb{E} [\mathbf{x}_n] - \boldsymbol{\theta}) + \mathbb{E} [(\mathbf{x}_n - \mathbb{E} [\mathbf{x}_n])' (\mathbf{x}_n - \mathbb{E} [\mathbf{x}_n])] \\ &= \sum_{k=1}^K \text{bias}^2(x_{kn}) + \sum_{k=1}^K \mathbb{V}(x_{kn}) \end{aligned} \quad (3.4)$$

where x_{kn} is the k th element of \mathbf{x}_n that is assumed to be K dimensional. Thus from (3.4) \mathbf{x}_n converges to $\mathbf{0}$ in quadratic mean if and only if the bias and variance of \mathbf{x}_n approach zero

as $n \rightarrow \infty$. This result, and the fact that Chebyshev's inequality can be used to prove that convergence in quadratic mean implies convergence in probability. See below.

An useful theorem is the following:

Theorem 3.4 — Consistency of the sample mean. The mean of a random sample from any population with finite mean μ and finite variance σ^2 is a consistent estimator of μ .

Proof of consistency of the sample mean. Since $\mathbb{E}(\bar{X}_n) = \mu$ and $\mathbb{V}(\bar{X}_n) = \sigma^2/n$. Therefore, using Theorem 3.3 (Convergence in quadratic mean)

$$\bar{X}_n \xrightarrow{q.m.} \mu \implies \bar{X}_n \xrightarrow{p} \mu$$

■

Theorem 3.5 — Sufficient Conditions for Consistency. Chebyshev's inequality implies that a sufficient conditions for an estimator based on a sample of size n , say $\hat{\theta}_n$, say to be consistent for θ are:

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{E}(\hat{\theta}_n) &= \theta_0 \\ \lim_{n \rightarrow \infty} \mathbb{V}(\hat{\theta}_n) &= 0 \end{aligned}$$

If these two requirements are met, then:

$$\hat{\theta}_n \xrightarrow{p} \theta$$

Proof of consistency of unbiased estimator. Since $\hat{\theta}_n$ is unbiased, using Chebyshev's inequality 3.A.4 we obtain:

$$\Pr [|\hat{\theta}_n - \theta| \geq \delta] \leq \frac{\mathbb{V}(\hat{\theta}_n)}{\delta^2}$$

If $\lim_{n \rightarrow \infty} \mathbb{V}(\hat{\theta}_n) = 0$, then $\Pr [|\hat{\theta}_n - \theta| \geq \delta] \rightarrow 0$, so $\hat{\theta}_n \xrightarrow{p} \theta$ ■

■ **Example 3.8** For the normal case, we have that $\mathbb{E}(s^2) = \sigma^2$ and $\mathbb{V}(s^2) = 2\sigma^4/(n-1) \rightarrow 0$ as $n \rightarrow \infty$, hence $s^2 \xrightarrow{p} \sigma^2$ ■

■ **Example 3.9** For the Bernoulli case, we know that $\mathbb{E}(\bar{X}) = \theta$ and $\mathbb{V}(\bar{X}) = \theta(1-\theta)/n \rightarrow 0$ as $n \rightarrow \infty$, hence $\bar{X} \xrightarrow{p} \theta$ ■

Therefore, another alternative method for proving that some estimator $\hat{\theta}$ is consistent is to demonstrate that its unbiased and its covariance matrix approaches zero as $n \rightarrow \infty$.



Theorem 3.5 (Consistency of Unbiased Estimator) is only a sufficient condition for consistency. Failing to satisfy this condition does not necessarily imply that the estimator is inconsistent.

Theorem 3.6 — Rules for probability limits. If X_n and Y_n are random variables with $X_n \xrightarrow{p} c$ and $Y_n \xrightarrow{p} d$, then:

(a) Sum rule:

$$X_n + Y_n \xrightarrow{p} c + d \quad (3.5)$$

(b) Product rule:

$$X_n Y_n \xrightarrow{p} cd \quad (3.6)$$

(c) Ratio rule:

$$X_n/Y_n \xrightarrow{p} c/d \quad \text{if } d \neq 0 \quad (3.7)$$

(d) Matrix inverse rule: If \mathbf{W}_n is a matrix whose elements are random variables and if $\mathbf{W}_n \xrightarrow{p} \mathbf{\Omega}$, then

$$\mathbf{W}_n^{-1} \xrightarrow{p} \mathbf{\Omega}^{-1} \quad (3.8)$$

(e) Matrix product rule: If \mathbf{X}_n and \mathbf{Y}_n are random matrices with $\mathbf{X}_n \xrightarrow{p} \mathbf{A}$ and $\mathbf{Y}_n \xrightarrow{p} \mathbf{B}$, then

$$\mathbf{X}_n \mathbf{Y}_n \xrightarrow{p} \mathbf{AB} \quad (3.9)$$

■ **Example 3.10 — Plims of Scalar Additive and Multiplicative Functions.** Let $\mathbf{A} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$, and $\{\mathbf{X}_n\}$ be such that $\text{plim } \mathbf{X}_n = \begin{pmatrix} 2 \\ 5 \end{pmatrix}$. Then,

$$\text{plim } (\mathbf{AX}_n) = \mathbf{A} \text{plim } (\mathbf{X}_n) = \begin{pmatrix} 9 \\ 7 \end{pmatrix}$$

■

■ **Example 3.11 — Plims of Matrix Functions to Constant Matrices.** Let $\{\mathbf{Y}_n\}$ be such that $\text{plim } \mathbf{Y}_n = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$ and $\{\mathbf{X}_n\}$ be such that $\text{plim } \mathbf{X}_n = \begin{pmatrix} 3 & 1 \\ 2 & 1 \end{pmatrix}$. Then:

$$\text{plim } (\mathbf{X}_n \mathbf{Y}_n) = \text{plim } (\mathbf{X}_n) \text{plim } (\mathbf{Y}_n) = \begin{pmatrix} 3 & 1 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 5 & 7 \\ 4 & 5 \end{pmatrix}$$

and

$$\text{plim } (\mathbf{X}_n^{-1} \mathbf{Y}_n) = \text{plim } (\mathbf{X}_n)^{-1} \text{plim } (\mathbf{Y}_n) = \begin{pmatrix} 1 & -1 \\ -2 & 3 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} -1 & 1 \\ 4 & -1 \end{pmatrix}$$

■

3.3 Law of Large Numbers

Much of the work of an econometrician, and also of a student of econometrics, is to determine whether an estimator is consistent. Fortunately, the ‘law of large numbers’ will greatly simplify this work. Roughly speaking, law of large numbers (LLN) are theorems for convergence in probability in the special case where the sequence $\{X_n\}$ is a sample average, i.e., $X_n = \bar{X}_n$ where:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i.$$

Thus, a LLN provides a much easier way to establish the probability limit of a sequence than the alternatives of the (δ, ϵ) definition of the probability given previously.

Let us start with the simplest LLN's definition.

Theorem 3.7 — Khinchine's Weak Law of Large Numbers. Let $\{X_n\}$ be an i.i.d random sample with $\mathbb{E}(X_i) = \mu$, and let $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$. Then:

$$\lim_{n \rightarrow \infty} \Pr \left[\left| \bar{X}_n - \mu \right| > \epsilon \right] = 0$$

or equivalently,

$$\lim_{n \rightarrow \infty} \Pr \left[\left| \bar{X}_n - \mu \right| \leq \epsilon \right] = 1.$$

In other words,

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{p} \mathbb{E}(X_i)$$

or $\text{plim } \bar{X}_n = \mu$

The WLLN shows that the estimator $\hat{\mu} = \bar{X}_n$ converges in probability to the true population mean μ . Another important feature of this theorem is that it does not require the existence of moments higher order than the mean. This is a powerful result that is very convenient when we have an i.i.d sample. Moreover, this theorem will simplify our proofs when we encounter sample moments such as $n^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'$.

■ **Example 3.12** Consider a random sampling from a population with mean μ_n and variance σ_n^2 . What is the probability limit of $\hat{\theta}_n = \bar{x}_n^2 / s_n^2$? By the ratio rule in Theorem 3.6

$$\text{plim } \frac{\bar{x}_n^2}{s_n^2} = \frac{\text{plim } \bar{x}_n^2}{\text{plim } s_n^2}$$

Note that

$$\begin{aligned} \text{plim } \bar{x}_n^2 &= (\text{plim } \bar{x}_n)^2 \quad \text{by Theorem 3.2} \\ &= \mu^2 \quad \text{by LLN 3.7} \end{aligned}$$

Since s_n^2 is consistent $s_n^2 \xrightarrow{p} \sigma^2$, then

$$\text{plim } \frac{\bar{x}_n^2}{s_n^2} = \frac{\mu^2}{\sigma^2}$$

■

R Theorem 3.7 (Khinchine's Weak Law of Large Numbers) is widely used in econometric because the estimators involve averages. Note also that LLN is much easier way to get the plim than use of Definition 3.2.1 (Convergence in Probability) or Theorem 3.3 (Convergence in Quadratic Mean).

■ **Example 3.13 — Example of mean from normal.** Consider we have n different samples with pdf $N(1, 0.5^2)$. For example X_1 is the first sample with just one observation that comes from a $N(1, 0.5^2)$, X_2 is the second sample with two observations (X_1, X_2) which also comes from

a $N(1, 0.5^2)$; X_3 with three observations (X_1, X_2, X_3) and so on. Note that each sample (or sequence) is a i.i.d. random sample with $\mathbb{E}(X_i) = \mu$. The mean for each sequence is also a sequence:

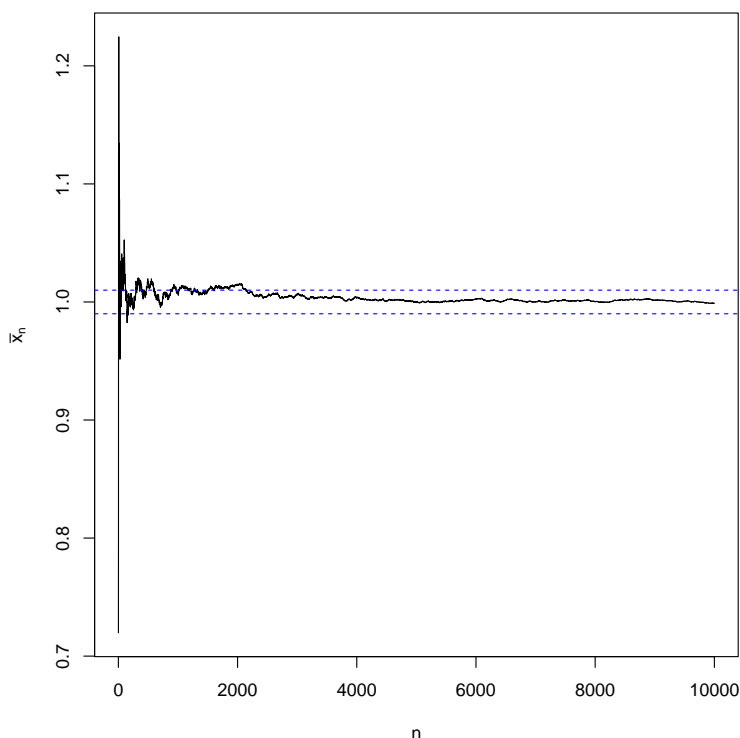
$$\begin{aligned}\bar{X}_1 &= g(X_1) = X_1 \\ \bar{X}_2 &= g(X_1, X_2) = \frac{1}{2} \sum_{i=1}^2 X_i \\ \bar{X}_3 &= g(X_1, X_2, X_3) = \frac{1}{3} \sum_{i=1}^3 X_i \\ &\vdots \\ \bar{X}_n &= g(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i\end{aligned}$$

Now in R, we will show that the mean will converge to the true population mean $\mu = 1$, as $n \rightarrow \infty$.

```
# Setup
set.seed(123) # set the seed
N <- 10000    # total number of observations
n <- 1:N      # vector: n = 1, 2, ..., N

n_dat <- rnorm(n = n, mean = 1, sd = 0.5) # Sample from N(1, 0.5^2)
xbar <- cumsum(n_dat) / n                  # Cumulated mean
plot(n, xbar, type = "l", ylab = expression(bar(x)[n]))
abline(h = 1.01, col = "blue", lty = 2)
abline(h = 0.99, col = "blue", lty = 2)
```

Figure 3.4: Convergence of mean from normal distribution



Notes: This graphs shows the convergence of \bar{X} as $n \rightarrow \infty$ for a normal distribution.

From Figure 3.4 we can see that \bar{X}_n gets arbitrarily close to μ as n increases indefinitely. In words, as the sample size n increases, the sample mean converges to the theoretical mean.

■

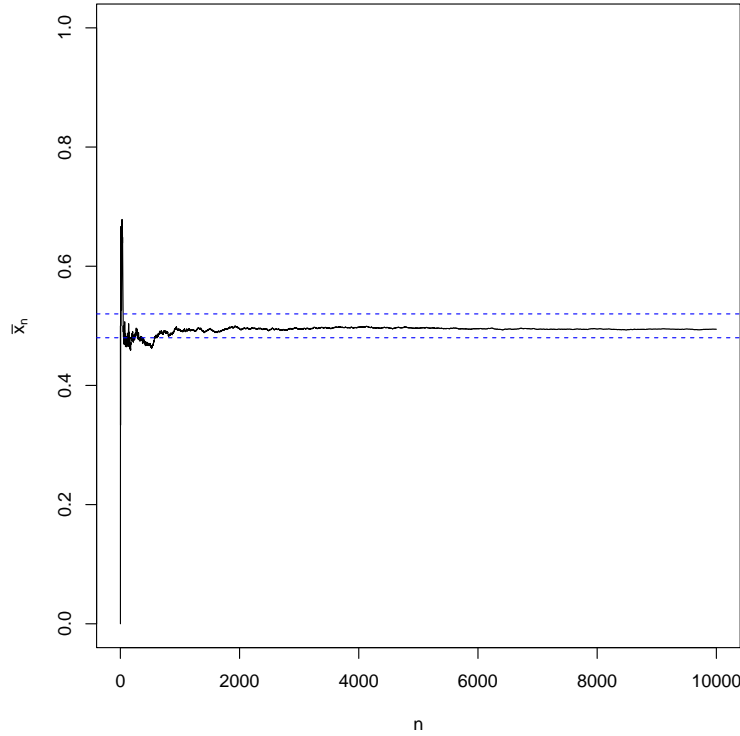
■ **Example 3.14 — Tossing a fair coin.** Now we simulate $n = 1000$ coin tosses. After each simulated toss, we plot the proportion X_n of heads obtained so far against the number n of tosses so far. The LLN says we should see a trace that gets very close to $1/2$ as n increases.

In R, the code is:

```
# Set up
set.seed(123)                                # set seed
N <- 10000                                    # total number of tosses
n <- 1:N                                       # vector: n = 1, 2, ..., N; Toss number

# Simulate and plot
h <- rbinom(n = n, size = 1, prob = 1/2)      # vector: H = 0 or 1 each with p = 1./2
x <- cumsum(h) / n                           # vector: proportion of heads
plot(n, x, type = "l", ylim = c(0, 1), ylab = expression(bar(x)[n]))
abline(h = 0.52, col = "blue", lty = 2)
abline(h = 0.48, col = "blue", lty = 2)
```


Figure 3.5: Convergence of mean from binomial distribution



Notes: This graphs shows the convergence of \bar{X} as $n \rightarrow \infty$ for a binomial distribution.

Note that the n th element of the vector \mathbf{x} is the mean of the first n elements of \mathbf{h} . Figure 3.5 shows that the mean from a binomial distribution converges to the population mean $\mu = p = 1/2$, as $n \rightarrow \infty$. Note that the dashed lines at 0.48 and 0.52 illustrate the LLN with $\epsilon = 0.02$. ■

To apply LLN for several variables we have to know that summands of iid different random variables are also i.i.d.

Proposition 3.8 Let $\mathbf{g} : \mathbb{R}^K \rightarrow \mathbb{R}^l$ be a continuous function. (i) Let \mathbf{X}_i and \mathbf{X}_t be identically distributed. Then $\mathbf{g}(\mathbf{X}_i)$ and $\mathbf{g}(\mathbf{X}_t)$ are identically distributed. (ii) Let \mathbf{X}_i and \mathbf{X}_t be independent. Then $\mathbf{g}(\mathbf{X}_i)$ and $\mathbf{g}(\mathbf{X}_t)$ are independent.

Therefore, using this proposition we can state the following proposition:

Proposition 3.9 If $\{(\mathbf{Z}_i^\top, \mathbf{X}_i, \epsilon_i)\}$ is an i.i.d random sequence, then $\{\mathbf{X}_i \mathbf{X}_i^\top\}$, $\{\mathbf{X}_i \epsilon_i\}$, $\{\mathbf{Z}_i \mathbf{X}_i^\top\}$, $\{\mathbf{Z}_i \epsilon_i\}$, and $\{\mathbf{Z}_i \mathbf{Z}_i^\top\}$ are also i.i.d sequences.

This result is useful in situations in which we have observations from a random sample, as in a simple cross section. The result does not apply to stratified cross sections since there the observations are not identically distributed across strata, and generally will not apply to time-series data since there the observations $(\mathbf{X}_i, \epsilon_i)$ generally are not independent. For these situations, we need laws of large numbers that do not impose the i.i.d assumption.

■ **Example 3.15** In this example, we will show that the OLS estimator $\hat{\beta}_n \xrightarrow{a.s.} \beta_0$. Assume the following:

- (a) $y_i = \mathbf{x}_i^\top \beta_0 + \epsilon_i$, $i = 1, \dots, n$; $\beta_0 \in \mathbb{R}^k$;

- (b) the sample $\{y_i, \mathbf{x}_i^\top\}$ is an i.i.d sequence;
- (c) $\mathbb{E}(\mathbf{x}_i \epsilon_i) = \mathbf{0}$;
- (d) $\mathbb{E}(\mathbf{x}_i \mathbf{x}_i^\top) = \mathbf{M}$ is positive definite;

Since $\{\mathbf{x}_i\}$ is a i.i.d random sample by Assumption (b) (Random Sample), then $\{\mathbf{x}_i \mathbf{x}_i^\top\}$ is also i.i.d sequence by Proposition 3.9.

Note that each (g, j) element of the $k \times k$ matrix $\mathbf{x}_i \mathbf{x}_i^\top$ is given by

$$\sum_{h=1}^p x_{ihg} x_{ihj}.$$

By triangle inequality 3.A.2:

$$\left| \sum_{h=1}^p x_{ihg} x_{ihj} \right| \leq \sum_{h=1}^p |x_{ihg} x_{ihj}|.$$

Then, by Cauchy-Schwarz inequality 3.A.6:

$$\begin{aligned} \mathbb{E} \left| \sum_{h=1}^p x_{ihg} x_{ihj} \right| &\leq \sum_{h=1}^p \mathbb{E} |x_{ihg} x_{ihj}| \\ &\leq \sum_{h=1}^p \left\{ \left(\mathbb{E} |x_{ihg}|^2 \right)^{1/2} \left(\mathbb{E} |x_{ihj}|^2 \right)^{1/2} \right\} \end{aligned}$$

It follows that the elements of the $\mathbf{x}_i \mathbf{x}_i^\top$ will have $\mathbb{E} |\sum_{h=1}^p x_{ihg} x_{ihj}| < \infty$ provided simply that $\mathbb{E} |x_{ihg}|^2 < \infty$ for all $h = 1, \dots, p$ and $g = 1, \dots, k$. Thus, by Theorem 3.7 and assuming that $\mathbb{E} |x_{ihg}|^2 < \infty$, then

$$n^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \xrightarrow{p} \mathbf{M} \quad (3.10)$$

Similarly, $\{\mathbf{x}_i \epsilon_i\}$ is also i.i.d sequence by Proposition 3.9. Using our previous reasoning:

$$n^{-1} \mathbf{X}^\top \boldsymbol{\epsilon} = n^{-1} \sum_{i=1}^n \mathbf{x}_i \epsilon_i \xrightarrow{p} \mathbb{E}(\mathbf{x}_i \epsilon_i) = \mathbf{0} \quad (3.11)$$

if $\mathbb{E} |x_{ihg} \epsilon_{ih}| < \infty$ for all $h = 1, \dots, p$ and $g = 1, \dots, k$. From here, we proceed as Example 3.7. ■

Another important feature is that Khinchine's WLLN is broader than Theorem 3.5 (Consistency of Unbiased Estimator), as **it does not require that the variance of the distribution be finite**. On the other hand, it is not broad enough, because most of the situations we encounter where we will need a result such as this will not involve i.i.d. random sampling. A broader LLN Theorem is the following:

Theorem 3.10 — Chebychev's Weak Law of Large Numbers. If $X_i, i = 1, \dots, n$ is a sample of observations such that $\mathbb{E}(X_i) = \mu_i < \infty$ and $\mathbb{V}(X_i) = \sigma_i^2 < \infty$ such that

$$\frac{\bar{\sigma}_n^2}{n} = \frac{\sum_{i=1}^n \sigma_i^2}{n^2} \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

then

$$\bar{X}_n - \bar{\mu}_n \xrightarrow{p} 0.$$

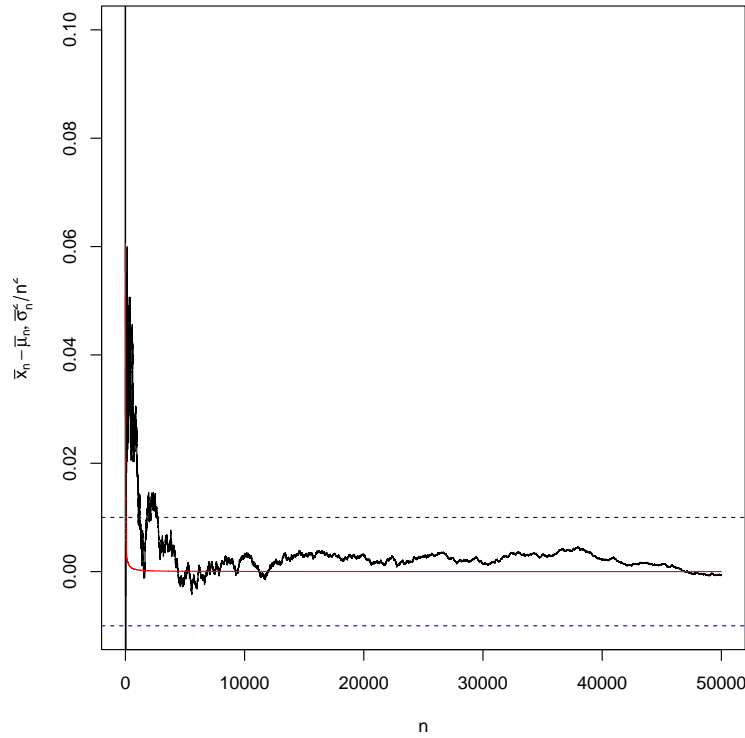
The Chebychev's theorem does not state that \bar{X}_n converges to $\bar{\mu}_n$, or even that it converges to a constant at all. The theorem states that as n increases without bound, these two quantities will be arbitrarily close to each other. In other words, the difference between them converges to a constant, zero. The more important difference between the Khinchine and Chebyshev theorems is that the second allows for heterogeneity in the distributions of the random variables that enter the mean. This will be very useful in cases where the independence assumption may hold but the identical distribution assumption does not (such as random sampling with cross-sectional data). For example, the X_i 's may have different means and/or variances for each i . If we retain the independent assumption but relax the identical distribution assumption, then we can still get convergence of the sample mean.

It is important to stress that the behavior of the variance of \bar{X}_n is the key element in this **LLN**. Independence implies that all covariances among the X_i are zero, so that the variance of \bar{X}_n simplifies to the sum of the variances of the X_i divided by n^2 . Then the key mechanism is that the variance of \bar{X}_n converges to zero:

$$\lim_{n \rightarrow \infty} \mathbb{V}[\bar{X}_n] = \lim_{n \rightarrow \infty} \frac{1}{n^2} \sum_{i=1}^n \mathbb{V}(X_i) = \lim_{n \rightarrow \infty} \frac{\sum_{i=1}^n \sigma_i^2}{n} = 0$$

To illustrate Chebychev's WLLN we have created artificial data sets from independent normal distributions with different mean and standard deviations. Figure 3.6 displays the sequence $\bar{X}_n - \bar{\mu}_n$ in black line and $\bar{\sigma}^2/n^2$ as $n \rightarrow \infty$ in a red line. It can be observed that both sequences change with sample size, but as the number of observations increases both settle down to zero. However, note that the sequence $\bar{X}_n - \bar{\mu}_n$ converges in probability, whereas $\bar{\sigma}^2/n^2 \rightarrow 0$ in a deterministic way.

Figure 3.6: Chebychev's Convergence



Notes: This graphs shows the convergence of $\bar{X}_n - \bar{\mu}_n$ and $\bar{\sigma}^2/n^2$ as $n \rightarrow \infty$.



When the iid assumption is relaxed, stronger restrictions need to be place on the variances of each of the random variables. If some assumption are weakened then other assumptions must be strengthened.

3.4 Convergence in Distribution

Definition 3.4.1 — Convergence in Distribution. If the cdfs F_{X_n} of the sequence of random variables $\{X_n\}$ converge to the cdf F_X as $n \rightarrow \infty$ at all points z where $F_X(z)$ is continuous, then $\{X_n\}$ converges in distribution to X . This will be denoted

$$X_n \xrightarrow{d} X$$

or

$$\lim_{n \rightarrow \infty} |F_{X_n} - F_X| = 0$$

This theorem states that the distribution of X_n gets closer and closer to that of the random variable X , so that the distribution of X , the cdf F_X , can be used as an **approximation** to the distribution of F_{X_n} . We can also say that X is the **limiting distribution** of X_n .

Convergence in distribution can be extended to random vectors and matrices although not in the element by element manner that we extended the earlier convergence forms. The reason is that convergence in distribution is a property of the CDF of the random variable,

not the variable itself. Thus, $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$ if $\lim_{n \rightarrow \infty} |F_{\mathbf{x}_n} - F_{\mathbf{x}}| = 0$ and likewise for a random matrix.

R One important case in which the limiting cdf F is discontinuous is when X is generate, meaning that it is identically equal to a constant c , so that $\Pr(X = c) = 1$.

R In most applications, X is either a normal or chi-square distributed random variable.

As an example, it is well know that

$$t_{n-1} \xrightarrow{d} N(0, 1)$$

as $n \rightarrow \infty$.

Theorem 3.11 — Convergence in probability implies convergence in distribution. If the sequence of random variables $\{X_n\}$ convergences in probability to a random variable X , the sequence also converges in distribution to X . In other words:

$$X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X$$

Convergence in distribution is a weaker form of convergence than convergence in probability, in the sense that $\xrightarrow{p} \implies \xrightarrow{d}$. Intuitively, when X_n converges to X in probability as $n \rightarrow \infty$, the random variable X_n will be arbitrarily close to random variable X for n sufficiently large. Therefore, the probability law of X_n will be arbitrarily close to the probability law of X for n sufficiently large. That is, X_n will converge in distribution to X as $n \rightarrow \infty$.

However, \xrightarrow{d} does not imply \xrightarrow{p} . When $\mathbf{x} = \boldsymbol{\theta}$ is a vector of constants the converse does hold. That is, it is also true that

$$\mathbf{x}_n \xrightarrow{d} \boldsymbol{\theta} \implies \mathbf{x}_n \xrightarrow{p} \boldsymbol{\theta}$$

In this case the limiting distribution of \mathbf{x}_n is degenerate since it collapses to the single point $\boldsymbol{\theta}$.

■ **Example 3.16 — Defining Limiting Distribution Through Convergence in Probability.** Let $\{Y_n\}$ be defined by $Y_n = (2 + n^{-1})X + 3$, where $X \sim N(1, 2)$. Using properties of plim operator it follows that

$$\text{plim}(Y_n) = \text{plim}[(2 + n^{-1})X] + \text{plim}(3) = 2X + 2 \sim N(5, 8).$$

Then, Theorem 3.11 implies that $Y_n \xrightarrow{d} N(5, 8)$. ■

Another important result is that the moments of the asymptotic distribution of a random variable are not necessarily equal to the limits of the moments of the random variable's finite sample distribution. That is, in terms of the first two moments, $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$ does not necessarily imply that $\lim \mathbb{E}(\mathbf{x}_n) = \mathbb{E}(\mathbf{x})$ and $\lim \mathbb{E}(\mathbf{x}_n \mathbf{x}_n') = \mathbb{E}(\mathbf{x} \mathbf{x}')$. For example, in simultaneous equation estimation, we frequently encounter estimator that do not possess finite moments of any order, but that, nevertheless, possess asymptotic distributions with well-defined moments.

■ **Example 3.17** Consider a random sample (y_1, y_2, \dots, y_n) from a normal distribution with mean $\mu \neq 0$ and variance σ^2 . As an estimator for μ^{-1} , the inverse of the sample mean \bar{y}_n^{-1} is a natural choice. To establish its statistical properties we note that, from Khinchine's theorem, $\text{plim } \bar{y}_n = \mu$, and then, from the continuous mapping theorem, $\text{plim } \bar{y}_n^{-1} = \mu^{-1}$. Also because $\sqrt{n}(\bar{y}_n - \mu) \sim N(0, \sigma^2)$ for all n , it follows that

$$\sqrt{n}(\bar{y}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$$

Then,

$$\sqrt{n}(\bar{y}_n^{-1} - \mu^{-1}) \xrightarrow{d} N(0, \sigma^2 \mu^{-4})$$

Thus the mean of the asymptotic distribution of \bar{y}_n^{-1} is μ^{-1} , but $\lim \mathbb{E}(\bar{y}_n^{-1}) \neq \mu^{-1}$ because it can be shown that $\mathbb{E}(\bar{y}_n^{-1})$ does not exist. Note that this example also demonstrate that an estimator can be consistent, that is $\text{plim } \bar{y}_n^{-1} = \mu^{-1}$, without its bias and variance going to zero as $n \rightarrow \infty$ ($\mathbb{E}(\bar{y}_n^{-1})$ and $\mathbb{V}(\bar{y}_n^{-1})$ do not exist.) ■

Some useful results that combine both probability and limiting distribution are as follows.

Theorem 3.12 — Rules for limiting distribution. Consider the following rules

(a) If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then

$$X_n Y_n \xrightarrow{d} cX \quad (3.12)$$

which means that the limiting distribution of $X_n Y_n$ is cX . Also,

$$X_n + Y_n \xrightarrow{d} X + c \quad (3.13)$$

$$X_n / Y_n \xrightarrow{d} X/c, \quad \text{if } c \neq 0 \quad (3.14)$$

(b) If $X_n \xrightarrow{d} X$ and $g(X_n)$ is a continuous function, then

$$g(X_n) \xrightarrow{d} g(X) \quad (3.15)$$

(c) $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$, $\mathbf{A}_n \xrightarrow{p} \mathbf{A} \implies \mathbf{A}_n \mathbf{x}_n \xrightarrow{d} \mathbf{A} \mathbf{x}$, provided that \mathbf{A}_n and \mathbf{x}_n are conformable. In particular, if $\mathbf{x} \sim N(\mathbf{0}, \Sigma)$, then $\mathbf{A}_n \mathbf{x}_n \xrightarrow{d} N(\mathbf{0}, \mathbf{A} \Sigma \mathbf{A}')$.

(d) $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$, $\mathbf{A}_n \xrightarrow{p} \mathbf{A} \implies \mathbf{x}_n' \mathbf{A}_n^{-1} \mathbf{x}_n \xrightarrow{d} \mathbf{x}' \mathbf{A}^{-1} \mathbf{x}$, provided that \mathbf{A}_n and \mathbf{x}_n are conformable and \mathbf{A} is nonsingular.

■ **Example 3.18 — Plims of Matrix Functions to Vector Random Variables.** Let $\{\mathbf{X}_n\}$ and $\{\mathbf{Y}_n\}$ be such that $\text{plim } (\mathbf{X}_n) = \begin{pmatrix} 3 & 2 \\ 2 & 4 \end{pmatrix}$ and $\mathbf{Y}_n \xrightarrow{d} \mathbf{Y} \sim N(\mathbf{0}, \mathbf{I})$. Then,

$$\mathbf{X}_n \mathbf{Y}_n \xrightarrow{d} [\text{plim } (\mathbf{X}_n)] \mathbf{Y} \sim N\left(\mathbf{0}, \begin{pmatrix} 13 & 14 \\ 14 & 20 \end{pmatrix}\right)$$

and

$$\mathbf{X}_n^{-1} \mathbf{Y}_n \xrightarrow{d} [\text{plim}(\mathbf{X}_n)]^{-1} \mathbf{Y} \sim N\left(\mathbf{0}, \begin{pmatrix} .3125 & -.2188 \\ -.2188 & .2031 \end{pmatrix}\right)$$

■

R An useful example of Equation (3.15) of Theorem 3.12 is the following. The exact distribution of t_n^2 is $F(1, n)$. But as $n \rightarrow \infty$, t_n converges to a standard normal variable. According to this result, the limiting distribution of t_n^2 will be that of the square of a standard normal, which is $\chi^2(1)$. Therefore, we conclude that:

$$F(1, n) \xrightarrow{d} \chi^2(1)$$

Lemma 3.13 — Asymptotic Equivalence. If $Y_n - X_n \xrightarrow{p} 0$ and $X_n \xrightarrow{d} X$ as $n \rightarrow \infty$, then $Y_n \xrightarrow{d} X$

Intuitively, if two random variables Y_n and X_n are very close with probability approaching one as $n \rightarrow \infty$, they will follow the same large sample probability distribution. This lemma is very useful when one is interested in deriving the asymptotic distribution of Y_n . We can establish the asymptotic equivalence (in probability) between Y_n and X_n in the sense that $Y_n - X_n \xrightarrow{p} 0$ as $n \rightarrow \infty$, then the asymptotic distributions of Y_n and X_n will be identical.¹

Theorem 3.14 — Cramer-Wold device. If $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$, then $\mathbf{c}'\mathbf{x}_n \xrightarrow{d} \mathbf{c}'\mathbf{x}$ for all conformable vectors \mathbf{c} with real valued elements.

3.5 Central Limit Theorems

Recall that we are interested in a way to describe the statistical properties of estimators when their exact distribution are unknown. However the previous tools do not allow us to find the limiting distribution. From Theorem 3.11 (Convergence in probability implies convergence in distribution), we know that:

$$\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta} \implies \hat{\boldsymbol{\theta}}_n \xrightarrow{d} \boldsymbol{\theta}.$$

That is, the limiting distribution of $\hat{\boldsymbol{\theta}}_n$ is a spike (the asymptotic distribution of $\hat{\theta}_j$ collapses to a single point) and not very informative. The ‘trick’ is to apply some normalization. For example, whereas $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}$, we often find that

$$z_n = \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} f(z), \quad (3.16)$$

where $f(z)$ is a well-defined distribution with mean and positive variance. An estimator which has this property is said to be **root-n consistent**.

For example, consider the sequence of sample means $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$, such that $\bar{X}_1, \bar{X}_2, \bar{X}_3, \dots$. We would like that the probability that \bar{X}_n will deviate from μ by any amount

¹For example, this lemma is useful when deriving the distribution of spatial GLS is the same as the spatial FGLS.

that decreases to zero as $n \rightarrow \infty$. We now that $\mathbb{E}(\bar{X}_n) = \mu$ and $\mathbb{V}(\bar{X}_n) = \sigma^2/n$. Thus, \bar{X}_n will eventually converge to a constant μ since its variance will go to zero eventually for a large enough n . In other words, because $\mathbb{V}(\bar{X}_n) \rightarrow 0$ the distribution shrinks as $n \rightarrow \infty$.

Now consider the sequence of variables:

$$Z_n = \sqrt{n}(\bar{X}_n - \mu), \quad n = 1, 2, \dots \quad (3.17)$$

For this variable we have $\mathbb{E}(Z_n) = 0$ and $\mathbb{V}(Z_n) = \sigma^2$ so that, as $n \rightarrow \infty$, the mean of Z_n remains at zero, but its variance does not converges to zero.

Central limit theorems, establish that, under some conditions, the arithmetic mean of a sufficiently large number of independent random variables, each with a finite expected value and finite variance, will be approximately normally distributed, regardless of the underlying distribution.

The following Theorem gives the most classical (Central Limit Theorem) CLT.

Theorem 3.15 — Lindberg-Levy CLT (Univariate). Let $\{X_n\}$ be a sequence of i.i.d. random variables such that $\mathbb{E}(X_n) = \mu$ and the variance is strictly positive and finite, $0 < \sigma^2 < \infty$. Define $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. Then the distribution of

$$\begin{aligned} Z_n &= \frac{\bar{X}_n - \mathbb{E}(\bar{X}_n)}{\sqrt{\mathbb{V}(\bar{X}_n)}} \\ &= \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \\ &= \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0, 1) \end{aligned}$$

as n approaches infinity. This is the same as:

$$\sqrt{n}(\bar{X}_n - \mu) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (X_i - \mu) \xrightarrow{d} N(0, \sigma^2)$$

This theorem tell us that if a large random sample is taken from any population distribution with finite variance, regardless of whether this population distribution is discrete or continuous, then the distribution of the standardized sample mean

$$Z_n = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$$

will approximately follow a $N(0, 1)$. Therefore, for each finite n , the distribution of \bar{X}_n will be approximately a $N(\mu, \sigma^2/n)$.

It is important to stress that CLT does not say that a large population is approximately normally distributed. It says nothing about the distribution of the population; it is only a statement about the approximate distribution of a standardized sample mean Z_n .

R Sometimes CLT is interpreted incorrectly as implying that the distribution of \bar{X}_n approaches a normal distribution as $n \rightarrow \infty$. This is incorrect because $\mathbb{V}(\bar{X}_n) \rightarrow 0$ and \bar{X}_n converges to a degenerate distribution $F(\cdot)$ such that $F(x) = 0$ if $x < \mu$ and $F(x) = 1$ if $x \geq \mu$.

Multivariate versions of the CLTs can be obtained where each individual \mathbf{x}_i is a random vector in \mathbb{R}^K ,

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iK} \end{pmatrix}$$

with mean vector:

$$\boldsymbol{\mu} = \mathbb{E}(\mathbf{x}_i) = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_K \end{pmatrix},$$

and covariance matrix \mathbf{Q} . Then the sum of the random vectors will be componentwise, that is:

$$\begin{pmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1K} \end{pmatrix} + \begin{pmatrix} x_{21} \\ x_{22} \\ \vdots \\ x_{2K} \end{pmatrix} + \dots + \begin{pmatrix} x_{n1} \\ x_{n2} \\ \vdots \\ x_{nK} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n x_{i1} \\ \sum_{i=1}^n x_{i2} \\ \vdots \\ \sum_{i=1}^n x_{iK} \end{pmatrix} = \sum_{i=1}^n \mathbf{x}_i.$$

The multivariate version of Theorem 3.15 is the following:

Theorem 3.16 — Multivariate Lindberg-Levy CLT. Let $\{\mathbf{x}_n\}$ be a sequence of i.i.d. random variables from a multivariate distribution. If $\mathbb{E}(\mathbf{x}_n) = \boldsymbol{\mu}$ and finite and positive covariance matrix \mathbf{Q} . Then the distribution of

$$Z_n = \sqrt{n}(\bar{\mathbf{x}}_n - \boldsymbol{\mu}) \xrightarrow{d} N(\mathbf{0}, \mathbf{Q}),$$

as n approaches infinity, where $\bar{\mathbf{x}}_n = (1/n) \sum_{i=1}^n \mathbf{x}_i$.

The Linderbeg-Levy CLT is one of the several forms of this extremely powerful result. An important extension allow us to relax the assumption of equal variances. The Linderberg-Feller CLT allows for this extension:

Theorem 3.17 — Univariate Lindberg-Feller CLT. Let $\{X_n\}, i = 1, 2, \dots, n$ be a sequence of i.i.d. random variables. If $\mathbb{E}(X_i) = \mu_i$ and the variance is strictly positive and finite, $0 < \sigma_i^2 < \infty$. Define

$$\bar{\mu}_n = \frac{1}{n}(\mu_1 + \mu_2 + \dots + \mu_n), \quad \text{and} \quad \bar{\sigma}_n^2 = \frac{1}{n}(\sigma_1^2 + \sigma_2^2 + \dots + \sigma_n^2)$$

If no single term dominates this average variance, which we could state as

$$\lim_{n \rightarrow \infty} \frac{\max(\sigma_i)}{n\bar{\sigma}_n} = 0,$$

and if the average variance converges to a finite constant,

$$\lim_{n \rightarrow \infty} \bar{\sigma}_n^2 = \bar{\sigma}^2$$

then

$$Z_n = \frac{\sqrt{n}(\bar{X}_n - \bar{\mu})}{\bar{\sigma}} \xrightarrow{d} N(0, 1)$$

as n approaches infinity.

In practical terms, the theorem states that sums of random variables, regardless of their form, will tend to be normally distributed.

Theorem 3.18 — Multivariate Lindberg-Feller CLT. Suppose that $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ are a sample of random vectors such that $\mathbb{E}(\mathbf{x}_i) = \boldsymbol{\mu}_i$, $\mathbb{V}(\mathbf{x}_i) = \mathbf{Q}_i$, and all mixed third moments of the multivariate distribution are finite. Let

$$\bar{\boldsymbol{\mu}}_n = \frac{1}{n} \sum_{i=1}^n \boldsymbol{\mu}_i,$$

$$\bar{\mathbf{Q}}_n = \frac{1}{n} \sum_{i=1}^n \mathbf{Q}_i.$$

We assume that

$$\lim_{n \rightarrow \infty} \bar{\mathbf{Q}}_n = \mathbf{Q},$$

where \mathbf{Q} is a finite, positive definite matrix, and that for every i ,

$$\lim_{n \rightarrow \infty} (n\bar{\mathbf{Q}}_n)^{-1} \mathbf{Q}_i = \lim_{n \rightarrow \infty} \left(\sum_{i=1}^n \mathbf{Q}_i \right)^{-1} \mathbf{Q}_i = \mathbf{0}$$

We allow the means of the random vectors to differ, although in the cases that will analyze, they will generally be identical. The second assumption states that individual components of the sum must be finite and diminish in significance. There is also an implicit assumption that the sum of matrices is nonsingular. Because the limiting matrix is nonsingular, the assumption must hold for large enough n , which is all that concerns us here. With these in place, the result is

$$\sqrt{n}(\bar{\mathbf{x}}_n - \bar{\boldsymbol{\mu}}_n) \xrightarrow{d} N(\mathbf{0}, \mathbf{Q})$$

Theorem 3.19 — Liapounov Central Limit Theorem. Suppose that $\{X_n\}, i = 1, 2, \dots, n$ is a sequence of independent random variables with finite mean μ_i and finite positive variances σ_i^2 such that $\mathbb{E}[|X_i - \mu_i|^{2+\delta}]$ is finite for some $\delta > 0$. If $\bar{\sigma}_n$ is positive and finite for all n sufficiently large, then

$$\frac{\sqrt{n}(\bar{X}_n - \bar{\mu}_n)}{\bar{\sigma}_n} \xrightarrow{d} N(0, 1)$$

This version of the central limit theorem requires only that moments slightly larger than two be finite and it is generally used when the variables are fixed.

We end this section by defining the concept of **asymptotic variance**.

Definition 3.5.1 — Asymptotic Variance. Let $\{\mathbf{x}_n\}$ be a sequence of random vectors. If there exists a sequence of matrices $\{\mathbf{V}_n\}$ such that \mathbf{V}_n is nonsingular for all n sufficiently large and $\mathbf{V}_n^{-1/2}\mathbf{x}_n \stackrel{a}{\sim} N(\mathbf{0}, \mathbf{I})$, then \mathbf{V}_n is called the asymptotic covariance matrix of \mathbf{x}_n , denoted $\text{Avar}(\mathbf{x}_n)$.

3.6 Orders in Probability

Similarly to the nonstochastic sequences, we can make similar statement about o and O when we have random variables. The following theorem state the definition of unboudedness and convergence for random variables:

Definition 3.6.1 — Order in Probability. Consider the following two definition:

- (a) **Stochastically Bounded (Big O):** The sequence of random variables $\{X_n\}$ is at most of order in probability n^λ , and we write

$$X_n = O_p(n^\lambda) \quad (3.18)$$

if, for every $\epsilon > 0$, there exists a real number n_0 such that:

$$\Pr \left[n^{-\lambda} |X_n| \geq n_0 \right] \leq \epsilon \quad (3.19)$$

for all n .

- (b) **Stochastic Convergence:** Also, we say that $\{X_n\}$ is of smaller order in probability than n^λ and we write

$$X_n = o_p(n^\lambda) \quad (3.20)$$

if

$$\text{plim } n^{-\lambda} X_n = 0 \quad (3.21)$$

When $\lambda = 0$, X_n converges to zero, and we also write $X_n = o_p(1)$.

Intuitively, for $X_n = O_p(n^\lambda)$ with $\lambda > 0$, the order n^λ is the fastest growth rate at which X_n goes to infinity with probability approaching 1. When $\lambda < 0$, the order n^λ is the fastest convergence rate at which X_n vanishes to 0 with probability approaching 1. Thus, $X_n = O_p(1) = O_p(n^0)$ implies that for n sufficiently large, $|X_n|$ takes value larger than a very large constant has a tiny probability. In other words, $|X_n|$ is bounded by a constant with a very high probability for all n sufficiently large.

Definition 3.6.2 — Stochastically Negligible. If $X_n \xrightarrow{p} 0$, then $X_n = o_p(1)$. If $X_n = n^\lambda o_p(1)$, then $X_n = o_p(n^{-\lambda})$

To give some intuition about these definitions, consider $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$ and $\mathbf{y}_n \xrightarrow{p} \mathbf{0}$. Then:

$$\mathbf{x}_n + \mathbf{y}_n \xrightarrow{d} \mathbf{x} \quad \text{by 3.12 in Theorem 3.12}$$

That is, if $\mathbf{z}_n = \mathbf{x}_n + \mathbf{y}_n$ and $\mathbf{y}_n \xrightarrow{p} \mathbf{0}$, implying that $\mathbf{z}_n - \mathbf{x}_n \xrightarrow{p} \mathbf{0}$, then the asymptotic distribution of \mathbf{z}_n is the same as that of \mathbf{x}_n . Note that this is the same as Lemma 3.13 (asymptotic equivalence). So we can write:

$$\mathbf{z}_n \stackrel{a}{\sim} \mathbf{x}_n \quad \text{or} \quad \mathbf{z}_n = \mathbf{x}_n + o_p(1)$$

where $o_p(1)$ is some variable (\mathbf{y}_n in this case) that is stochastically negligible, that is, it converges to zero in probability.

This is more intuitive if we think in the consistency of OLS estimator. Given the OLS consistency, it is the same to write:

$$\hat{\beta}_n \xrightarrow{p} \beta_0 \quad \text{as } n \rightarrow \infty$$

as

$$\hat{\beta}_n = \beta_0 + o_p(1) \quad \text{as } n \rightarrow \infty$$

In other words, a consistent estimator is equal to the true estimator plus something that converges to 0 in probability.

Lemma 3.20 — Convergence in distribution implies boundedness. Let X_n be a random variable with CDF $F_n(\cdot)$, and let X be a random variable with continuous CDF $F(\cdot)$. If $X_n \xrightarrow{d} X$ as $n \rightarrow \infty$, then $X_n = O_p(1)$

Intuitively, if the probability distribution of X_n converges to a well-defined continuous probability distribution as $n \rightarrow \infty$, then X_n is bounded in probability. This result is very useful for establishing that a sequence of random variables is bounded in probability. Often it is easier to verify that a sequence of random variables converges in distribution.

When do we use the O_p ? If a random vector converges in distribution $\mathbf{x}_n \xrightarrow{d} \mathbf{x}$ (for example $\mathbf{x} \sim N(\mathbf{0}, \mathbf{V})$) then $\mathbf{x}_n = O_p(1)$.

■ **Example 3.19** If $X_n \sim N(0, 1)$ for all $n \geq 1$. Then $X_n = O_p(1)$ because for any given $\delta > 0$, there exists a finite constant $M = \Phi^{-1}(1 - \delta/2) < \infty$, where Φ is the $N(0, 1)$ CDF, such that

$$\Pr(|X_n| > M) = 2[1 - \Phi(M)] = \delta < 2\delta$$

for all $n \geq 1$ ■

$O_p(1)$ is weaker than $o_p(1)$ in the sense that $X_n = o_p(1)$ implies $X_n = O_p(1)$ but not the reverse.

There are many simple rules for manipulating $o_p(1)$ and $O_p(1)$ sequences which can be deduced from the continuous mapping theorem or Slutsky's Theorem.

Proposition 3.21 — Properties of stochastic big and little O. Let a_n and b_n random scalars.

- (a) If $a_n = O_p(n^\lambda)$ and $b_n = O_p(n^\mu)$, then $a_n b_n = O_p(n^{\lambda+\mu})$ and $a_n + b_n = O_p(n^\kappa)$, where $\kappa = \max[\lambda, \mu]$.
- (b) If $a_n = o_p(n^\lambda)$ and $b_n = o_p(n^\mu)$, then $a_n b_n = o_p(n^{\lambda+\mu})$ and $a_n + b_n = o_p(n^\kappa)$, where $\kappa = \max[\lambda, \mu]$.

- (c) If $a_n = O_p(n^\lambda)$ and $b_n = o_p(n^\mu)$, then $a_n b_n = o_p(n^{\lambda+\mu})$ and $a_n + b_n = O_p(n^\kappa)$, where $\kappa = \max[\lambda, \mu]$.

One of the most common uses of this concept of stochastic order is “root- n ” (\sqrt{n}) consistency:

Definition 3.6.3 — \sqrt{N} -Consistent. if $\sqrt{n}(\boldsymbol{\theta}_n - \boldsymbol{\theta}_0) = O_p(1)$, then $\boldsymbol{\theta}_n$ is \sqrt{n} consistent for $\boldsymbol{\theta}_0$

■ **Example 3.20 — OLS and O_p and o_p .** Recall that:

$$\hat{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_0 + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}.$$

Under appropriate assumption, we know that:

$$\frac{1}{n}\mathbf{X}'\mathbf{X} \xrightarrow{p} \mathbb{E}(\mathbf{X}'\mathbf{X}),$$

which is finite and positive definite. The fact that the elements of $n^{-1}\mathbf{X}'\mathbf{X}$ converge to finite limits in probability implies that $N^{-1}\mathbf{X}'\mathbf{X}$ is **bounded** in the sense that the sequences of the elements within $n^{-1}\mathbf{X}'\mathbf{X}$ are bounded, and under these circumstances we say that $\mathbf{X}'\mathbf{X}$ is at most of order n , that is, $\mathbf{X}'\mathbf{X} = O_p(n)$, or we can say:

$$n^{-1}\mathbf{X}'\mathbf{X} = O_p(1).$$

We also assume that $n^{-1/2}\mathbf{X}'\boldsymbol{\varepsilon}$ has probability limit which a normally distributed random variable with expectation zero and finite variance. So, we can write:

$$\mathbf{X}'\boldsymbol{\varepsilon} = O_p(n^{1/2}).$$

Thus:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \\ \hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 &= (n^{-1}\mathbf{X}'\mathbf{X})^{-1}n^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \\ \hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 &= O_p(1) \cdot o_p(1) \\ \hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 &= o_p(1)\end{aligned}$$

Also:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \\ \hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 &= [O_p(n)]^{-1} O_p(n^{1/2}) \\ \hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 &= O_p(n^{-1})O_p(n^{1/2}) \quad \because [O_p(n)]^{-1} = O_p(n^{-1}) \\ \hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 &= O_p(n^{-1/2})\end{aligned}$$

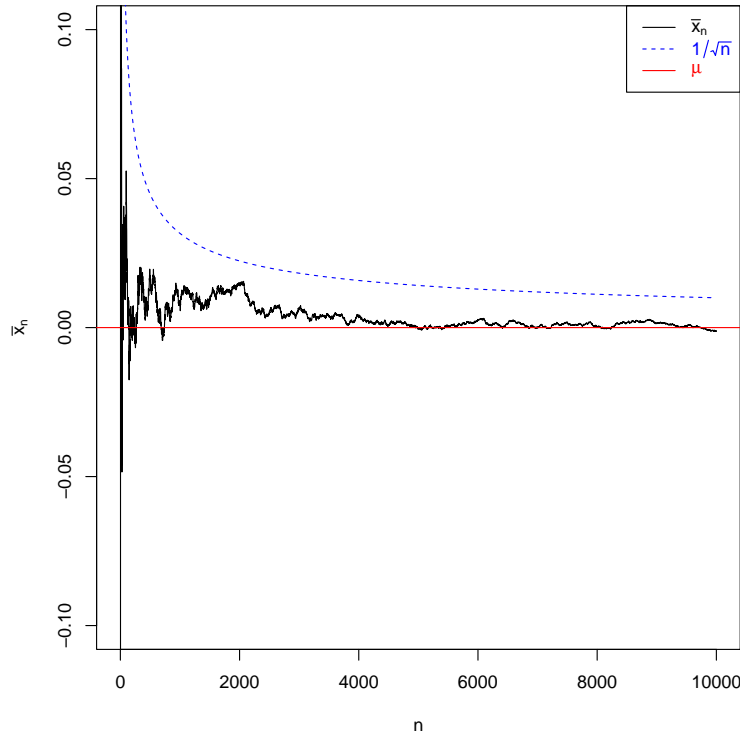
So, we might then say that $\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0$ is converging to zero at the rate $1/\sqrt{n}$; and the rate tell us what multiplier of the variable $\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0$ stabilizes it so that it converges to a well-defined random variables rather than to 0 or ∞ .

Note also that:

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_n - \beta_0) &= n^{1/2}O_p(n^{-1})O_p(n^{1/2}) \\
&= n^{1/2}O_p(n^{-1/2}) \\
&= O_p(1) \quad \because X_n = O_p(1/\sqrt{n}) \implies X_n/(1/\sqrt{n}) = \sqrt{n}X_n = O_p(1)
\end{aligned}$$

■ **Example 3.21 — Rate of convergence for sample mean.** Consider an iid random sample X_i with mean $\mu = 0$ and variance $\sigma^2 = 0,25$. Then, by the CLT the sample mean \bar{X} we know that $\sqrt{n}(\bar{X} - \mu)/\sigma \xrightarrow{d} N(0,1)$. That is, $\sqrt{n}(\bar{X} - \mu)/\sigma = O_p(1)$. This implies that $\bar{X} - \mu = O_p(1/\sqrt{n})$. Figure 3.7 shows how \bar{X} converges towards μ to the speed of $1/\sqrt{n}$. ■

Figure 3.7: Convergence of the sample mean and speed of convergence



Notes: This graph shows the convergence of \bar{X} to μ as $n \rightarrow \infty$ for a normal distribution as fast as $1/\sqrt{n}$.

3.7 Triangular Arrays

An important question in the context of asymptotic theory is the following: *What is the meaning of $n \rightarrow \infty$ in a spatial context?* Will it increase the geographical area or will it increase the number of spatial units in a given geographical area?

For spatial data, two distinct asymptotic frameworks have been studied: **increasing domain** and **infill asymptotic**. Increasing domain consists of a sampling structure where new observations (spatial units) are added at the edges (boundary points), similar to the underlying asymptotic in time series analysis. That is, increasing domain asymptotic refers to more and more observations being sampled over an increasing domain. The problem here

is what the boundary is. When referring to increasing domain asymptotic, it is assumed that the spatial locations of the observations do not become dense. Infill asymptotic are appropriate when the spatial domain is bounded, and new observations (points) are added in between existing ones, generating denser surface. In most applications of spatial econometric, the implied structure is that of an increasing domain.

The increasing domain framework requires the knowledge of **triangular arrays**. The following definition give us a simply definition of Triangular Arrays.

Definition 3.7.1 — Triangular Array of Random Variables. The ordered collection of random variables

$$\{X_{11}, X_{21}, X_{22}, X_{31}, X_{32}, X_{33}, \dots, X_{nn}, \dots\},$$

or

$$\begin{pmatrix} X_{11} \\ X_{21} & X_{22} \\ X_{31} & X_{32} & X_{33} \\ \vdots & \vdots & \vdots & \ddots \\ X_{n1} & X_{n2} & X_{n3} & X_{n4} & \dots & X_{nn} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

is called a triangular array of random variables, and will be denoted by $\{X_{nn}\}$.

Central limit theorems that are applied to triangular arrays of random variables are concerned with limiting distributions of appropriately defined function of the row average $S_n = n^{-1} \sum_{i=1}^n X_{ni}$. For example, for $n = 3$ (third row) we have $S_3 = (1/3)(X_{31} + X_{32} + X_{33})$. Note that the traditional CLTs deal with functions of average of the type $n^{-1} \sum_{i=1}^n X_i$, the X_i 's being elements of the sequence $\{X_n\}$. However, the triangular array $\{X_{nn}\}$ is more general than a sequence $\{X_n\}$ in the sense that the random variables in a row of the array need not be the same as random variables in other rows. Thus, the triangular nature of a random variable leads to certain statistical problems, especially with respect to the relevant CLT that should be applied. In other words, we will need a CLT applicable to triangular array. Both the LLN and CLT require slightly stronger conditions than the LLN and CLT for i.i.d sequence of random variables.

What are the conditions on the random variables so that a properly S_n converges to a normal distribution as $n \rightarrow \infty$. In a nutshell, assume:

- Independence: assume all random variables in the array are independent.
- Centering: assume $\mathbb{E}(X_{j,i}) = 0$ for all j, i .
- Variances converge: assume $\sum_{i=1}^n \mathbb{E}(X_{n,i}^2) \rightarrow \sigma^2 > 0$ as $n \rightarrow \infty$
- No single variance is too large.

Then $S_n \xrightarrow{d} N(0, \sigma^2)$ as $n \rightarrow \infty$.

Probably you're asking yourself, why triangular arrays are important in the spatial context? Note that if we adopt the increasing domain approach, it is clear that as n increases, \mathbf{W} itself changes as observations are added. To see this, let $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^\top, \rho_0, \sigma_0^2)^\top$ be the true

parameter vector. We further assume that variables and estimates depend on the sample size n . This will allow us to study their behavior as $n \rightarrow \infty$. Therefore, denote $\mathbf{A}_n(\rho) = \mathbf{I}_n - \rho \mathbf{W}_n$ for any value of ρ . The “equilibrium” vector is

$$\mathbf{y}_n = \mathbf{A}_n^{-1}(\mathbf{X}_n \boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}_n) \quad (3.22)$$

where $\mathbf{A}_n = \mathbf{A}_n(\rho_0)$ is nonsingular. Let $\boldsymbol{\varepsilon}_n(\boldsymbol{\delta}) = \mathbf{y}_n - \mathbf{X}_n \boldsymbol{\beta} - \rho \mathbf{W}_n \mathbf{y}_n$, where $\boldsymbol{\delta} = (\boldsymbol{\beta}^\top, \rho)^\top$. Thus, $\boldsymbol{\varepsilon}_n = \boldsymbol{\varepsilon}_n(\boldsymbol{\delta}_0)$. Since the matrices $(\mathbf{I}_n - \rho \mathbf{W})^{-1}$ generally depend upon the sample size n , the vectors \mathbf{y} and $\boldsymbol{\varepsilon}$ will also depend upon n , and they will form a **triangular arrays**. This is due to the fact that for the “boundary” elements the sample weights matrix changes as new spatial units — or new data points — are added. That is, new spatial units change the structure for the existing spatial units (see for example [Kelejian and Prucha, 1999, 2001](#); [Anselin, 2007](#)). For example, the outcome for the first spatial unit, $y_{1,n}$, will be different if we consider a total $n = 10$ or $n = 15$ observations because of the changing nature of \mathbf{W} as n changes and given the DGP in Equation (3.22). This implies that these elements and the vector \mathbf{y} should be indexed by n :

$$\mathbf{y}_n = (y_{11}, y_{21}, y_{22}, \dots, y_{nn})$$

For example, for $n = 1, 2, 3$, then (by row):

$$\begin{aligned} n = 1 &\implies y_{11} \\ n = 2 &\implies y_{12} \ y_{22} \\ n = 3 &\implies y_{13} \ y_{23} \ y_{33} \\ &\vdots \\ n = n &\implies y_{13} \ y_{23} \ y_{33} \dots y_{3n} \end{aligned}$$

where $y_{11} \neq y_{12} \neq y_{13}$ and $y_{22} \neq y_{23}$. Note that the dependent variable in the same row are mutually independent (spatial units are independent) and have the same distribution. But the distribution of the random variable y (and ϵ) in different rows are allowed to be different.

The triangular array structure of y is partly a consequence of allowing a triangular array structure for the disturbances in the model. But there is a more fundamental reason for it, and for treating the \mathbf{X} observations as a triangular array also. In allowing for the elements of \mathbf{X}_n to depend on n we allow explicitly for some of the regressors to be spatial lags.

We can identify each of the indices $i = 1, \dots, n$ with a location in space. In regularly-observed time series settings, these indices correspond to equidistant points on the real line, and it is evident what we usually mean by letting n increase. However there is ambiguity when these points are in space. For example, consider n points on a 2 dimensional regularly-spaced lattice, where both the number (n_1) of rows and the number (n_2) of columns increases with $n = n_1 \cdot n_2$. If we choose to list these points in lexicographic order (say first left to right, then second row, etc) then as n increases there would have to be some re-labeling, as the triangular array permits. Another consequence of this listing is that dependence between locations i and j is not always naturally expressed as a function of the difference $i - j$. For example, this is so if the dependence is isotropic.

3.8 Matrix

Let $\mathbf{C} = \mathbf{AB}$, then the (i, j) element of \mathbf{C} is

$$c_{ij} = \sum_{k=1}^n a_{ik} b_{kj} \quad (3.23)$$

3.9 Matrix Norm

Definition 3.9.1 — Matrix Norm. Given a square complex or real matrix \mathbf{A} , a matrix norm $\|\mathbf{A}\|$ is a nonnegative number associated with \mathbf{A} having the properties

- (a) $\|\mathbf{A}\| > 0$ when $\mathbf{A} \neq 0$ and $\|\mathbf{A}\| = 0$ if and only if $\mathbf{A} = 0$,
- (b) $\|k\mathbf{A}\| = |k| \|\mathbf{A}\|$ for any scalar k ,
- (c) $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$,
- (d) $\|\mathbf{AB}\| \leq \|\mathbf{A}\| \|\mathbf{B}\|$

The maximum absolute column sum norm $\|\mathbf{A}\|_1$ is defined as

$$\|\mathbf{A}\|_1 = \max_j \sum_{i=1}^n |a_{ij}| \quad (3.24)$$

The spectral norm $\|\mathbf{A}\|_2$ is defined as

$$\|\mathbf{A}\|_2 = \sigma_{\max}(\mathbf{A}) = \sqrt{\lambda_{\max}(\mathbf{A}^\top \mathbf{A})} \quad (3.25)$$

The maximum absolute row sum norm is defined by

$$\|\mathbf{A}\|_\infty = \max_i \sum_{j=1}^n |a_{ij}| \quad (3.26)$$

$\|\mathbf{A}\|_1$, $\|\mathbf{A}\|_2$ and $\|\mathbf{A}\|_\infty$ satisfy the inequality $\|\mathbf{A}\|_2^2 \leq \|\mathbf{A}\|_1 \|\mathbf{A}\|_\infty$.

3.10 Bounded Matrices and Useful Lemmas for Spatial Econometrics

Definition 3.10.1 — Bounded Matrices. Let $\{\mathbf{A}_n\}$ be a sequence of n -dimensional square matrices, where $\mathbf{A}_n = [a_{n,ij}]$,

- (a) The column sums of $\{\mathbf{A}_n\}$ are uniformly bounded (in absolute value) if there exists a finite constant c_a that does not depend on n such that

$$\|\mathbf{A}_n\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^n |a_{n,ij}| \leq c_a$$

- (b) The row sums of $\{\mathbf{A}_n\}$ are uniformly bounded (in absolute value) if there exists a finite constant c_a that does not depend on n such that

$$\|\mathbf{A}_n\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{n,ij}| \leq c_a$$

Then $\{\mathbf{A}_n\}$ is said to be **uniformly bounded** in row sums if $\{\|\mathbf{A}_n\|_\infty\}$ is a bounded sequence. Similarly, $\{\mathbf{A}_n\}$ is said to be **uniformly bounded** in column sums if $\{\|\mathbf{A}_n\|_1\}$ is a bounded sequence.

The following lemmas will be very useful:

Lemma 3.22 If $\{\mathbf{A}_n\}$ and $\{\mathbf{B}_n\}$ are uniformly bounded in row sums (column sums), then $\{\mathbf{A}_n \mathbf{B}_n\}$ is also uniformly bounded in row sums (column sums).

Proof. Suppressing the index n , let $\mathbf{D} = \mathbf{AB}$, then

$$d_{ij} = \sum_{r=1}^n a_{ir} b_{rj} \quad (3.27)$$

Let r_i be the i th row sum, then

$$\begin{aligned} r_i &= \sum_{j=1}^n |d_{ij}| \\ &= \sum_{j=1}^n \left| \sum_{r=1}^n a_{ir} b_{rj} \right| \\ &\leq \sum_{j=1}^n \sum_{r=1}^n |a_{ir} b_{rj}| \text{ by triangle inequality 3.A.2} \\ &= \sum_{j=1}^n \sum_{r=1}^n |a_{ir}| |b_{rj}| \text{ by multiplicativity 3.A.1} \\ &= \sum_{r=1}^n \sum_{j=1}^n |a_{ir}| |b_{rj}| \text{ by property of summation} \\ &= \sum_{r=1}^n |a_{ir}| \sum_{j=1}^n |b_{rj}| \\ &\leq c_a c_b, \text{ for all } i = 1, \dots, n \text{ and } n \geq 1 \text{ by Def. 3.10.1} \end{aligned} \quad (3.28)$$

Similarly, we can show that

$$\sum_{i=1}^n |d_{ij}| \leq c_a c_b, \text{ for all } j = 1, \dots, n \text{ and } n \geq 1. \quad (3.29)$$

■

Lemma 3.23 If $\{\mathbf{A}_n\}$ is absolutely summable (uniformly bounded in either row or column sums), and \mathbf{Z}_n has bounded elements, then the elements of $\mathbf{Z}_n^\top \mathbf{A}_n \mathbf{Z}_n = O(n)$

Proof. Suppressing the index n , let Z_{ij} be the (i, j) th element of \mathbf{Z} , and let $|Z_{ij}| \leq c_z$ for all i, j and $n \geq 1$. Let δ_{ij} be the (i, j) th element of $\mathbf{Z}^\top \mathbf{A} \mathbf{Z}$, then

$$\begin{aligned}
\delta_{ij} &= \sum_{r=1}^n \sum_{s=1}^n Z_{si} a_{sr} Z_{rj} \\
|\delta_{ij}| &= \left| \sum_{r=1}^n \sum_{s=1}^n Z_{si} a_{sr} Z_{rj} \right| \\
|\delta_{ij}| &\leq \sum_{r=1}^n \sum_{s=1}^n |Z_{si}| |a_{sr}| |Z_{rj}| \text{ by 3.A.2} \\
&\leq c_z^2 \sum_{r=1}^n \sum_{s=1}^n |a_{sr}| \\
&\leq c_z^2 \sum_{r=1}^n c_a \\
&\leq c_z^2 c_a n \\
&= O(n)
\end{aligned} \tag{3.30}$$

■

Lemma 3.24 If $\{\mathbf{A}_n\}$ is absolutely summable (uniformly bounded in either row or column sums), then

- (a) elements $a_{n,ij}$ of \mathbf{A}_n are uniformly bounded in i and j ,
- (b) $\text{tr}(\mathbf{A}^m) = O(n)$ for $m \geq 1$, and
- (c) the elements of $\text{tr}(\mathbf{A}_n \mathbf{A}_n^\top) = O(n)$.

Proof. Proof of (c). Suppressing the index n

$$\begin{aligned}
\text{tr}(\mathbf{A}_n \mathbf{A}_n^\top) &= \sum_{i=1}^n \sum_{j=1}^n a_{ij}^2 \\
\left| \text{tr}(\mathbf{A}_n \mathbf{A}_n^\top) \right| &= \left| \sum_{i=1}^n \sum_{j=1}^n a_{ij}^2 \right| \\
&\leq \sum_{i=1}^n \sum_{j=1}^n |a_{ij}^2| \\
&\leq \sum_{i=1}^n \left(\sum_{j=1}^n |a_{ij}| \right)^2 \\
&\leq n c_1^2 \quad \text{if } \mathbf{A}_n \text{ is uniformly bounded in row sums}
\end{aligned} \tag{3.31}$$

$$\begin{aligned}
\left| \text{tr}(\mathbf{A}_n \mathbf{A}_n^\top) \right| &\leq \sum_{j=1}^n \left(\sum_{i=1}^n |a_{ij}| \right)^2 \\
&\leq n c_2^2 \quad \text{if } \mathbf{A}_n \text{ is uniformly bounded in column sums}
\end{aligned} \tag{3.32}$$

■

3.11 Quadratic forms

Definition 3.11.1 — Quadratic form. For a $n \times n$ symmetric matrix $\mathbf{A}_n = [a_{ij}]$ the quadratic function of n variables $\boldsymbol{\varepsilon}$ defined by:

$$\boldsymbol{\varepsilon}^\top \mathbf{A}_n \boldsymbol{\varepsilon} = \sum_{i=1}^n \sum_{j=1}^n a_{ij} \varepsilon_i \varepsilon_j \quad (3.33)$$

is called the quadratic form with matrix \mathbf{A}_n . If \mathbf{A}_n is not symmetric, we can replace \mathbf{A}_n by $(\mathbf{A}_n + \mathbf{A}_n^\top)/2$.

Lemma 3.25 — First and Second Moments. Let $\mathbf{A}_n = [a_{ij}]$ be an n -dimensional square matrix. Then, it can be shown that:

$$\mathbb{E}[\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n] = \text{tr}(\mathbf{A}_n \boldsymbol{\Sigma}) + \boldsymbol{\mu}^\top \mathbf{A}_n \boldsymbol{\mu}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the expected value and variance-covariance matrix of $\boldsymbol{\varepsilon}_n$, respectively. This result only depends on the existence of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$; it does not require normality of $\boldsymbol{\varepsilon}$.

Assume that $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \sigma_0^2 \mathbf{I}$, then

- (a) $\mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) = \sigma_0^2 \text{tr}(\mathbf{A}_n)$,
- (b) $\mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n)^2 = (\mu_4 - 3\sigma_0^4) \sum_{i=1}^n a_{ii}^2 + \sigma_0^4 [\text{tr}^2(\mathbf{A}_n) + \text{tr}(\mathbf{A}_n \mathbf{A}_n^\top) + \text{tr}(\mathbf{A}_n^2)]$, and
- (c) $\mathbb{V}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) = (\mu_4 - 3\sigma_0^4) \sum_{i=1}^n a_{ii}^2 + \sigma_0^4 [\text{tr}(\mathbf{A}_n \mathbf{A}_n^\top) + \text{tr}(\mathbf{A}_n^2)]$.

For the moment assume that \mathbf{A} is symmetric and $\boldsymbol{\varepsilon}$ is normally distributed, then:

$$\mathbb{V}(\boldsymbol{\varepsilon}^\top \mathbf{A} \boldsymbol{\varepsilon}) = 2 \text{tr}(\mathbf{A} \boldsymbol{\Sigma} \mathbf{A} \boldsymbol{\Sigma}) + 4 \boldsymbol{\mu}^\top \mathbf{A} \boldsymbol{\Sigma} \mathbf{A} \boldsymbol{\mu},$$

and the covariance:

$$\text{Cov}(\boldsymbol{\varepsilon}^\top \mathbf{A}_1 \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}^\top \mathbf{A}_2 \boldsymbol{\varepsilon}) = 2 \text{tr}(\mathbf{A}_1 \boldsymbol{\Sigma} \mathbf{A}_2 \boldsymbol{\Sigma}) + 4 \boldsymbol{\mu}^\top \mathbf{A}_1 \boldsymbol{\Sigma} \mathbf{A}_2 \boldsymbol{\mu}$$

If \mathbf{A} is not symmetric, then:

$$\text{Cov}(\boldsymbol{\varepsilon}^\top \mathbf{A}_1 \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}^\top \mathbf{A}_2 \boldsymbol{\varepsilon}) = 2 \text{tr} \left[\frac{1}{2} (\mathbf{A}_1 + \mathbf{A}_1^\top) \boldsymbol{\Sigma} \frac{1}{2} (\mathbf{A}_2 + \mathbf{A}_2^\top) \boldsymbol{\Sigma} \right] + 4 \boldsymbol{\mu}^\top \frac{1}{2} (\mathbf{A}_1 + \mathbf{A}_1^\top) \boldsymbol{\Sigma} \frac{1}{2} (\mathbf{A}_2 + \mathbf{A}_2^\top) \boldsymbol{\mu} \quad (3.34)$$

In particular, if $\boldsymbol{\varepsilon}$'s are normally distributed with mean 0 and variance σ_0^2 , then

- $\mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n)^2 = \sigma_0^4 [\text{tr}^2(\mathbf{A}_n) + \text{tr}(\mathbf{A}_n \mathbf{A}_n^\top) + \text{tr}(\mathbf{A}_n^2)]$, and
- $\mathbb{V}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) = \sigma_0^4 [\text{tr}(\mathbf{A}_n \mathbf{A}_n^\top) + \text{tr}(\mathbf{A}_n^2)]$

Lemma 3.26 Suppose that $\{\mathbf{A}_n\}$ is uniformly bounded in either row and column sums. Then:

- (a) $\mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) = O(n)$,
- (b) $\mathbb{V}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) = O(n)$
- (c) $\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n = O_p(n)$

Proof. Proof of (a). By Lemma 3.25, $\mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) = \sigma_0^2 \text{tr}(\mathbf{A}_n)$. By Lemma 3.24, $\text{tr}(\mathbf{A}) = O(n)$. Then $\mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) = \sigma_0^2 O(n) = O(n)$ since σ_0^2 is a constant.

Proof of (b). From Lemma 3.25

$$\mathbb{V}(n^{-1} \boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) = (\mu_4 - 3\sigma_0^4) \sum_{i=1}^n a_{ii}^2 + \sigma_0^4 [\text{tr}(\mathbf{A}_n \mathbf{A}_n^\top) + \text{tr}(\mathbf{A}_n^2)] \quad (3.35)$$

From Lemma 3.24 $\text{tr}(\mathbf{A}_n \mathbf{A}_n^\top) = O(n)$ and $\text{tr}(\mathbf{A}_n^2) = O(n)$. Since

$$\sum_{i=1}^n a_{n,ii}^2 \leq \text{tr}(\mathbf{A}_n \mathbf{A}_n^\top) = O(n),$$

then

$$\begin{aligned} \mathbb{V}(n^{-1} \boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) &= (\mu_4 - 3\sigma_0^4) O(n) + \sigma_0^4 [O(n) + O(n)] \\ &= O(n) \end{aligned} \quad (3.36)$$

Proof of (c). Since $\mathbb{E}((\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n)^2) = \mathbb{V}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) + (\mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n))^2 = O(n) + O(n)O(n) = O(n^2)$ by Property 3.1. The Chebyshev inequality 3.A.4 implies that

$$\Pr\left(\frac{1}{n} |\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n| \geq \delta\right) \leq \frac{1}{\delta^2 n^2} \mathbb{E}[(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n)^2] = \frac{1}{\delta^2} O(1)$$

Then $\frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n = O_p(1)$. Thus, $n^{-1} \boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n$ is bounded. ■

Theorem 3.27 — Consistency of quadratic forms in spatial models. Let \mathbf{A}_n be an $n \times n$ nonstochastic matrix whose row and columns sums are uniformly bounded in absolute value. Let $\boldsymbol{\varepsilon}_n^\top = (\varepsilon_{n1}, \dots, \varepsilon_{nn})$ where ε_{ni} are iid $(0, \sigma_0^2)$ and $\mathbb{E}(\varepsilon_{ni}^4) < \infty$. Then

$$\frac{\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n}{n} \xrightarrow{p} \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) = \sigma_0^2 \frac{\text{tr}(\mathbf{A}_n)}{n}$$

If the limit of $\text{tr}(\mathbf{A}_n)/n$ exists, then:

$$\lim_{n \rightarrow \infty} \frac{\text{tr}(\mathbf{A}_n)}{n} = \mathbf{A}^*$$

and

$$\frac{\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n}{n} \xrightarrow{p} \sigma_0^2 \mathbf{A}^*.$$

Proof. This result follows from Theorem 3.5. By Chebyshev inequality 3.A.4

$$\begin{aligned}
\Pr \left[\frac{1}{n} \left| \boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n - \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) \right| \geq \delta \right] &\leq \frac{\mathbb{V}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n)}{n^2 \delta^2} \\
\Pr \left[\left| \frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n - \sigma_0^2 \frac{\text{tr}(\mathbf{A}_n)}{n} \right| \geq \delta \right] &\leq \frac{\mathbb{V}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n)}{n^2 \delta^2} \\
&\leq \frac{1}{\delta^2} n^{-2} \mathbb{V}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n)
\end{aligned}$$

By Lemma 3.24, $\text{tr}(\mathbf{A}_n) = O(n)$. Hence $\lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) = \sigma_0^2 \mathbf{A}^* = O(1)$ and the expectation exists.

By Lemma 3.26 $\mathbb{V}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) = O(n)$. Therefore, $n^{-2} \mathbb{V}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) = \mathbb{V}(n^{-1} \boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) = O(n^{-1}) = o(1)$. Thus $\mathbb{V}(n^{-1} \boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) \rightarrow 0$ as $n \rightarrow \infty$ and $\Pr \left[\frac{1}{n} \left| \boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n - \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) \right| \geq \delta \right] \rightarrow 0$, so

$$\frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n - \frac{1}{n} \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) = o_p(1)$$

which can also be written as

$$\frac{\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n}{n} \xrightarrow{p} \sigma_0^2 \mathbf{A}^*.$$

■

3.12 CLT for Spatial Models

The following theorem states the limiting distribution for triangular arrays with homokedastic errors in linear forms:

Theorem 3.28 — CLT for triangular arrays with homokedastic errors, (Kelejian and Prucha, 1998). Let $\{v_{i,n}, 1 \leq i \leq n, n \geq 1\}$ be a triangular array of identically distributed random variables. Assume that the random variables $\{v_{i,n}, 1 \leq i \leq n\}$ are jointly independently distributed for each n with $\mathbb{E}(v_{i,n}) = 0$ and $\mathbb{E}(v_{i,n}^2) = \sigma^2 < \infty$. Let $\{a_{ij,n}, 1 \leq i \leq n, n \geq 1\}, j = 1, \dots, k$ be triangular arrays of real numbers that are bounded in absolute value. Further let

$$\mathbf{v}_n = \begin{pmatrix} v_{1,n} \\ \vdots \\ v_{n,n} \end{pmatrix}, \quad \mathbf{A}_n = \begin{pmatrix} a_{11,n} & \dots & a_{1k,n} \\ \vdots & & \vdots \\ a_{n1,n} & \dots & a_{nk,n} \end{pmatrix}$$

Then:

$$\frac{1}{\sqrt{n}} \mathbf{A}_n^\top \mathbf{v}_n = O_p(1)$$

Furthermore, assume that $\lim_{n \rightarrow \infty} n^{-1} \mathbf{A}_n^\top \mathbf{A}_n = \mathbf{Q}_{AA}$ is finite and nonsingular matrix. Then

$$\frac{1}{\sqrt{n}} \mathbf{A}_n^\top \mathbf{v}_n \xrightarrow{d} N(\mathbf{0}, \sigma^2 \mathbf{Q}_{AA})$$

Theorem 3.29 — CLT for Vectors of Linear Quadratic Forms with Heterokedastic Innovations.

Assume the following:

- (a) For $r = 1, \dots, m$ let $\mathbf{A}_{r,n}$ with elements $(a_{ijr})_{i,j=1,\dots,n}$ be an $n \times n$ non-stochastic symmetric real matrix with $\sup_{1 \leq j \leq n, n \geq 1} \sum_{i=1}^n |a_{ijr}| < \infty$,
- (b) and let $\mathbf{a}_r = (a_{ir}, \dots, a_{nr})^\top$ be a $n \times 1$ non-stochastic real vector with $\sup_n \frac{\sum_{i=1}^n |a_{ir}|^{\delta_1}}{n} < \infty$ for some $\delta_1 > 2$.
- (c) Let $\boldsymbol{\varepsilon} = (\epsilon_1, \dots, \epsilon_n)^\top$ be an $n \times 1$ random vector with the ϵ_i distributed totally independent with $\mathbb{E}[\epsilon_i] = 0$, $\mathbb{E}[\epsilon_i^2]$, and $\sup_{1 \leq i \leq n, n \geq 1} \mathbb{E}|\epsilon_i|^{\delta_2} < \infty$ for some $\delta_2 > 4$.

Consider the $m \times 1$ vector of linear quadratic forms $\mathbf{v}_n = [Q_{1n}, \dots, Q_{mn}]'$ with:

$$Q_{rn} = \boldsymbol{\varepsilon}' \mathbf{A}_r \boldsymbol{\varepsilon} + \mathbf{a}_r' \boldsymbol{\varepsilon} = \sum_{i=1}^n \sum_{j=1}^n a_{ijr} \epsilon_i \epsilon_j + \sum_{i=1}^n a_{ir} \epsilon_i. \quad (3.37)$$

Let $\mu_v = \mathbf{E}[\mathbf{v}_n] = [\mu_{Q_1}, \dots, \mu_{Q_m}]^\top$ and $\boldsymbol{\Sigma}_{v_n} = [\sigma_{Q_{rs}}]_{r,s=1,\dots,m}$ denote the mean and VC matrix of \mathbf{v}_n , respectively, then:

$$\begin{aligned} \mu_{Q_r} &= \sum_{i=1}^n a_{iir} \sigma_i^2 \\ \sigma_{Q_{rs}} &= 2 \sum_{i=1}^n \sum_{j=1}^n a_{ijr} a_{ijs} \sigma_i^2 \sigma_j^2 + \sum_{i=1}^n a_{ir} a_{is} \sigma_i^2 \\ &\quad + \sum_{i=1}^n a_{iir} a_{iis} [\mu_i^{(4)} - 3\mu_i^4] + \sum_{i=1}^n (a_{ir} a_{iis} + a_{is} a_{iir}) \mu_i^{(3)} \end{aligned}$$

with $\mu_i^{(3)} = \mathbf{E}(\epsilon_i^3)$ and $\mu_i^{(4)} = \mathbf{E}(\epsilon_i^4)$. Furthermore, given that $n^{-1} \lambda_{\min}(\boldsymbol{\Sigma}_{v_n}) \geq c$ for some $c > 0$, then

$$\boldsymbol{\Sigma}_{v_n}^{-1/2}(\mathbf{v}_n - \mu_{v_n}) \xrightarrow{d} \mathbf{N}(\mathbf{0}, \mathbf{I}_m)$$

and thus:

$$n^{-1/2}(\mathbf{v}_n - \mu_{v_n}) \overset{a}{\sim} \mathbf{N}(\mathbf{0}, n^{-1} \boldsymbol{\Sigma}_{v_n})$$

Kelejian and Prucha (2001) introduced a CLT for a single quadratic form under the assumptions useful for spatial models. The generalization to vectors of linear quadratic forms is given in Kelejian and Prucha (2010).

3.13 Exercises

Exercise 3.1 Provide a proof of Proposition 3.1.

Exercise 3.2 Let \mathbf{A}_n be a $k \times k$ matrix and let \mathbf{b}_n be a $k \times 1$ vector. If $\mathbf{A}_n = o(1)$ and $\mathbf{b}_n = O(1)$, show that $\mathbf{A}_n \mathbf{b}_n = o(1)$.

Exercise 3.3 Prove the following result for the 2SLS estimator. Suppose: (i) $y_i = \mathbf{x}_i^\top \boldsymbol{\beta}_0 + \epsilon_i$, $i = 1, \dots, n$, $\boldsymbol{\beta}_0 \in \mathbb{R}^k$; (ii) $\mathbf{Z}^\top \boldsymbol{\varepsilon} / n \xrightarrow{p} \mathbf{0}$; (iii) $\mathbf{Z}^\top \mathbf{X} / n \xrightarrow{p} \mathbf{Q}$, finite with full column rank;

(iv) $\widehat{\mathbf{P}}_n \xrightarrow{p} \mathbf{P}$, finite, symmetric, and positive definite. Then $\widehat{\boldsymbol{\beta}}_n$ exists in probability, and $\widehat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$.

Appendix

3.A Inequalities

Definition 3.A.1 — Multiplicativity of absolute value. For any two random variables a and b

$$|ab| = |a| |b| \quad (3.38)$$

Definition 3.A.2 — Triangle Inequality. For any real numbers x_j ,

$$\left| \sum_{j=1}^n x_j \right| \leq \sum_{j=1}^n |x_j| \quad (3.39)$$

Definition 3.A.3 — Chebyshev's inequality. If X_n is a random variable with mean μ and finite variance, then, for every $\delta > 0$,

$$\Pr [|X_n - \mu| \geq \delta] \leq \frac{\mathbb{E} [(X_n - \mu)^2]}{\delta^2}$$

To prove the Chebyshev inequality, we use the Markov's Inequality

Definition 3.A.4 — Markov's inequality. If X_n is a nonnegative random variable, then for every $\delta > 0$,

$$\Pr [X_n \geq \delta] \leq \frac{\mathbb{E} [X_n]}{\delta}$$

Definition 3.A.5 — Jensen's Inequality. If $g(X_n)$ is a concave function of X_n then

$$g [\mathbb{E}(X_n)] \geq \mathbb{E} [g(X_n)]$$

Definition 3.A.6 — Cauchy-Schwarz Inequality. For two random variables

$$\mathbb{E} [|X_n Y_n|] \leq \left\{ \mathbb{E} [X_n^2] \right\}^{1/2} \left\{ \mathbb{E} [Y_n^2] \right\}^{1/2}$$

Maximum Likelihood Estimation

In this chapter we begin the study of the estimation methods for spatial models. In particular, we focus in the maximum likelihood estimation method. However, it is important to know some basic of the different estimation methods.

Spatial econometric models can be estimated by maximum likelihood (ML) (Ord, 1975), quasi-maximum likelihood (QML) (Lee, 2004), instrumental variables (IV) (Anselin, 1988, pp. 82-86), generalized method of moments (GMM) (Kelejian and Prucha, 1998, 1999), or by Bayesian Markov Chain Monte Carlo method (Bayesian MCMC) (LeSage, 1997).

As we will see in this chapter, the main drawback of the ML estimation is the assumption of normality of the error terms. QML and IV/GMM have the advantage that they do not rely on the assumption of normality of the disturbances. However, both estimators assume that the disturbance terms are independently and identically distributed for all i with zero mean and variance σ^2 . IV/GMM estimator has the disadvantage that the estimate for ρ or λ may be out of the parameter space. These coefficients are restricted to the interval $(1/r_{\min})$ by the Jacobian term in the ML estimation. This issue motivated the development of the IV/GMM, which do not require the Jacobian term. To instrument the spatially lagged dependent variable, Kelejian et al. (2004) suggest $[\mathbf{X}, \mathbf{W}\mathbf{X}, \dots, \mathbf{W}^g\mathbf{X}]$, where g is a pre-selected constant.

4.1 What Are The Consequences of Applying OLS?

We start this chapter analyzing the consequences of applying OLS model on a sample that follows a SAR process. The main result is that the estimated coefficients will be biased and inconsistent. This means that the estimated parameters will not be close to the true parameters, even if you have a very large data set, which is a serious problem.

4.1.1 Finite and Asymptotic Properties

First, we will show that an OLS estimate of ρ will be biased under the SLM. To do so, and not get lost with the notation, consider the following pure first order spatial autoregressive model:

$$\underset{(n \times 1)}{\mathbf{y}} = \rho_0 \underset{(n \times 1)}{\mathbf{W}\mathbf{y}} + \underset{(n \times 1)}{\boldsymbol{\varepsilon}}, \quad (4.1)$$

where ρ_0 is the true population parameter of the data generating process (DGP). The reduced form for the **pure SLM** in (4.1) is:

$$\mathbf{y} = (\mathbf{I}_n - \rho_0 \mathbf{W})^{-1} \boldsymbol{\varepsilon}. \quad (4.2)$$

As a result, the spatial lag term equals:

$$\mathbf{W}\mathbf{y} = \mathbf{W} (\mathbf{I}_n - \rho_0 \mathbf{W})^{-1} \boldsymbol{\varepsilon}. \quad (4.3)$$

This result will be useful later. Now, recall that if the model is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, then the OLS estimator is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$. Then, considering (4.1) the OLS estimator for ρ_0 is:

$$\hat{\rho}_{OLS} = \left[\underbrace{(\mathbf{W}\mathbf{y})^\top}_{(1 \times n)} \underbrace{(\mathbf{W}\mathbf{y})}_{(n \times 1)} \right]^{-1} \underbrace{(\mathbf{W}\mathbf{y})^\top}_{(1 \times n)} \underbrace{\mathbf{y}}_{(n \times 1)}. \quad (4.4)$$

Substituting the expression for \mathbf{y} from the population Equation (4.1) into Equation (4.4) gives us the following sampling error equation:

$$\begin{aligned} \hat{\rho}_{OLS} &= \rho_0 + [(\mathbf{W}\mathbf{y})^\top (\mathbf{W}\mathbf{y})]^{-1} (\mathbf{W}\mathbf{y})^\top \boldsymbol{\varepsilon} \\ &= \rho_0 + \left(\sum_{i=1}^n \mathbf{y}_{Li}^2 \right)^{-1} \left(\sum_{i=1}^n \mathbf{y}_{Li} \epsilon_i \right), \end{aligned}$$

where \mathbf{y}_{Li} is the i th element of the spatial lag operator $\mathbf{W}\mathbf{y} = \mathbf{y}_L$. Assuming that \mathbf{W} is nonstochastic, the mathematical expectation of $\hat{\rho}_{OLS}$ is

$$\begin{aligned} \mathbb{E}(\hat{\rho}_{OLS} | \mathbf{W}) &= \rho_0 + \mathbb{E} \left([(\mathbf{W}\mathbf{y})^\top (\mathbf{W}\mathbf{y})]^{-1} (\mathbf{W}\mathbf{y})^\top \boldsymbol{\varepsilon} \middle| \mathbf{W} \right) \\ &= \rho_0 + \left(\sum_{i=1}^n \mathbf{y}_{Li}^2 \right)^{-1} \mathbb{E} \left(\sum_{i=1}^n \mathbf{y}_{Li} \epsilon_i \middle| \mathbf{W} \right). \end{aligned} \quad (4.5)$$

From (4.5) it is clear that if the expectation of the last term is zero, then $\hat{\rho}_{OLS}$ will be unbiased. However, note that

$$\begin{aligned} \mathbb{E} \left(\sum_{i=1}^n \mathbf{y}_{Li} \epsilon_i \middle| \mathbf{W} \right) &= \mathbb{E} [(\mathbf{W}\mathbf{y})^\top \boldsymbol{\varepsilon} | \mathbf{W}] \\ &= \mathbb{E} \left[\boldsymbol{\varepsilon}^\top (\mathbf{I} - \rho \mathbf{W}^\top)^{-1} \mathbf{W}^\top \boldsymbol{\varepsilon} \middle| \mathbf{W} \right] \quad \text{using (4.3)} \\ &= \mathbb{E} [\boldsymbol{\varepsilon}^\top \mathbf{C}^\top \boldsymbol{\varepsilon} | \mathbf{W}] \\ &= \mathbb{E} [\text{tr} \boldsymbol{\varepsilon}^\top \mathbf{C}^\top \boldsymbol{\varepsilon} | \mathbf{W}] \\ &= \mathbb{E} [\text{tr} \mathbf{C}^\top \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top | \mathbf{W}] \\ &= \text{tr}(\mathbf{C}) \mathbb{E} (\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top | \mathbf{W}) \quad \text{since } \text{tr}(\mathbf{A}) = \text{tr}(\mathbf{A}^\top) \\ &\neq 0, \end{aligned} \quad (4.6)$$

where $\mathbf{C} = \mathbf{W} (\mathbf{I} - \rho \mathbf{W})^{-1}$. Therefore, given the result in (4.6) we have that $\mathbb{E}(\hat{\rho}_{OLS} | \mathbf{W}) = \rho_0$ if and only if $\text{tr}(\mathbf{C}) = 0$, which occurs if $\rho_0 = 0$. If $\rho = 0$, $\mathbf{C} = \mathbf{W}$, and $\text{tr}(\mathbf{C}) = \text{tr}(\mathbf{W}) = 0$

because the diagonal elements of \mathbf{W} are zeros (See Definition 4.1.1 for properties of the trace). In other words, if the true model follows a spatial autoregressive structure, the OLS estimate of ρ will be biased.

Definition 4.1.1 — Some useful results on trace. The **trace** of a squared matrix \mathbf{A} , denoted $\text{tr}(\mathbf{A})$, is defined to be the sum of the elements on the main diagonal of \mathbf{A} :

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^n a_{ii} = a_{11} + a_{22} + \dots + a_{nn} \quad (4.7)$$

where a_{ii} denotes the entry on the i th row and i th column of \mathbf{A} .

Some properties:

(a) Let \mathbf{A} and \mathbf{B} be square matrices and c a scalar. Then:

$$\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B}) \quad (4.8)$$

$$\text{tr}(c\mathbf{A}) = c \text{tr}(\mathbf{A}) \quad (4.9)$$

(b) $\text{tr}(\mathbf{A}) = \text{tr}(\mathbf{A}^\top)$.

(c) $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$.

(d) Trace of an idempotent matrix: Let \mathbf{A} be an idempotent matrix, then $\text{tr}(\mathbf{A}) = \text{rank}(\mathbf{A})$.

What about consistency? Note that we can write:

$$\hat{\rho}_{OLS} = \rho_0 + \left(\frac{1}{n} \sum_{i=1}^n \mathbf{y}_{Li}^2 \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{y}_{Li} \epsilon_i \right). \quad (4.10)$$


Under ‘some conditions’ we can show that:

$$\frac{1}{n} \sum_{i=1}^n \mathbf{y}_{Li}^2 \rightarrow q, \quad (4.11)$$

where q is some finite scalar (We need some assumptions here about ρ and the structure of the spatial weight matrix). However, for the second term we obtain

$$\frac{1}{n} \sum_{i=1}^n \mathbf{y}_{Li} \epsilon_i \xrightarrow{p} \mathbb{E}(\mathbf{y}_{Li} \epsilon_i) = \text{tr}(\mathbf{C}) \mathbb{E}(\epsilon \epsilon^\top) \neq 0. \quad (4.12)$$

As a result, the presence of the spatial weight matrix results in a quadratic form in the error terms, which in turns introduces a form of endogeneity because the spatial lag $\mathbf{W}\mathbf{y}$ will be correlated with the disturbance vector ϵ . Therefore $\hat{\rho}_{OLS}$ is inconsistent, and we need to account for the simultaneity by either in a maximum likelihood estimation framework, or by using a proper set of instrumental variables.

 **Lee (2002)** shows that in some cases the OLS estimator may still consistent and even be asymptotically efficient relative to some other estimators.

4.1.2 Illustration of Bias

We will perform a simple simulation experiment to assess the properties of the OLS estimator when the data generating process follows a spatial lag model. The basic design of the experiment consists of generating simulated observations from a known data generating process (from a SLM model in this case), from known parameters, and then estimate the parameters for each simulated sample. If the parameter is biased then, on average, the estimated parameters should be far away from the true parameter.

For our simulation experiment, we will assume that the true DGP is:

$$\mathbf{y} = \rho_0 \mathbf{W} \mathbf{y} + \boldsymbol{\varepsilon} \quad (4.13)$$

where the true value $\rho_0 = 0.7$; the sample size for each sample is $n = 225$; $\boldsymbol{\varepsilon} \sim N(0, 1)$ and \mathbf{W} is an artificial $n \times n$ weight matrix. The \mathbf{W} is constructed from a neighbor list for rook contiguity on a 500×500 regular lattice.

The syntax for creating the global parameters for the simulation in R is the following:

```
# Global parameters
library("spdep")
library("spatialreg")
set.seed(123)                                # Set seed
S      <- 100                                # Number of simulations
n      <- 225                                # Spatial units
rho    <- 0.7                                # True rho
w      <- cell2nb(sqrt(n), sqrt(n))          # Create artificial W matrix
iw     <- invIrM(w, rho)                     # Compute inverse of (I - rho*W)
rho_hat <- vector(mode = "numeric", length = S) # Vector to save results.
```

The function `cell2nb` creates a list of neighbors for a grid of cells. By default it creates neighbors based on rook criteria. The `invIrM` function generates the full weights \mathbf{W} , checks that ρ lies in its feasible range between $1/\min \boldsymbol{\omega}$ and $1/\max \boldsymbol{\omega}$, where $\boldsymbol{\omega} = \text{eigen}(\mathbf{W})$, and returns the $n \times n$ inverted matrix $(\mathbf{I}_n - \rho \mathbf{W})^{-1}$.

The loop for the simulation is the following:

```
# Loop for simulation
for (s in 1:S) {
  e <- rnorm(n, mean = 0, sd = 1) # Create error term
  y <- iw %*% e                    # True DGP
  Wy <- lag.listw(nb2listw(w), y) # Create spatial lag
  out <- lm(y ~ Wy)                # Estimate OLS
  rho_hat[s] <- coef(out)["Wy"]    # Save results
}
```

Note that since \mathbf{W} is considered as fixed (nonstochastic) it is created out of simulation loop. The results are the following:

```
# Summary of rho_hat
```

```
summary(rho_hat)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.8309  0.9981  1.0331  1.0332  1.0751  1.1680
```

It can be noticed that the estimated ρ ranges from 0.8 to 1.2, that is, the range does not include the true parameter $\rho_0 = 0.7$. Moreover, the mean of the estimated parameters is 1, which is very far away from 0.7! We can conclude that the OLS estimator of the pure SLM model is highly biased.

Finally, we can plot the sampling distribution of the estimated parameters in the following way:

```
# Plot density of estimated rho_hat.
```

```
plot(density(rho_hat),
```

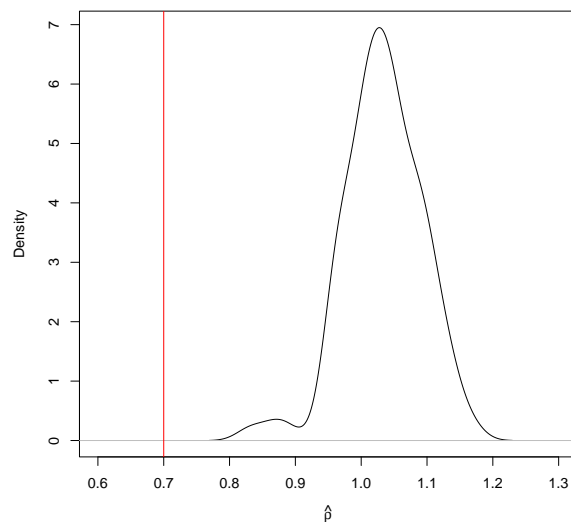
```
      xlab = expression(hat(rho)),
```

```
      main = "")
```

```
abline(v = rho, col = "red")
```

Figure 4.1 present the sampling distribution of ρ estimated by OLS for each sample in the Monte Carlo simulation study. The observed pattern is the same as previously discussed: the distribution does not contain $\rho_0 = 0.7$.

Figure 4.1: Distribution of $\hat{\rho}$



Notes: This graph shows the sampling distribution of ρ estimated by OLS for each sample in the Monte Carlo simulation study. The true DGP follows a pure Spatial Lag Model where the true parameter is $\rho_0 = 0.7$

4.2 Maximum Likelihood Estimation of SLM

Maximum Likelihood (ML) estimation of spatial lag and spatial error regression models was first derived by [Ord \(1975\)](#). The starting point is the assumption of normality for the error terms. The joint likelihood then follows from the multivariate normal distribution for the dependent variable \mathbf{y} . But unlike the classic OLS, the joint log likelihood for a spatial regression does not equal the sum of the log likelihoods associated with the individual observations. This is due to the spatial simultaneity of the system.

In this Section, we will give further insights about these issues. In particular, we derived the ML estimation procedure for the Spatial Lag Model following very close to [Ord \(1975\)](#) and [Anselin \(1988, chapter 6\)](#).

4.2.1 Maximum Likelihood Function

The SLM model is given by the following structural model:

$$\begin{aligned}\mathbf{y} &= \rho_0 \mathbf{W} \mathbf{y} + \mathbf{X} \boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}, \\ \boldsymbol{\varepsilon} &\sim N(\mathbf{0}_n, \sigma_0^2 \mathbf{I}_n),\end{aligned}\tag{4.14}$$

where \mathbf{y} is a vector $n \times 1$ that collects the dependent variable for each spatial unit; \mathbf{W} is an $n \times n$ spatial weight matrix; \mathbf{X} is an $n \times k$ matrix of independent variables; $\boldsymbol{\beta}_0$ is a known k -dimensional vector of parameters; ρ_0 measures the degree of spatial correlation; and $\boldsymbol{\varepsilon}$ is an n -dimensional vector of error terms. Note that we are making the explicit assumption that the error terms follow a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\sigma_0^2 \mathbf{I}_n$. That is, we are assuming that all spatial units have the same error variance.

Since we are explicitly assuming the distribution of the error term, we will be able to use the maximum likelihood estimation procedure. Under the maximum likelihood criterion, the parameter estimates $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^\top, \hat{\rho}, \hat{\sigma}^2)^\top$ are chosen so as to maximize the probability of generating or obtaining the observed sample. However, it should be noted that ML estimation is a highly parametric approach, which means that it is based on strong assumptions. We will see that within these assumptions, it has optimal asymptotic properties (such as consistency and asymptotic efficiency), but when the assumptions are violated, the optimal properties may no longer hold. How can we estimate $\boldsymbol{\theta}_0$? Note that we can rearrange the model as:

$$\mathbf{y} - \rho_0 \mathbf{W} \mathbf{y} = \mathbf{X} \boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}.$$

Using the OLS estimator, an estimate for $\boldsymbol{\beta}_0$ would be

$$\hat{\boldsymbol{\beta}}(\rho_0) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{I}_n - \rho_0 \mathbf{W}) \mathbf{y},$$

which depend on ρ_0 . Similarly, an estimate for the variance parameter would be

$$\hat{\sigma}^2(\rho_0) = \frac{\hat{\boldsymbol{\varepsilon}}(\rho_0)^\top \hat{\boldsymbol{\varepsilon}}(\rho_0)}{n},$$

which also depends on ρ_0 , and where the residuals $\hat{\boldsymbol{\varepsilon}}(\rho_0)$ are given by $\hat{\boldsymbol{\varepsilon}}(\rho_0) = \mathbf{y} - \rho_0 \mathbf{W} \mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}$. Since $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ depend on ρ_0 , we can **concentrate** the full log-likelihood with respect to the

parameters β, σ^2 and reduce maximum likelihood to an univariate optimization problem in the parameter ρ . This will be very useful later in order to derive the ML algorithm.

In order to derive the joint distribution of the data, we need to find the probability density function $f(y_1, y_2, \dots, y_n | \mathbf{X}; \boldsymbol{\theta}) = f(\mathbf{y} | \mathbf{X}; \boldsymbol{\theta})$, that is, the joint conditional distribution of \mathbf{y} given \mathbf{X} . Using the **Transformation Theorem**, we know that

$$f(\mathbf{y} | \mathbf{X}; \boldsymbol{\theta}) = f(\boldsymbol{\varepsilon}(\mathbf{y}) | \mathbf{X}; \boldsymbol{\theta}) \left| \frac{\partial \boldsymbol{\varepsilon}}{\partial \mathbf{y}} \right|.$$

where $|\cdot|$ is the determinant function.

Recall that the error term can be written as $\boldsymbol{\varepsilon} = \mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ with $\mathbf{A} = \mathbf{I}_n - \rho\mathbf{W}$ where $\mathbf{A}\mathbf{y}$ is the **spatially filtered dependent variable**, i.e., with the effect of spatial autocorrelation taken out. Note that $\boldsymbol{\varepsilon} = f(\mathbf{y})$. That is, the error vector is a functional form of the observed \mathbf{y} .¹ To move from the the distribution of the error term to the distribution for the observable random variable \mathbf{y} we need the Jacobian transformation:

$$\det \left(\frac{\partial \boldsymbol{\varepsilon}}{\partial \mathbf{y}} \right) = \det(\mathbf{J}) = \det(\mathbf{A}) = \det(\mathbf{I}_n - \rho\mathbf{W}),$$

where $\mathbf{J} = \left(\frac{\partial \boldsymbol{\varepsilon}}{\partial \mathbf{y}} \right)$ is the $n \times n$ Jacobian matrix, and $\det(\mathbf{I}_n - \rho\mathbf{W})$ is the determinant of a $n \times n$ matrix. In contrast to the time-series case, the spatial Jacobian is not the determinant of a triangular matrix, but of a full matrix. This may complicate its computation considerably. The Jacobian reduces to a scalar 1 in the standard regression model, since the partial derivative becomes $|\partial(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/\partial \mathbf{y}| = |\mathbf{I}_n| = 1$.

Using the density function of the multivariate normal distribution we can find the joint pdf of $\boldsymbol{\varepsilon} | \mathbf{X}$.² By recognizing that $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, we can write:

$$f(\boldsymbol{\varepsilon} | \mathbf{X}) = (2\pi \cdot \sigma^2)^{-n/2} \exp \left[-\frac{1}{2\sigma^2} \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} \right].$$

Given an i.i.d sample of n observations, \mathbf{y} and \mathbf{X} , the joint density of the observed sample is:

$$f(\mathbf{y} | \mathbf{X}; \boldsymbol{\theta}) = (2\pi \cdot \sigma^2)^{-n/2} \exp \left[-\frac{1}{2\sigma^2} (\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right] \det \left(\frac{\partial(\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\partial \mathbf{y}} \right).$$

Note that the likelihood function is defined as the joint density treated as a function of the parameters: $L(\boldsymbol{\theta} | \mathbf{y}, \mathbf{X}) = f(\mathbf{y} | \mathbf{X}; \boldsymbol{\theta})$. Finally, the log-likelihood function, which will be

¹Since y_i , and not ϵ_i , are the observed quantities, the parameters must be estimated by maximizing $L(\mathbf{y})$, not $L(\boldsymbol{\varepsilon})$. For more details about this, see Mead (1967) and Doreian (1981).

²The multivariate normal distribution of an n -dimensional random vector \mathbf{x} with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ can be written as

$$(2\pi)^{-n/2} \det(\boldsymbol{\Sigma})^{-1/2} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right).$$

maximized, takes the form:³

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \log |\mathbf{A}| - \frac{n \log(2\pi)}{2} - \frac{n \log(\sigma^2)}{2} - \frac{1}{2\sigma^2} (\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \log |\mathbf{A}| - \frac{n \log(2\pi)}{2} - \frac{n \log(\sigma^2)}{2} - \frac{1}{2\sigma^2} [\mathbf{y}^\top \mathbf{A}^\top \mathbf{A} \mathbf{y} - 2 (\mathbf{A}\mathbf{y})^\top \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta}],\end{aligned}\tag{4.15}$$

where this development uses the fact that the transpose of a scalar is the scalar, i.e., $\mathbf{y}^\top \mathbf{A}^\top \mathbf{X} \boldsymbol{\beta} = (\mathbf{y}^\top \mathbf{A} \mathbf{X} \boldsymbol{\beta})^\top = \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{A} \mathbf{y}$. This is similar to the typical linear-normal likelihood, except that the transformation from $\boldsymbol{\varepsilon}$ to \mathbf{y} , is not by the usual factor of 1, but by $\log |\mathbf{A}|$.

As we will show in Section 4.7.1, we can directly estimate the $\boldsymbol{\theta}$ by maximizing the log-likelihood function (4.15) using a constrained optimization algorithm. However, as shown in the next Section, we can create a more easy estimation algorithm by concentrating the log-likelihood function.

4.2.2 Score Vector and Estimates

In order to find the ML estimates for the SLM model, we need to maximize $\ell(\boldsymbol{\theta})$ in Equation (4.15) with respect to $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2, \rho)^\top$. To do so, we need to find the first order condition (FONC) of this optimization problem.

Before taking derivatives, it is useful to review some important properties of matrix calculus given in the next definition.

Definition 4.2.1 — Some useful results on matrix calculus. Some important results are the followings:

$$\frac{\partial(\rho \mathbf{W})}{\partial \rho} = \mathbf{W}\tag{4.16}$$

$$\begin{aligned}\frac{\partial \mathbf{A}}{\partial \rho} &= \frac{\partial(\mathbf{I}_n - \rho \mathbf{W})}{\partial \rho} \\ &= \frac{\partial \mathbf{I}_n}{\partial \rho} - \frac{\partial \rho \mathbf{W}}{\partial \rho} \\ &= -\mathbf{W}\end{aligned}\tag{4.17}$$

$$\frac{\partial \log |\mathbf{A}|}{\partial \rho} = \text{tr}(\mathbf{A}^{-1} \partial \mathbf{A} / \partial \rho) = \text{tr}[\mathbf{A}^{-1}(-\mathbf{W})]\tag{4.18}$$

Let $\boldsymbol{\varepsilon} = \mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$, then:

$$\frac{\partial \boldsymbol{\varepsilon}}{\partial \rho} = \frac{\partial(\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\partial \rho} = -\mathbf{W}\mathbf{y}\tag{4.19}$$

$$\frac{\partial \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}}{\partial \rho} = \boldsymbol{\varepsilon}^\top (\partial \boldsymbol{\varepsilon} / \partial \rho) + (\partial \boldsymbol{\varepsilon}^\top / \partial \rho) \boldsymbol{\varepsilon} = 2\boldsymbol{\varepsilon}^\top (\partial \boldsymbol{\varepsilon} / \partial \rho) = 2\boldsymbol{\varepsilon}^\top (-\mathbf{W})\mathbf{y}\tag{4.20}$$

³Since the constant $-\frac{n \log(2\pi)}{2}$ is not a function of any of the parameters, some software programs do not include it when reporting maximized log-likelihood. See [Bivand and Piras \(2015\)](#).

$$\frac{\partial \mathbf{A}^{-1}}{\partial \rho} = -\mathbf{A}^{-1}(\partial \mathbf{A} / \partial \rho) \mathbf{A}^{-1} = \mathbf{A}^{-1} \mathbf{W} \mathbf{A}^{-1} \quad (4.21)$$

$$\frac{\partial \text{tr}(\mathbf{A}^{-1} \mathbf{W})}{\partial \rho} = \text{tr}(\partial \mathbf{A}^{-1} \mathbf{W} / \partial \rho) \quad (4.22)$$

Taking the derivative of Equation (4.15) respect to β , we obtain:

$$\frac{\partial \ell(\theta)}{\partial \beta} = -\frac{1}{2\sigma^2} \left[-2 \left((\mathbf{A}\mathbf{y})^\top \mathbf{X} \right)^\top + 2\mathbf{X}^\top \mathbf{X} \beta \right] = \frac{1}{\sigma^2} \mathbf{X}^\top (\mathbf{A}\mathbf{y} - \mathbf{X}\beta), \quad (4.23)$$

and with respect to σ^2 yields

$$\frac{\partial \ell(\theta)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (\mathbf{A}\mathbf{y} - \mathbf{X}\beta)^\top (\mathbf{A}\mathbf{y} - \mathbf{X}\beta). \quad (4.24)$$

Setting both (4.23) and (4.24) to 0 and solving, we obtain

$$\hat{\beta}_{ML}(\rho) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{A}\mathbf{y} \quad (4.25)$$

$$\hat{\sigma}_{ML}^2(\rho) = \frac{(\mathbf{A}\mathbf{y} - \mathbf{X}\hat{\beta}_{ML})^\top (\mathbf{A}\mathbf{y} - \mathbf{X}\hat{\beta}_{ML})}{n}. \quad (4.26)$$

Note that conditional on ρ (assuming we know ρ), these estimates are simply OLS applied to the *spatial filtered* dependent variable $\mathbf{A}\mathbf{y}$ and the exploratory variables \mathbf{X} . Moreover, after some manipulation, Equation (4.25) can be re-written as

$$\begin{aligned} \hat{\beta}_{ML}(\rho) &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} - \rho (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}\mathbf{y}, \\ &= \hat{\beta}_O - \rho \hat{\beta}_L. \end{aligned} \quad (4.27)$$

Note that the first term in (4.27) is just the OLS regression of \mathbf{y} on \mathbf{X} , whereas the second term is just ρ times the OLS regression of $\mathbf{W}\mathbf{y}$ on \mathbf{X} . Next, define the following:

$$\mathbf{e}_O = \mathbf{y} - \mathbf{X}\hat{\beta}_O \text{ and } \mathbf{e}_L = \mathbf{W}\mathbf{y} - \mathbf{X}\hat{\beta}_L. \quad (4.28)$$

Then, plugging (4.27) into (4.26)

$$\tilde{\sigma}^2(\rho) = \frac{(\mathbf{e}_O - \rho \mathbf{e}_L)^\top (\mathbf{e}_O - \rho \mathbf{e}_L)}{n}. \quad (4.29)$$

Note that both (4.27) and (4.29) rely only on observables, except for ρ , and so are readily calculable given some estimate of ρ . Therefore, plugging (4.27) and (4.29) back into the likelihood (4.15) we obtain the **concentrated log-likelihood function**:

$$\ell(\rho) = -\frac{n}{2} - \frac{n}{2} \log(2\pi) - \frac{n}{2} \log \left[\frac{(\mathbf{e}_O - \rho \mathbf{e}_L)^\top (\mathbf{e}_O - \rho \mathbf{e}_L)}{n} \right] + \log |\mathbf{I}_n - \rho \mathbf{W}|, \quad (4.30)$$

which is a **nonlinear** function of a single parameter ρ . A ML estimate for ρ is obtained from a numerical optimization of the concentrated log-likelihood function (4.30). Once we obtain $\hat{\rho}$, we can easily obtain $\hat{\beta}$. The procedure can be summarized in the following steps.

Algorithm 4.1 — ML estimation of SLM. The algorithm to perform the ML estimation of the SLM is the following:

- (a) Perform the two auxiliary regression of \mathbf{y} and $\mathbf{W}\mathbf{y}$ on \mathbf{X} to obtain $\hat{\beta}_O$ and $\hat{\beta}_L$ as in Equation (4.27).
- (b) Use $\hat{\beta}_O$ and $\hat{\beta}_L$ to compute the residuals in Equation (4.28).
- (c) Maximize the concentrated likelihood given in Equation (4.30) by numerical optimization to obtain an estimate of ρ .
- (d) Use the estimate of $\hat{\rho}$ to plug it back in to the expression for β (Equation 4.25) and σ^2 (Equation 4.26).

Since the score function will be important for understanding the asymptotic theory of MLE, we will derive also $\partial\ell(\theta)/\partial\rho$. Taking the derivative of Equation (4.15) respect to ρ , we obtain:

$$\begin{aligned}
 \frac{\partial\ell(\theta)}{\partial\rho} &= \left(\frac{\partial}{\partial\rho}\right) \log|\mathbf{A}| - \frac{1}{2\sigma^2} \left(\frac{\partial}{\partial\rho}\right) \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}, \\
 &= -\text{tr}(\mathbf{A}^{-1}\mathbf{W}) + \frac{1}{2\sigma^2} 2\boldsymbol{\varepsilon}^\top \mathbf{W}\mathbf{y}, \quad \text{using (4.18) and (4.20)} \\
 &= -\text{tr}(\mathbf{A}^{-1}\mathbf{W}) + \frac{1}{2\sigma^2} 2\boldsymbol{\varepsilon}^\top \mathbf{W}\mathbf{y}, \\
 &= -\text{tr}(\mathbf{A}^{-1}\mathbf{W}) + \frac{1}{\sigma^2} \boldsymbol{\varepsilon}^\top \mathbf{W}\mathbf{y}.
 \end{aligned} \tag{4.31}$$

Thus the complete gradient (or score function) is:

$$\nabla_{\theta} = \frac{\partial\ell(\theta)}{\partial\theta} = \begin{pmatrix} \frac{\partial\log L(\theta)}{\partial\beta} \\ \frac{\partial\log L(\theta)}{\partial\sigma^2} \\ \frac{\partial\log L(\theta)}{\partial\rho} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2} \mathbf{X}^\top \boldsymbol{\varepsilon} \\ \frac{1}{2\sigma^4} (\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} - n\sigma^2) \\ -\text{tr}(\mathbf{A}^{-1}\mathbf{W}) + \frac{1}{\sigma^2} \boldsymbol{\varepsilon}^\top \mathbf{W}\mathbf{y} \end{pmatrix}, \tag{4.32}$$

where $\boldsymbol{\varepsilon} = \mathbf{A}\mathbf{y} - \mathbf{X}\beta$.

Note that if we replace $\mathbf{y} = \mathbf{A}^{-1}\mathbf{X}\beta + \mathbf{A}^{-1}\boldsymbol{\varepsilon}$ in Equation (4.31), we get

$$\frac{\partial\ell(\theta)}{\partial\rho} = \frac{1}{\sigma^2} (\mathbf{C}\mathbf{X}\beta)^\top \boldsymbol{\varepsilon} + \frac{1}{\sigma^2} (\boldsymbol{\varepsilon}^\top \mathbf{C}\boldsymbol{\varepsilon} - \sigma^2 \text{tr}(\mathbf{C})),$$

where:

$$\mathbf{C} = \mathbf{W}\mathbf{A}^{-1}. \tag{4.33}$$

This expression will be useful later.

4.2.3 Hessian

The Hessian matrix will be very important in the following sections to obtain the asymptotic variance-covariance matrix. For this reason we devote a complete section in order to derive this matrix for the SLM. In this case, the Hessian is a $(k+2) \times (k+2)$ matrix of second

derivatives given by

$$\mathbf{H}(\boldsymbol{\beta}, \sigma^2, \rho) = \begin{pmatrix} \frac{\ell(\boldsymbol{\beta}, \sigma^2, \rho)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} & \frac{\ell(\boldsymbol{\beta}, \sigma^2, \rho)}{\partial \boldsymbol{\beta} \partial \sigma^2} & \frac{\ell(\boldsymbol{\beta}, \sigma^2, \rho)}{\partial \boldsymbol{\beta} \partial \rho} \\ \frac{\ell(\boldsymbol{\beta}, \sigma^2, \rho)}{\partial \sigma^2 \partial \boldsymbol{\beta}^\top} & \frac{\ell(\boldsymbol{\beta}, \sigma^2, \rho)}{\partial (\sigma^2)^2} & \frac{\ell(\boldsymbol{\beta}, \sigma^2, \rho)}{\partial \sigma^2 \partial \rho} \\ \frac{\ell(\boldsymbol{\beta}, \sigma^2, \rho)}{\partial \rho \partial \boldsymbol{\beta}^\top} & \frac{\ell(\boldsymbol{\beta}, \sigma^2, \rho)}{\partial \rho \partial \sigma^2} & \frac{\ell(\boldsymbol{\beta}, \sigma^2, \rho)}{\partial \rho^2} \end{pmatrix}.$$

Now, we work in the cross-derivatives for $\boldsymbol{\beta}$. From (4.23):

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = -\frac{1}{\sigma^2} (\mathbf{X}^\top \mathbf{X}), \quad (4.34)$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \sigma^2} = -\frac{1}{(\sigma^2)^2} \mathbf{X}^\top \boldsymbol{\varepsilon}, \quad (4.35)$$

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \rho} = -\frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{W} \mathbf{y}, \quad (4.36)$$

Using the first derivative (4.24) and working in the cross-derivatives for σ^2 , we obtain

$$\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial (\sigma^2)^2} = \frac{n}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3} \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}, \quad (4.37)$$

and:

$$\begin{aligned} \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \sigma^2 \partial \rho} &= \frac{1}{2\sigma^4} \left[2\boldsymbol{\varepsilon}^\top \left(\frac{\partial \boldsymbol{\varepsilon}}{\partial \rho} \right) \right] \quad \text{using Equation (4.20)} \\ &= -\frac{1}{\sigma^4} \boldsymbol{\varepsilon}^\top \mathbf{W} \mathbf{y} \\ &= -\frac{\boldsymbol{\varepsilon}^\top \mathbf{W} \mathbf{y}}{\sigma^4} \end{aligned} \quad (4.38)$$

Finally, working in the second derivative of ρ , and using (4.31), we obtain

$$\begin{aligned} \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \rho^2} &= -\left(\frac{\partial}{\partial \rho} \right) \text{tr}(\mathbf{A}^{-1} \mathbf{W}) + \frac{1}{\sigma^2} \left(\frac{\partial}{\partial \rho} \right) \boldsymbol{\varepsilon}^\top \mathbf{W} \mathbf{y}, \\ &= -\text{tr} \left(\frac{\partial \mathbf{A}^{-1} \mathbf{W}}{\partial \rho} \right) + \frac{1}{\sigma^2} \left(\frac{\partial}{\partial \rho} \right) (\mathbf{A} \mathbf{y})^\top \mathbf{W} \mathbf{y}, \\ &= -\text{tr} (\mathbf{A}^{-1} \mathbf{W} \mathbf{A}^{-1} \mathbf{W}) + \frac{1}{\sigma^2} (-\mathbf{y}^\top \mathbf{W}^\top \mathbf{W} \mathbf{y}), \\ &= -\text{tr} [(\mathbf{W} \mathbf{A}^{-1})^2] - \frac{1}{\sigma^2} (\mathbf{y}^\top \mathbf{W}^\top \mathbf{W} \mathbf{y}). \end{aligned} \quad (4.39)$$

Therefore, the Hessian is:

$$\mathbf{H}(\boldsymbol{\beta}, \sigma^2, \rho) = \begin{pmatrix} -\frac{1}{\sigma^2} (\mathbf{X}^\top \mathbf{X}) & -\frac{1}{(\sigma^2)^2} \mathbf{X}^\top \boldsymbol{\varepsilon} & -\frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{W} \mathbf{y} \\ \cdot & \frac{n}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3} \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} & -\frac{\boldsymbol{\varepsilon}^\top \mathbf{W} \mathbf{y}}{\sigma^4} \\ \cdot & \cdot & -\text{tr} [(\mathbf{W} \mathbf{A}^{-1})^2] - \frac{1}{\sigma^2} (\mathbf{y}^\top \mathbf{W}^\top \mathbf{W} \mathbf{y}) \end{pmatrix} \quad (4.40)$$

which is symmetric and \mathbf{C} is given in Equation (4.33).

4.2.4 Ord's Jacobian

An important feature of the concentrated log-likelihood function (4.30) is the Jacobian term $|\mathbf{I}_n - \rho \mathbf{W}|$. Computationally this is burdensome, since determining $\hat{\rho}$ rests on the evaluation of the $n \times n$ matrix $|\mathbf{I}_n - \rho \mathbf{W}|$ in each iteration. However, Ord (1975) note that:

$$|\omega \mathbf{I}_n - \mathbf{W}| = \prod_{i=1}^n (\omega - \omega_i).$$

Therefore:

$$|\mathbf{I}_n - \rho \mathbf{W}| = \prod_{i=1}^n (1 - \rho \omega_i),$$

and the log-determinant term follows as

$$\log |\mathbf{I}_n - \rho \mathbf{W}| = \sum_{i=1}^n \log(1 - \rho \omega_i). \quad (4.41)$$

The advantage of this approach is that the eigenvalues only need to be computed once, which carries some overhead, but greatly speeds up the calculation of the log-likelihood at each iteration. In practice, in all but the smallest data sets (< 4000 observations), the Ord's approach will be faster than the brute force approach.

This new formulation give us the possible domain of ρ . We need that $1 - \rho \omega_i \neq 0$, which occurs only if $1/\omega_{\min} < \rho < 1/\omega_{\max}$. For row-standardized matrix, the largest eigenvalues is 1.

With this new approximation, the new concentrated log-likelihood function is:

$$\ell(\rho) = -\frac{n}{2} - \frac{n}{2} \log(2\pi) - \frac{n}{2} \log \left[\frac{(\mathbf{e}_O - \rho \mathbf{e}_L)^\top (\mathbf{e}_O - \rho \mathbf{e}_L)}{n} \right] + \sum_{i=1}^n \log(1 - \rho \omega_i). \quad (4.42)$$

Another method approach is the characteristic root method outlined in Smirnov and Anselin (2001). This approach allows for the estimation of spatial lag models for very large data sets ($> 100,000$ observations) in a very short time. However, it is limited by the requirement that the weight matrix needs to be intrinsically symmetric. This precludes the use of asymmetric weight such as k -nearest neighbor weights. For other approximations see LeSage and Pace (2010, chapter 4).

4.3 Maximum Likelihood Estimation of SEM

4.3.1 What Are The Consequences of Applying OLS on a SEM Model?

As we reviewed in Section 2.1.3, a second way to incorporate spatial autocorrelation in a regression model is to specify a spatial process for the error term. The SEM model is given by

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta_0 + \mathbf{u} \\ \mathbf{u} &= \lambda_0 \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon} &\sim N(\mathbf{0}, \sigma_0^2 \mathbf{I}_n) \end{aligned} \quad (4.43)$$

where λ_0 is the spatial autoregressive coefficient for the error lag $\mathbf{W}\mathbf{u}$ (to distinguish the notation from the spatial autoregressive coefficient ρ in a spatial lag model), \mathbf{W} is the spatial weight matrix, $\boldsymbol{\varepsilon}$ is the error term such that $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I}_n)$. This model do not require a theoretical model for a spatial process, but instead, is consistent with a situation where determinants of the dependent variable omitted from the model are **spatially autocorrelated**, or with a situation where unobserved shocks follows a spatial pattern (Elhorst, 2014). In summary, SEM treats spatial correlation primarily as a nuisance.

If $\lambda_0 > 0$, then we face positive spatial correlation. This implies clustering of similar values; that is, the errors for spatial unit i tend to vary systematically with the errors for other nearby observations j so that smaller/larger errors for i would tend to go together with smaller/larger errors for j . This violates the typical assumption of no autocorrelation in the error term of the OLS.

The reduced form of the SEM is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + (\mathbf{I}_n - \lambda_0 \mathbf{W})^{-1} \boldsymbol{\varepsilon}.$$

Since $\mathbf{u} = (\mathbf{I}_n - \lambda_0 \mathbf{W})^{-1} \boldsymbol{\varepsilon}$, it can be shown that $\mathbb{E}(\mathbf{u}|\mathbf{W}, \mathbf{X}) = \mathbf{0}$. Furthermore, the variance-covariance matrix of \mathbf{u} is:

$$\mathbb{V}(\mathbf{u}|\mathbf{W}, \mathbf{X}) = \mathbb{E}(\mathbf{u}\mathbf{u}^\top|\mathbf{W}, \mathbf{X}) = \sigma_0^2 (\mathbf{I}_n - \lambda_0 \mathbf{W})^{-1} (\mathbf{I} - \lambda_0 \mathbf{W}^\top)^{-1} = \sigma_0^2 \boldsymbol{\Omega}_u^{-1}, \quad (4.44)$$

where $\boldsymbol{\Omega}_u = (\mathbf{I}_n - \lambda_0 \mathbf{W})(\mathbf{I}_n - \lambda_0 \mathbf{W}^\top)$. The variance covariance (4.44) is a full matrix implying a spatial autoregressive error process leading to a nonzero error covariance between every pair of observations, but decreasing in magnitude with the order of contiguity (Anselin and Bera, 1998). Furthermore, the complex structure in the inverse matrix in (4.44) yields non constant diagonal elements in the error covariance matrix, thus inducing heteroskedasticity in \mathbf{u} , irrespective of the heteroskedasticity of $\boldsymbol{\varepsilon}$. Finally, $\mathbf{u} \sim N(\mathbf{0}, \sigma_0^2 \boldsymbol{\Omega}_u^{-1})$.

R The OLS estimates of model in Equation (4.43) are unbiased, but inefficient if $\lambda_0 \neq 0$.

Given the previous Remark, we might used generalized least squares (GLS) for a more efficient parameters estimation. Recall that the inefficiency of OLS estimates of the regression coefficient would invalidate the statistical inference in the spatial error model. The invalidity of significance test arises from biased estimation of the variance and standard errors of the OLS estimates for $\boldsymbol{\beta}$ and λ .

4.3.2 Log-likelihood function

The model in Equation (4.43) implies that

$$\boldsymbol{\varepsilon} = (\mathbf{I}_n - \lambda_0 \mathbf{W}) \mathbf{y} - (\mathbf{I}_n - \lambda_0 \mathbf{W}) \mathbf{X} \boldsymbol{\beta}_0 = \mathbf{B}_0 \mathbf{y} - \mathbf{B}_0 \mathbf{X} \boldsymbol{\beta}_0,$$

where $\mathbf{B}_0 = (\mathbf{I}_n - \lambda_0 \mathbf{W})$. Recall that in order to create the log-likelihood function we need the joint density function. Using the Transformation Theorem we are able to find the joint conditional function:

$$f(y_1, \dots, y_n | \mathbf{X}; \boldsymbol{\theta}) = f(\boldsymbol{\varepsilon}(\mathbf{y}) | \mathbf{X}; \boldsymbol{\theta}) \cdot |\mathbf{J}|$$

Again, the Jacobian term is not equal to one, but instead is

$$\mathbf{J} = \frac{\partial \boldsymbol{\varepsilon}}{\partial \mathbf{y}} = \mathbf{B}_0.$$

Thus, the joint density function of $\boldsymbol{\varepsilon}$ —which is a function of \mathbf{y} — equals

$$f(\boldsymbol{\varepsilon}(\mathbf{y})|\mathbf{X}; \boldsymbol{\theta}) = (2\pi\sigma^2)^{-n/2} \exp \left[-\frac{[(\mathbf{I}_n - \lambda\mathbf{W})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})]^\top [(\mathbf{I}_n - \lambda\mathbf{W})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})]}{2\sigma^2} \right],$$

and the joint density function of \mathbf{y} , $f(y_1, \dots, y_n|\mathbf{X}; \boldsymbol{\theta})$ equals

$$f(\mathbf{y}|\mathbf{X}; \boldsymbol{\theta}) = (2\pi\sigma^2)^{-n/2} \exp \left[-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{B}^\top \mathbf{B}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2} \right] \cdot |\mathbf{B}|$$

Finally, the log-likelihood can be expressed as

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Omega}(\lambda)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2} + \log |\mathbf{I}_n - \lambda\mathbf{W}|, \quad (4.45)$$

where

$$\boldsymbol{\Omega}(\lambda) = \mathbf{B}^\top \mathbf{B} = (\mathbf{I}_n - \lambda\mathbf{W})^\top (\mathbf{I}_n - \lambda\mathbf{W})$$

Again, we run into complications over the log of the determinant $|\mathbf{I}_n - \lambda\mathbf{W}|$, which is an n th-order polynomial that is cumbersome to evaluate.

4.3.3 Score Function and ML Estimates

Maximizing the log-likelihood function (4.45) is the same as to minimizing the sum of the transformed errors, $\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}$, corrected by the log of the Jacobian, $\log |\mathbf{I}_n - \lambda\mathbf{W}|$. Since we are accounting for this correction, the ML estimates will differ from the OLS estimates. They will coincide if $\lambda \rightarrow 0$.


To obtain the ML estimates, we apply FONC to the log-likelihood function (4.45). Taking the derivative respect to $\boldsymbol{\beta}$ yields:

$$\begin{aligned} \boldsymbol{\beta}_{ML}(\lambda) &= [\mathbf{X}^\top \boldsymbol{\Omega}(\lambda) \mathbf{X}]^{-1} \mathbf{X}^\top \boldsymbol{\Omega}(\lambda) \mathbf{y}, \\ &= [(\mathbf{B}\mathbf{X})^\top (\mathbf{B}\mathbf{X})]^{-1} (\mathbf{B}\mathbf{X})^\top \mathbf{B}\mathbf{y}, \\ &= [\mathbf{X}(\lambda)^\top \mathbf{X}(\lambda)]^{-1} \mathbf{X}(\lambda)^\top \mathbf{y}(\lambda), \end{aligned} \quad (4.46)$$

where:

$$\begin{aligned} \mathbf{X}(\lambda) &= \mathbf{B}\mathbf{X} = (\mathbf{I} - \lambda\mathbf{W})\mathbf{X} = (\mathbf{X} - \lambda\mathbf{W}\mathbf{X}), \\ \mathbf{y}(\lambda) &= (\mathbf{y} - \lambda\mathbf{W}\mathbf{y}). \end{aligned} \quad (4.47)$$

If λ is known, this estimator is equal to the GLS estimator— $\hat{\boldsymbol{\beta}}_{ML} = \hat{\boldsymbol{\beta}}_{GLS}$ —and it can be thought as the OLS estimator resulting from a regression of $\mathbf{y}(\lambda)$ on $\mathbf{X}(\lambda)$. In other words, for a known value of the spatial autoregressive coefficient, λ , this is equivalent to OLS on the transformed variables.

 In the literature, the transformations:

$$\begin{aligned} \mathbf{X}(\lambda) &= (\mathbf{X} - \lambda\mathbf{W}\mathbf{X}) \\ \mathbf{y}(\lambda) &= (\mathbf{y} - \lambda\mathbf{W}\mathbf{y}) \end{aligned}$$

are known as the *Cochrane-Orcutt transformation*.

In the same way, a first-order condition resulting from the spatial derivative of (4.45) with respect to σ^2 gives the ML estimator for the error variance:

$$\sigma_{ML}^2(\lambda) = \frac{1}{n} \left(\hat{\boldsymbol{\varepsilon}}^\top \mathbf{B}^\top \mathbf{B} \hat{\boldsymbol{\varepsilon}} \right) = \frac{1}{n} \hat{\boldsymbol{\varepsilon}}^\top(\lambda) \hat{\boldsymbol{\varepsilon}}(\lambda), \quad (4.48)$$

where $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML}$ and $\hat{\boldsymbol{\varepsilon}}(\lambda) = \mathbf{B}(\lambda)(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML}) = \mathbf{B}(\lambda)\mathbf{y} - \mathbf{B}(\lambda)\mathbf{X}\hat{\boldsymbol{\beta}}_{ML}$.

First order condition derived from the expression of the likelihood are highly non-linear and therefore the likelihood in Equation (4.45) cannot be directly maximized. Again, a concentrated likelihood approach is necessary.

The estimators for $\boldsymbol{\beta}$ and σ^2 are both functions of the value of λ . A concentrated log-likelihood can then be obtained as:

$$\ell(\lambda) = \text{const} + \frac{n}{2} \log \left[\frac{1}{n} \hat{\boldsymbol{\varepsilon}}^\top \mathbf{B}^\top \mathbf{B} \hat{\boldsymbol{\varepsilon}} \right] + \log |\mathbf{B}| \quad (4.49)$$

The residual vector of the concentrated likelihood is also, indirectly, a function of the spatial autoregressive parameter.

A one-time optimization will in general not be sufficient to obtain maximum likelihood estimates for all the parameters. Therefore an interactive procedure will be needed.

Alternate back and forth between the estimation of the spatial autoregressive coefficient conditional upon residuals (for a value of $\boldsymbol{\beta}$), and a estimation of the parameter vector (conditional upon the s.a.c).

Algorithm 4.2 — ML estimation of SEM. Following Anselin (1988), the procedure can be summarize in the following steps:

- (a) Carry out an OLS of $\mathbf{B}\mathbf{X}$ on $\mathbf{B}\mathbf{y}$; get $\hat{\boldsymbol{\beta}}_{OLS}$
- (b) Compute initial set of residuals $\hat{\boldsymbol{\varepsilon}}_{OLS} = \mathbf{B}\mathbf{y} - \mathbf{B}\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}$
- (c) Given $\hat{\boldsymbol{\varepsilon}}_{OLS}$, find $\hat{\lambda}$ that maximizes the concentrated likelihood.
- (d) If the convergence criterion is met, proceed, otherwise repeat steps 1, 2 and 3.
- (e) Given $\hat{\lambda}$, estimate $\hat{\boldsymbol{\beta}}(\lambda)$ by GLS and obtain a new vector of residuals, $\hat{\boldsymbol{\varepsilon}}(\lambda)$
- (f) Given $\hat{\boldsymbol{\varepsilon}}(\lambda)$ and $\hat{\lambda}$, estimate $\hat{\sigma}(\lambda)$.

Finally, the asymptotic variance-covariance matrix is:

$$\text{AsyVar}(\boldsymbol{\beta}, \sigma^2, \lambda) = \begin{pmatrix} \frac{\mathbf{X}(\lambda)^\top \mathbf{X}(\lambda)}{\sigma^2} & 0 & 0 \\ k \times k & & \\ 0 & \frac{n}{2\sigma^4} & \frac{\text{tr}(\mathbf{W}_B)}{\sigma^2} \\ 0 & \frac{\text{tr}(\mathbf{W}_B)}{\sigma^2} & \text{tr}(\mathbf{W}_B)^2 + \text{tr}(\mathbf{W}_B^\top \mathbf{W}_B) \end{pmatrix}^{-1} \quad (4.50)$$

where $\mathbf{W}_B = \mathbf{W}(\mathbf{I} - \lambda\mathbf{W})^{-1}$.

4.4 Asymptotic Properties of SLM

In this section we review the asymptotic properties of the ML and Quasi ML for the SLM. In particular, we follow Lee (2004).

4.4.1 Consistency of QMLE

We now write the true SLM as

$$\mathbf{y}_n = \mathbf{X}_n \boldsymbol{\beta}_0 + \lambda_0 \mathbf{W}_n \mathbf{y}_n + \boldsymbol{\varepsilon}_n,$$

We now index the variables of the model, and \mathbf{W} , by n to indicate that they depend on the sample size.

Lee (2004) derives the asymptotic properties (consistency and asymptotic normality) of the ML and QML estimator for the SLM. Lee (2004) starts with the following assumption about the error terms ϵ_i .

Assumption 4.3 — Errors (Lee, 2004). Assume the following

- (a) The disturbances $\{\epsilon_i\}, i = 1, \dots, n$, in $\boldsymbol{\varepsilon}_n = (\epsilon_1, \dots, \epsilon_n)^\top$ are iid with mean zero and variance σ^2 . Its moment $\mathbb{E}(|\epsilon|^{4+\gamma})$ for some $\gamma > 0$ exists.

Note that Assumption 4.3 states that the error variance is homoskedastic. Moreover, because statistics involving quadratic forms of $\boldsymbol{\varepsilon}_n$ will be present in the estimation, the existence of the fourth order moment of ϵ_i will guarantee finite variances for the quadratic forms and we will be able to apply a CLT.

In order to understand the asymptotic behavior of \mathbf{W}_n under some **regularity conditions**, we need to understand some useful terminologies.

Definition 4.4.1 — Triangular array of constants. Let $\{b_{ni}\}, i = 1, \dots, n$ be a triangular array of constants.

- (a) $\{b_{ni}\}$ are at most of order $(1/h_n)$, denoted by $O(1/h_n)$ uniformly in i if there exists a finite constant c independent of i and n such that $|b_{ni}| \leq \frac{c}{h_n}$ for all i and n .
- (b) $\{b_{ni}\}$ are bounded away from zero uniformly in i at rate of h_n if there exists a positive sequence $\{h_n\}$ and a constant $c > 0$ independent of i and n such that $c \leq |b_{ni}|/h_n$ for all i for sufficiently large n .

Again, we must think the \mathbf{W}_n matrices as triangular arrays of constants. As explained in Section 3.7, as we add more spatial units the spatial structure changes. Thus, it might be the case that the element w_{ij} is not the same when $n = 50$ or $n = 55$. By considering triangular arrays, we make explicit this possibility. This explains why we index the elements of \mathbf{W}_n as $w_{n,ij}$.

Another question is whether each element of \mathbf{W}_n —or sequences—are bounded. That is, they are limited as $n \rightarrow \infty$. In this context, Definition 4.4.1 provides a specific setting for sequences bounded away from zero. If sequences are divergent, this definition describes how fast the sequences tend to infinity. Now, we apply this definition to the spatial weight matrices:

Assumption 4.4 — Weight Matrix (Lee, 2004). The elements $w_{n,ij}$ of \mathbf{W}_n are at most of order h_n^{-1} , denoted by $O(1/h_n)$, uniformly in all i, j , where the rate sequence h_n can be **bounded** or **divergent**. As a normalization, $w_{n,ii} = 0$ for all i .

Recall that in econometric we are often interested in the asymptotic behavior of variables (see Section 3.1). For example we say that:

$$X_n = O(b_n) \implies \lim_{n \rightarrow \infty} \frac{X_n}{b_n} = -\infty < c < \infty.$$

This implies that X_n is a bounded sequence of rate b_n . Assumption 4.4 states that the elements of \mathbf{W}_n are sequences that might be bounded or divergent at rate h_n . That is, we do not know if $h_n w_{n,ij}$ is bounded or divergent.

Assumption 4.5 — (Lee, 2004). The ratio $h_n/n \rightarrow 0$ as n goes to infinity

Assumptions 4.4 and 4.5 link directly the spatial weight matrix to the sample size n . The intuition tell us that as the sample size n increases, the row sum of the weight matrices will also tend to increase, since one region could have more neighbors (see our discussion in Section 3.7). The rate at which the spatial weights $w_{n,ij}$ increases as n increases can be bounded (limit on the number of neighbors) or can be divergent (not limit in the number of neighbors). Therefore, Assumptions 4.4 and 4.5 are intended to cover weight matrices whose elements are not restricted to be nonnegative and those that might not be row-standardized.

What are the implications of those assumption? These assumptions deal with the row and column sums of the matrix \mathbf{W}_n . In particular, the row and column sums of \mathbf{W}_n before \mathbf{W}_n is row-normalized should not diverge to infinity at a rate equal to or faster than the rate of the sample size n . This condition is slightly different in Kelejian and Prucha (1998, 1999). Their condition states that the row and columns sums of the matrices \mathbf{W} and $(\mathbf{I}_n - \rho \mathbf{W})^{-1}$ before \mathbf{W} is row-normalized should be uniformly bounded in absolute value as n goes to infinity. In both cases these conditions limit the cross-sectional correlation to a manageable degree, i.e., the correlation between two spatial units should converge to zero as the distance separating them increases to infinity.

In addition to the technicality, these assumptions have applied implications. Normally, no spatial unit is assumed to be a neighbor to more than a given number, say q , of other units. Therefore, the number of neighbors is limited and Lee (2004)'s and Kelejian and Prucha (1998, 1999)'s assumption is satisfied.

By contrast, when the spatial weights matrix is an inverse distance matrix Kelejian and Prucha (1998, 1999)'s condition may not be satisfied. To see this, consider an infinite number of spatial units that are arranged linearly. Let the distance of each spatial unit to its first left- and right-hand neighbor be d ; to its second left- and right-hand neighbor, the distance $2d$; and so on. See for example Figure 4.2.

Figure 4.2: Distances from R3 to all Regions

$$\text{R1} \xleftarrow{2d} \text{R2} \xleftarrow{d} \text{R3} \xrightarrow{d} \text{R4} \xrightarrow{2d} \text{R5}$$

When \mathbf{W} is an inverse distance matrix and its off-diagonal elements are of the form $1/d_{ij}$, where d_{ij} is the distance between two spatial units i and j , each row sum is

$$1/d + 1/d + 1/2d + 1/2d + \dots = 2 \times (1/d + 1/2d + 1/3d + \dots)$$

representing a series that is not finite. This is perhaps the main motivation of why some empirical applications introduce a cut-off point d^* such that $w_{ij} = 0$ if $d_{ij} > d^*$. However, since the ratio $2 \times (1/d + 1/2d + 1/3d + \dots)/n \rightarrow 0$ as $n \rightarrow \infty$, Lee (2004)'s condition is satisfied, which implies that an inverse distance matrix without cut-off point does not necessarily have

to be excluded in an empirical study for reasons of consistency. Thus, Assumption 4.5 excludes cases where the row sums, $\sum_{j=1}^n w_{ij}$, for $i = 1, \dots, n$, diverges to infinity at a rate equal to or faster than the rate of the sample size n , because the ML estimator would likely be inconsistent for those cases. Another case where $\{h_n\}$ is a bounded sequence is when we fixed the number of neighbors, such as in the case of k -neighbors approach. Nevertheless our distance example explains why it sometimes leads to numerical problems or unexpected outcomes in empirical applications. This is because the number of unit in the sample generally does not go to infinity, but is finite.

What if h_n is unbounded? Under this case $\sum_{j=1}^n d_{ij}$ is uniformly bounded away from zero at the rate h_n , where $\lim_{n \rightarrow \infty} h_n = \infty$. This particular case **rules out** cases where each unit has only a (fixed) finite number of neighbors even when the total number of unit increases to infinity. For example, it rules out the case where units correspond to counties and neighbors are defined as counties with contiguous border.

In which cases $h_n \rightarrow \infty$? This case requires that each unit in the limit has infinitely many neighbors. As stated by Lee (2002), in economic applications where either the neighbors of any unit are dense in a relevant space or each unit is influenced by many of its neighboring units, which represents a significant proportion of the total population units, it is likely that $\sum_{j=1}^n d_{ij}$ will diverge and $(1/n) \sum_{j=1}^n d_{ij}$ will converge as n becomes large. Consider the case where $d_{ij} = 1/|r_i - r_j|$, where r_i is the proportion of state i 's population that is of African descent. As no state in USA has zero proportion of African-Americans in its population, d_{ij} will be positive, and $(1/n) \sum_{j=1}^n d_{ij}$ will be bounded away from zero and $\sum_{j=1}^n d_{ij}$ will be likely to possess the n rate of divergence in this example.

Another example occurs when all cross-sectional units are assumed to be neighbors of each other and are given equal weights. In that case all off-diagonal elements of the spatial weights matrix are $w_{ij} = 1$. Since the row and column sums are $n - 1$, these sums diverge to infinity as $n \rightarrow \infty$. In contrast to the previous case, however, $(n - 1)/n \rightarrow 1$ instead of 0 as $n \rightarrow \infty$. This implies that a spatial weight matrix that has equal weights and that is row-normalized subsequently, $w_{ij} = 1/(n - 1)$ must be excluded for reasons of consistency since it satisfies neither Lee (2004)'s and Kelejian and Prucha (1998, 1999)'s condition. The alternative is a group interaction matrix, introduced by Case (1991). Here “neighbors” refer to farmers who live in the same district. Suppose that there are R districts and there are m farmers in each district. The sample size is $n = mR$. Case assumed that in a district, each neighbor of a farmer is given equal weight. In that case, $\mathbf{W}_n = \mathbf{I}_R \otimes \mathbf{B}_m$, where $\mathbf{B}_m = (\mathbf{z}_m \mathbf{z}_m^\top - \mathbf{I}_m)/(m - 1)$. For this example, $h_n = (m - 1)$ and $h_n/n = (m - 1)/(mR) = O(1/R)$. If sample size n increases by increasing both R and m , then h_n goes to infinity and h_n/n goes to zero as n tends to infinity. Thus, this matrix satisfies Lee (2004)'s condition.

R Whether $\{h_n\}$ is a bounded or divergent sequence has interesting implications on the OLS estimation. The OLS estimators of β and ρ are inconsistent when $\{h_n\}$ is bounded, but they can be consistent when $\{h_n\}$ is divergent (see Lee, 2002).

In summary, when $\{h_n\}$ is a bounded sequence, it implies a cross sectional unit has only a small number of neighbors, where the spatial dependence is usually defined based on geographical implications. When $\{h_n\}$ is divergent, it corresponds to the scenario where each unit has a large number of neighbors that often emerges in empirical studies of social interactions or cluster sampling data.

Assumption 4.6 — Non-singularity of A_n (Lee, 2004). The matrix A_n is nonsingular.

Under Assumption 4.6, the SLM (system) has the reduced form (equilibrium) given by

$$\mathbf{y} = \mathbf{A}_{n0}^{-1}(\mathbf{X}_n\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}_n) \quad (4.51)$$

where $\mathbf{A}_{n0}^{-1} = (\mathbf{I}_n - \rho_0\mathbf{W}_n)$ and with the following expectation and variance:

$$\mathbb{E}(\mathbf{y}_n) = (\mathbf{I}_n - \rho_0\mathbf{W}_n)^{-1} \mathbf{X}_n\boldsymbol{\beta} = \mathbf{A}_{n0}^{-1} \mathbf{X}_n\boldsymbol{\beta}_0 \quad (4.52)$$

$$\mathbb{V}(\mathbf{y}_n) = \mathbb{E}(\mathbf{y}_n\mathbf{y}_n^\top) = \sigma_0^2 (\mathbf{I}_n - \rho_0\mathbf{W}_n)^{-1} [(\mathbf{I}_n - \rho_0\mathbf{W}_n)^{-1}]^\top = \sigma_0^2 \mathbf{A}_{n0}^{-1} (\mathbf{A}_{n0}^{-1})^\top \quad (4.53)$$

We can also write the reduced-form equation as follows:

$$\begin{aligned} \mathbf{y}_n &= \mathbf{X}_n\boldsymbol{\beta}_0 + \rho_0\mathbf{W}_n\mathbf{y}_n + \boldsymbol{\varepsilon}_n \\ &= \mathbf{X}_n\boldsymbol{\beta}_0 + \rho_0\mathbf{W}_n [\mathbf{A}_n^{-1} \mathbf{X}_n\boldsymbol{\beta}_0 + \mathbf{A}_n^{-1} \boldsymbol{\varepsilon}_n] + \boldsymbol{\varepsilon}_n \\ &= \mathbf{X}_n\boldsymbol{\beta}_0 + \rho_0\mathbf{W}_n\mathbf{A}_n^{-1} \mathbf{X}_n\boldsymbol{\beta}_0 + \rho_0\mathbf{W}_n\mathbf{A}_n^{-1} \boldsymbol{\varepsilon}_n + \boldsymbol{\varepsilon}_n \\ &= \mathbf{X}_n\boldsymbol{\beta}_0 + \rho_0\mathbf{W}_n\mathbf{A}_n^{-1} \mathbf{X}_n\boldsymbol{\beta}_0 + (\mathbf{I}_n + \rho_0\mathbf{W}_n\mathbf{A}_n^{-1}) \boldsymbol{\varepsilon}_n \\ &= \mathbf{X}_n\boldsymbol{\beta}_0 + \rho_0\mathbf{C}_n\mathbf{X}_n\boldsymbol{\beta}_0 + (\mathbf{I}_n + \rho_0\mathbf{C}_n) \boldsymbol{\varepsilon}_n \\ &= \mathbf{X}_n\boldsymbol{\beta}_0 + \rho_0\mathbf{C}_n\mathbf{X}_n\boldsymbol{\beta}_0 + \mathbf{A}_n^{-1} \boldsymbol{\varepsilon}_n \end{aligned} \quad (4.54)$$

because $\mathbf{I}_n + \rho_0\mathbf{C}_n = \mathbf{A}_n^{-1}$ (see Exercise 4.6), where $\mathbf{C}_n = \mathbf{W}_n\mathbf{A}_n^{-1}$. This expression will also be useful later.

Assumption 4.7 — Uniform boundedness (Lee, 2004). The sequences of matrices $\{\mathbf{W}_n\}$ and $\{\mathbf{A}_n^{-1}\}$ are uniformly bounded in both row and column sums

The uniform boundedness of the matrices is a condition to limit the spatial correlation to a manageable degree. For example, it guarantees that the variances of \mathbf{y}_n are bounded as n goes to infinity. See our discussion in Section 3.10.

Why do we care about this? Because we need the variance goes to zero when the sample size goes to infinity in order to apply some consistency theorem.⁴

Lemma 4.8 — Uniform Boundedness of Matrices in Row and Column Sums. Suppose that the spatial weights matrix \mathbf{W}_n is a non-negative matrix with its (i, j) th element being

$$w_{n,ij} = \frac{d_{ij}}{\sum_{l=1}^n d_{il}}$$

and $d_{ij} > 0$ for all i, j .

- (a) If the row sums $\sum_{j=1}^n d_{ij}$ are bounded away from zero at the rate h_n uniformly in i , and the column sums $\sum_{i=1}^n d_{ij}$ are $O(h_n)$ uniformly in j , then $\{\mathbf{W}_n\}$ are uniformly bounded in column sums.
- (b) (Symmetric Matrix) If $d_{ij} = d_{ji}$ for all i and j and the row sums $\sum_{j=1}^n d_{ij}$ are $O(h_n)$

⁴Equivalently, this assumption rules out the unit root case in time series.

and bounded away from zero at the rate h_n uniformly in i , then $\{\mathbf{W}_n\}$ are uniformly bounded in column sums.

Assumption 4.9 — No asymptotic multicollinearity (Lee, 2004). The elements of \mathbf{X}_n are uniformly bounded constants for all n . The $\lim_{n \rightarrow \infty} \mathbf{X}_n^\top \mathbf{X}_n / n$ exists and is nonsingular.

This rules out multicollinearity among the regressors. Note also that we are assuming that \mathbf{X}_n is **nonstochastic**. If \mathbf{X}_n were stochastic, then we will require:

$$\text{plim}_{n \rightarrow \infty} \mathbf{X}_n^\top \mathbf{X}_n / n,$$

to exists.

Assumption 4.10 — Uniform Boundedness of $\mathbf{A}_n^{-1}(\rho)$ Lee (2004). $\mathbf{A}_n^{-1}(\rho)$ are uniformly bounded in either row or column sums, uniformly in ρ in a compact parameter space Γ . The true parameter ρ_0 is in the interior of Γ

This assumption is needed to deal with the nonlinearity of $\log |(\mathbf{I}_n - \rho \mathbf{W})^{-1}|$ in the log-likelihood function. Recall that if $\|\mathbf{W}\| < 1$, then $\mathbf{I}_n - \rho \mathbf{W}_n$ is invertible for all n . Then if $\|\mathbf{W}\| < 1$, then the sequence of matrices $\|(\mathbf{I}_n - \mathbf{W}_n)^{-1}\|$ are uniformly bounded in any subset of $(-1, 1)$ bounded away from the boundary. As we previously see, if \mathbf{W}_n is row-standardized $(\mathbf{I}_n - \mathbf{W})^{-1}$ is uniformly bounded in row sums norm uniformly in any closed subset of $(-1, 1)$. Therefore, Γ from Assumption 4.10 can be considered as a single closed set contained in $(-1, 1)$.

What if \mathbf{W}_n is not row-normalized but its eigenvalues are real? Then, the Jacobian of $|(\mathbf{I}_n - \mathbf{W})^{-1}|$ will be positive if $-1/\omega_{\min} < \rho < 1/\omega_{\max}$, where ω_{\min} and ω_{\max} are the minimum and maximum eigenvalues of \mathbf{W}_n , and Γ will be a closed interval contained in $(-1/\omega_{\min}, 1/\omega_{\max})$ for all n . Thus, Assumption 4.10 rules out models where ρ_0 is close to -1 and 1.

Assumption 4.11 — Identification (Lee, 2004). The

$$\lim_{n \rightarrow \infty} \frac{1}{n} (\mathbf{X}_n, \mathbf{C}_n \mathbf{X}_n \beta_0)' (\mathbf{X}_n, \mathbf{C}_n \mathbf{X}_n \beta_0)$$

exists and is nonsingular.

This is a sufficient condition for global identification of θ_0

Theorem 4.12 — Consistency. Let $\theta_0 = (\beta_0^\top, \rho_0, \sigma_0^2)^\top$. Under assumption 4.3-4.11, θ_0 is globally identifiable and $\hat{\theta}_n$ is a consistent estimator of θ_0 .

Identification of ρ_0 can be based on the maximum values of the concentrated log-likelihood function $Q_n(\rho)/n$. With identification and uniform convergence of $[\log L_n(\rho) - Q_n(\rho)]/n$ to zero on Γ , consistency of the QMLE $\hat{\theta}_n$ follows. The sketch of the proof for Theorem 4.12 is given in Appendix 4.A.

For a proof without compactness of the parameter space (proving concavity of the log-likelihood function) see Liu et al. (2022).

4.4.2 Asymptotic Normality

To derive the asymptotic distribution of the QML and ML we need the asymptotic behavior of the gradient. Taking a Taylor series expansion around $\boldsymbol{\theta}_0$ of $\partial \ell_n(\hat{\boldsymbol{\theta}}_n)/\partial \boldsymbol{\theta} = 0$ at $\boldsymbol{\theta}_0$, we get

$$\frac{\partial \ell_n(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}} = \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + \frac{\partial^2 \ell_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0), \quad (4.55)$$

where $\tilde{\boldsymbol{\theta}}_n = \alpha_n \hat{\boldsymbol{\theta}}_n + (1 - \alpha_n) \boldsymbol{\theta}_0$ and $\alpha_n \in [0, 1]$, therefore:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = - \left[\frac{1}{n} \frac{\partial^2 \ell_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right]^{-1} \frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}. \quad (4.56)$$

As standard in asymptotic theory of MLE, we need to show that the first element of the rhs of (4.56) converges to something. We also need to find the limiting distribution of $\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}$. Recall that the first-order derivatives of the log-likelihood function **evaluated at** $\boldsymbol{\theta}_0$ are given by (see Section 4.2.2):

$$\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma_0^2 \sqrt{n}} \mathbf{X}_n^\top \boldsymbol{\varepsilon}_n \quad (4.57)$$

$$\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \sigma^2} = \frac{1}{2\sigma_0^4 \sqrt{n}} (\boldsymbol{\varepsilon}_n' \boldsymbol{\varepsilon}_n - n\sigma_0^2) \quad (4.58)$$

$$\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \rho} = \frac{1}{\sigma_0^2 \sqrt{n}} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \boldsymbol{\varepsilon}_n + \frac{1}{\sigma_0^2 \sqrt{n}} (\boldsymbol{\varepsilon}_n^\top \mathbf{C}_{n0} \boldsymbol{\varepsilon}_n - \sigma_0^2 \text{tr}(\mathbf{C}_{n0})) \quad (4.59)$$

As explained by Lee (2004, pag. 1905), these are linear and quadratic functions of $\boldsymbol{\varepsilon}_n$. In particular, the asymptotic distribution of (4.59) may be derived from central limit theorem for linear-quadratic forms. The matrix \mathbf{C}_{n0} is uniformly bounded in row sums. As the elements of \mathbf{X}_n are bounded, the elements of $\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0$ for all n are uniformly bounded by Lemma 3.22. With the existence of high order moments of ϵ in Assumption 4.3, the central limit theorem for quadratic forms of double arrays of Kelejian and Prucha (2001) can be applied and the limit distribution of the score vector follows.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \rho, \sigma^2)^\top$ be the $k + 2$ -dimensional vector. Since $\mathbb{E}[(1/\sqrt{n})\partial \ell_n(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}] = \mathbf{0}$, the variance matrix of $(1/\sqrt{n})\partial \ell_n(\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}$ is:

$$\mathbb{E} \left[\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \cdot \frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^\top} \right] = -\mathbb{E} \left(\frac{1}{n} \frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right) + \boldsymbol{\Omega}_{\boldsymbol{\theta},n}, \quad (4.60)$$

where

$$-\mathbb{E} \left(\frac{1}{n} \frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right) = \begin{pmatrix} \frac{1}{n\sigma_0^2} (\mathbf{X}_n^\top \mathbf{X}_n) & \frac{1}{n\sigma_0^2} \mathbf{X}_n^\top (\mathbf{C}_n \mathbf{X}_n \boldsymbol{\beta}_0) & \mathbf{0}^\top \\ \frac{1}{n} \text{tr}(\mathbf{C}_{n0}^s \mathbf{C}_{n0}) + \frac{1}{n\sigma_0^2} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0) & \frac{1}{n\sigma_0^2} \text{tr}(\mathbf{C}_{n0}) & \frac{1}{2\sigma_0^4} \end{pmatrix} \quad (4.61)$$

and $\mathbf{C}_{n0}^s = \mathbf{C}_{n0} + \mathbf{C}_{n0}^\top$. Equation (4.61) represents the average Hessian matrix (or information matrix when $\boldsymbol{\varepsilon}$'s are **normal**). The matrix $\boldsymbol{\Omega}_{\boldsymbol{\theta},n}$ is a matrix with the second, third, and fourth moments of $\boldsymbol{\varepsilon}$. If $\boldsymbol{\varepsilon}_n$ is normally distributed, then $\boldsymbol{\Omega}_{\boldsymbol{\theta},n} = \mathbf{0}$.

Derivation of (4.61) is given in Appendix 4.B and the variance of the score function is given in Appendix 4.C.

Theorem 4.13 — Asymptotic Normality. Under Assumptions 4.3-4.11,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(\theta, \Sigma_\theta^{-1} + \Sigma_\theta^{-1} \Omega_\theta \Sigma_\theta^{-1}), \quad (4.62)$$

where $\Omega_\theta = \lim_{n \rightarrow \infty} \Omega_{\theta,n}$ and

$$\Sigma_\theta = - \lim_{n \rightarrow \infty} \mathbb{E} \left[\frac{1}{n} \frac{\partial^2 \log L_n(\theta_0)}{\partial \theta \partial \theta^\top} \right], \quad (4.63)$$

which are assumed to exist. If the ϵ_i 's are **normally distributed**, then:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(\theta, \Sigma_\theta^{-1}). \quad (4.64)$$

A sketch of the proof of Theorem 4.13 is given in Appendix 4.D.

4.5 Computing the Standard Errors For The Marginal Effects

In section 2.3.2, we explain how to obtain summary measures for the direct, indirect and total effects. However, we did not explain how to obtain standard errors for such measures. For example, we would like to have confidence intervals for the indirect effects and to be able to say whether they are significant.

Recall that our three summary measures are:

$$\begin{aligned} \bar{M}(\theta)_{\text{direct}} &= n^{-1} \text{tr}(\mathbf{S}_r(\theta)) \\ \bar{M}(\theta)_{\text{total}} &= n^{-1} \mathbf{1}_n^\top \mathbf{S}_r(\theta) \mathbf{1}_n \\ \bar{M}(\theta)_{\text{indirect}} &= \bar{M}(r)_{\text{total}} - \bar{M}(r)_{\text{direct}}, \end{aligned}$$

which are highly nonlinear due to $\mathbf{S}_r(\theta)$.⁵ Therefore, a procedure such as the Delta Method is not feasible. Instead, we use a Monte Carlo approximation which takes into account the sampling distribution of θ . To show this procedure, consider the SDM where:

$$\mathbf{S}(\theta)_r = (\mathbf{I}_n - \rho \mathbf{W})^{-1} (\mathbf{I}_n \beta_r + \mathbf{W} \gamma_r)$$

Let $g(\theta) = \bar{M}(\theta)$ be a function representing the marginal (direct, indirect or total) effect that depends on the population parameters θ . If $N(\theta|\bar{\theta}, \Sigma_\theta)$ denotes the multivariate normal density of θ with mean $\bar{\theta}$ and asymptotic variance-covariance matrix Σ_θ , then the expected value of the marginal effects conditional on the population parameters $\bar{\theta}$ and Σ_θ is:

$$\mathbb{E}(g(\theta)|\bar{\theta}, \Sigma_\theta) = \int_{\theta} \mathbb{E}(g(\theta)|\mathbf{y}, \mathbf{X}, \theta) N(\theta|\bar{\theta}, \Sigma_\theta) d\theta. \quad (4.65)$$

A Monte Carlo approximation to this expectation is obtained by calculation of the empirical marginal effects evaluated at pseudo draws of θ from the asymptotic distribution of the estimator. The algorithm is the following:

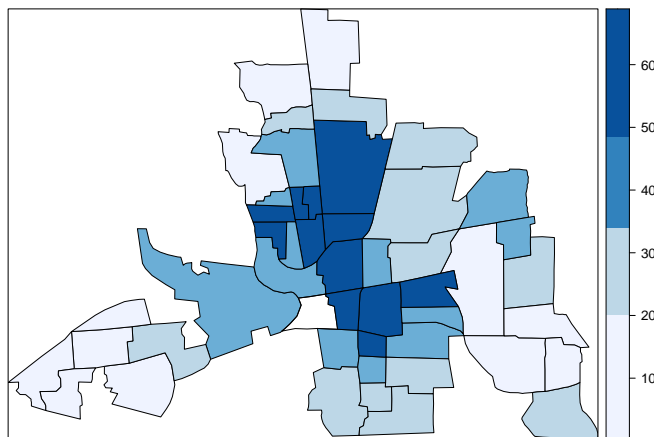
⁵Note that we have replaced the parameter for the spatially lagged independent variable to let θ be the vector parameters of the model.

As usual in applied work, we start the analysis by asking whether there exists a spatial pattern in the variable we are interested in. To get some insights about the spatial distribution of `CRIME` we use the following quantile choropleth graph:

```
# Spatial distribution of crime
spplot(columbus, "CRIME",
       at = quantile(columbus$CRIME, p = c(0, .25, .5, .75, 1), na.rm = TRUE),
       col.regions = brewer.pal(5, "Blues"),
       main = "")
```

Figure 4.3 shows the spatial pattern of crime. It can be observed that the spatial distribution of crime follows a clear pattern of positive autocorrelation. However, we must corroborate this statement by using a global test of spatial autocorrelation. To do so, we use a row-normalized binary contiguity matrix \mathbf{W} , `col.gal.nb`, based on the queen criteria and carry out a Moran's I test. In particular, we use a Moran test based on Monte Carlo simulations using the `moran.mc` function with 99 simulations.

Figure 4.3: Spatial Distribution of Crime in Columbus, Ohio Neighborhoods



Notes: This graph shows the spatial distribution of crime on the 49 Columbus, Ohio neighborhoods. Darker color indicates greater rate of crime.

```
# Moran's I test
set.seed(1234)
listw <- nb2listw(col.gal.nb, style = "W")
moran.mc(columbus$CRIME, listw = listw,
         nsim = 99, alternative = 'greater')
##
```



```
## Monte-Carlo simulation of Moran I
##
## data:  columbus$CRIME
## weights: listw
## number of simulations + 1: 100
##
## statistic = 0.48577, observed rank = 100, p-value = 0.01
## alternative hypothesis: greater
```

The results show that the Moran's I statistic is 0.51 and the p-value is 0.01. This implies that we reject the null hypothesis of random spatial distribution and there exists evidence of positive global spatial autocorrelation in the crime variable: places with high (low) crime rate are surrounded by places with high (low) crime rate.

Our next step is to estimate different spatial models using the functions already programmed in **spatialreg**. First, we estimate the classical OLS model followed by the SLX, SLM, SDM, SEM and SAC models. The functions used for each models are the following:

- OLS: `lm` function.
- SLX: `lm` function, where $\mathbf{W}\mathbf{X}$ is constructed using the function `lag.listw` from **spdep** package. This model can also be estimated using the function `lmSLX` from **spatialreg** package as shown below.
- SLM: `lagsarlm` from **spatialreg** package.
- SDM: `lagsarlm` from **spatialreg** package, using the argument `type = "mixed"`. Note that `type = "Durbin"` may be used instead of `type = "mixed"`.
- SEM: `errorsarlm` from **spatialreg** package. Note that the Spatial Durbin Error Model (SDEM)—not shown here— can be estimated by using `type = "emixed"`.
- SAC: `sacsarlm` from **spatialreg** package.

All models are estimated using ML procedure outline in the previous section. In order to compute the determinant of the Jacobian we use the [Ord \(1975\)](#)'s procedure by explicitly using the argument `method = "eigen"` in each spatial model. That is, the Jacobian is computed as in [\(4.41\)](#).

```
# Models
columbus$lag.INC  <- lag.listw(listw,
                             columbus$INC)  # Create spatial lag of INC
columbus$lag.HOVAL <- lag.listw(listw,
                             columbus$HOVAL) # Create spatial lag of HOVAL
ols <- lm(CRIME ~ INC + HOVAL,
          data = columbus)
slx <- lm(CRIME ~ INC + HOVAL + lag.INC + lag.HOVAL,
          data = columbus)
slm <- lagsarlm(CRIME ~ INC + HOVAL,
```

```

      data = columbus,
      listw,
      method = "eigen")
sdm <- lagsarlm(CRIME ~ INC + HOVAL,
      data = columbus,
      listw,
      method = "eigen",
      type = "mixed")
sem <- errorsarlm(CRIME ~ INC + HOVAL,
      data = columbus,
      listw,
      method = "eigen")
sac <- sacsarlml(CRIME ~ INC + HOVAL,
      data = columbus,
      listw,
      method = "eigen")

```

Note that the SLX model can also be estimated as follows:

```

slx2 <- lmSLX(CRIME ~ INC + HOVAL,
      data = columbus,
      listw)
summary(slx2)

```

The models are presented in Table 4.1. The OLS estimates are presented in the first column. The results show that an increase of one thousand dollars in the income of the neighborhood is correlated, in average, with a decreased of 1.6 crimes per thousand households. Similarly, an increase of one thousand dollars in the housing value of the neighborhood is correlated, on average, with a decreased of 0.3 crimes per thousand households. Both correlations are statistically significant.⁶ Both results implies that crimes (residential burglaries and vehicle thefts) are lower in richer neighborhoods.

Column 2 of Table 4.1 show the results for the SLX. In particular, the model is given by $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$, where $\mathbf{W}\mathbf{X}$ is a 49×2 matrix, whose columns correspond to the spatial lag of *INC* and *HOVAL*. The coefficient for the spatial lag of *INC*, $\mathbf{W}.\mathbf{INC}$, is negative and significant. This implies that crime in spatial unit i is correlated with the income in its neighborhood: the higher the income of the neighbors of i the lower the crime in i . This result does not, however, hold for the housing value of the neighbors of i which is positive but not statistically different from zero.

The results for the SLM are shown in column 3. The spatial autoregressive parameter ρ is positive and significant indicating strong spatial autocorrelation. This implies evidence of spillover effects on crime. The coefficients for the other variables in the regression are similar to the OLS results, though smaller in absolute value.

The results for the SDM are presented in column 4. Whereas the estimated ρ parameter is positive and significant, the coefficient of the lagged explanatory variables are not. This

⁶Note that we refer to correlation since there may still be some sort of endogeneity problem in either of the two variables.

indicates that once we have take into account the endogenous interaction effects of crime, the neighbors' factors do not matter in explaining the crime in each location. Moreover, for the spatial lag of income, the wrong sign is obtained, since the common factor hypothesis would imply a positive sign, given a positive estimate for ρ and negative sign for INC. This provide some evidence that an omitted spatial lag may be the main spatial effect, rather than spatial dependence in the error term.

Column 5 shows the results for the the SEM model which confirm the conclusions from the previous models. It can be noticed that the autoregressive parameter for $\mathbf{W}\mathbf{u}$ is positive and significant indicating an important spatial transmission of the random shocks. This result may be explained by the fact of omitting important variables that are spatially correlated.

The SAC model, presented in column 6, considers both endogenous interactions effects and interactions effects among the error terms. From the results, we observe that the SAC model produces coefficients estimates of $\mathbf{W}\mathbf{y}$ and $\mathbf{W}\mathbf{u}$ variables that are not significantly different from zero. However, if endogenous interaction effects and interactions effects among the error terms are separated from each other, both coefficients turn out to be significant. This might be explained by the fact that the model is overparametrized, as a result of which the significance levels of all variables tend to go down.

Table 4.1: Spatial Models for Crime in Columbus, Ohio Neighborhoods.

	OLS	SLX	SLM	SDM	SEM	SAC
<i>Constant</i>	68.619*** (4.735)	74.029*** (6.722)	46.851*** (7.315)	45.593*** (13.129)	61.054*** (5.315)	49.051*** (10.055)
INC	-1.597*** (0.334)	-1.108** (0.375)	-1.074*** (0.311)	-0.939** (0.338)	-0.995** (0.337)	-1.069** (0.333)
HOVAL	-0.274* (0.103)	-0.295** (0.101)	-0.270** (0.090)	-0.300*** (0.091)	-0.308*** (0.093)	-0.283** (0.092)
<i>W.INC</i>		-1.383* (0.559)		-0.618 (0.577)		
<i>W.HOVAL</i>		0.226 (0.203)		0.267 (0.184)		
ρ			0.404*** (0.121)	0.383* (0.162)		0.353 (0.197)
λ					0.521*** (0.141)	0.132 (0.299)
AIC	382.754	380.197	376.337	378.032	378.310	378.146
N	49	49	49	49	49	49

Significance: *** $\equiv p < 0.001$; ** $\equiv p < 0.01$; * $\equiv p < 0.05$

4.6.2 Estimation of Marginal Effects in R

In this Section we expand our analysis from Section 2.5 in the sense that we now integrate the estimation of the marginal effects using a real estimation from R.

We begin our analysis with the following question: what would happen to crime in all regions if income rose from 13.906 to 14.906 in the 30th region ($\Delta\text{INC} = 1$)? Note that we

tried to answer a similar question in the commuting-time example from previous chapter. As we did in Section 2.5 we can use the reduced-form predictor given by the following formula:

$$\hat{\mathbf{y}} = \mathbb{E}(\mathbf{y}|\mathbf{X}, \mathbf{W}) = (\mathbf{I}_n - \hat{\rho}\mathbf{W})^{-1}\mathbf{X}\hat{\beta},$$

and estimate the predicted values pre- and post- the change in the income variable. In the following lines we use the reduced-form predictor and the observed values of the exogenous variables to obtain the predicted values for CRIME, $\hat{\mathbf{y}}^1$, using the SLM model previously estimated.

```
# The predicted values
rho      <- slm$rho                      # Estimated rho from SLM model
beta_hat <- coef(slm)[-1]                 # Estimated parameters
A        <- invIrW(listw, rho = rho)      # (I - rho*W)^{-1}
X        <- cbind(1, columbus$INC, columbus$HOVAL) # Matrix of observed variables
y_hat_pre <- A %*% crossprod(t(X), beta_hat) # y hat
```

Next we increase INC by 1 in spatial unit 30, and calculate the reduced-form predictions, $\hat{\mathbf{y}}^2$.

```
# The post-predicted values
col_new <- columbus # copy the data frame

# Change the income value
col_new@data[col_new@data$POLYID == 30, "INC"] <- 14.906

# The predicted values
X_d      <- cbind(1, col_new$INC, col_new$HOVAL)
y_hat_post <- A %*% crossprod(t(X_d), beta_hat)
```

Finally, we compute the difference between pre- and post-predictions: $\hat{\mathbf{y}}^2 - \hat{\mathbf{y}}^1$:

```
# The difference
delta_y      <- y_hat_post - y_hat_pre
col_new$delta_y <- delta_y

# Show the effects
summary(delta_y)

##           V1
##  Min.      :-1.1141241
##  1st Qu.   :-0.0074114
##  Median   :-0.0012172
##  Mean      :-0.0336341
##  3rd Qu.   :-0.0002604
##  Max.      :-0.0000081

sum(delta_y)

## [1] -1.648071
```

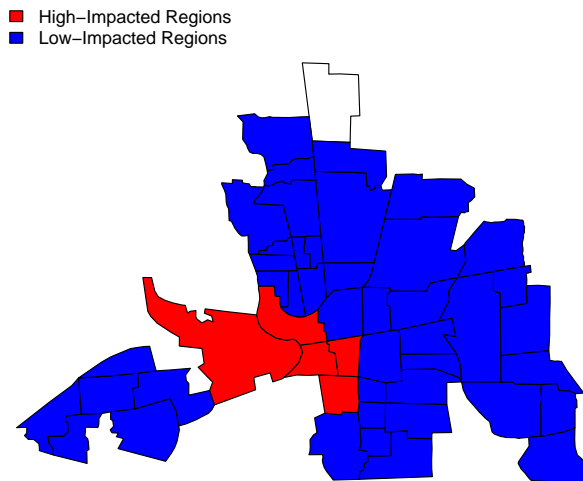
According to the result from `sum(delta_y)`, the predicted effect of the change would be a decrease of 1.65 in the crime rate, considering both direct and indirect effects. That is, increasing the income in US\$1,000 in region 30th might generate effects that will transmit through the whole system of region resulting in a new equilibrium where the the total crime will reduce in 1.7 crimes per thousand households.

Sometimes we would like to plot these effects. Suppose we wanted to show those regions that had low and high impact due to the increase in `INC`. Let's define "high impacted regions" those regions whose crime rate decrease more than 0.05. The following code produces Figure 4.4.

```
# Breaks
breaks <- c(min(col_new$delta_y), -0.05, max(col_new$delta_y))
labels <- c("High-Impacted Regions", "Low-Impacted Regions")
np <- findInterval(col_new$delta_y, breaks)
colors <- c("red", "blue")

# Draw Map
plot(col_new, col = colors[np])
legend("topleft", legend = labels, fill = colors, bty = "n")
points(38.29, 30.35, pch = 19, col = "black", cex = 0.5)
```

Figure 4.4: Effects of a Change in Region 30: Categorization

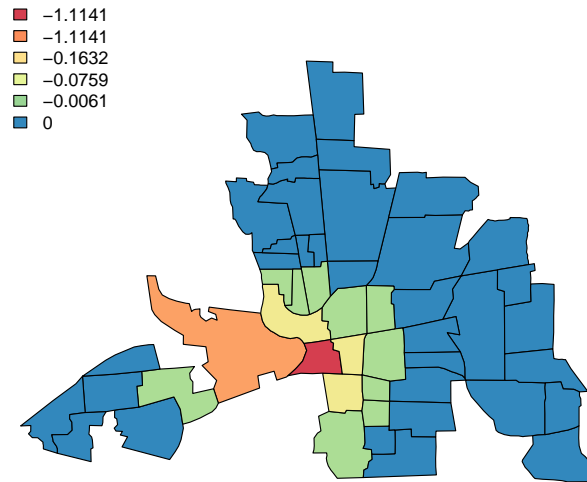


Notes: This graph shows those regions that had low and high impact due to increase in `INC` in 30th. Red-colored regions are those regions with a decrease of crime rate larger than 0.05, whereas blue-colored regions are those regions with lower decrease of crime rate.

Now we map the magnitude of the changes caused by altering `INC` in region 30. The code is the following and the graph is presented in Figure 4.5.

```
# Plot the magnitude of the ME
pal5    <- brewer.pal(6, "Spectral")
cats5    <- classIntervals(col_new$delta_y, n = 5, style = "jenks")
colors5 <- findColours(cats5, pal5)
plot(col_new, col = colors5)
legend("topleft", legend = round(cats5$brks, 2), fill = pal5, bty = "n")
```

Figure 4.5: Effects of a Change in Region 30: Magnitude



Notes: This graph shows the spatial distribution of the changes caused by altering `INC` in region 30.

In the rest of this Section we use the `impacts()` function from **spatialreg** package to understand the direct (local), indirect(spillover), and total effect of a unit change in each of the predictor variables. This function returns the direct, indirect and total impacts for the variables in the model. The spatial lag impact measures are computed using the reduced form:

$$\mathbf{y} = \sum_{r=1}^K \mathbf{A}(\mathbf{W})^{-1} (\mathbf{I}_n \beta_r) + \mathbf{A}(\mathbf{W})^{-1} \boldsymbol{\epsilon} \quad (4.66)$$

$$\mathbf{A}(\mathbf{W})^{-1} = \mathbf{I}_n + \rho \mathbf{W} + \rho^2 \mathbf{W}^2 + \dots$$

The exact $\mathbf{A}(\mathbf{W})^{-1}$ is computed when `listw` is given. When the traces are created by powering sparse matrices the approximation $\mathbf{I}_n + \rho \mathbf{W} + \rho^2 \mathbf{W}^2 + \dots$ is used. The exact and

the trace methods should give very similar results, unless the number of powers used is very small, or the spatial coefficient is close to its bounds.

```
impacts(slm, listw = listw)

## Impact measures (lag, exact):
##           Direct   Indirect   Total
## INC    -1.1225156 -0.6783818 -1.8008973
## HOVAL  -0.2823163 -0.1706152 -0.4529315
```

The output says that an increase of US\$1,000 in income leads to a decrease of 1.8 crimes per thousand households.

The direct effect of the income variable in the SLM model amounts to -1.123, while the coefficient estimate of this variable is -1.074. This implies that the feedback effect is $-1.123 - (-1.074) = -0.049$. This feedback effect corresponds to 4.5% of the coefficient estimate.

Let's corroborate these results by computing the impacts using matrix operations:

```
## Construct  $S_r(W) = A(W)^{-1} (I * \beta_r + W * \theta_r)$ 
Ibeta <- diag(length(listw$neighbours)) * coef(slm)["INC"]
S <- A %*% Ibeta

ADI <- sum(diag(S)) / nrow(A)
ADI

## [1] -1.122516

n <- length(listw$neighbours)
Total <- crossprod(rep(1, n), S) %*% rep(1, n) / n
Total

##           [,1]
## [1,] -1.800897

Indirect <- Total - ADI
Indirect

##           [,1]
## [1,] -0.6783818
```

Note that the results are the same as those computed by `impact`.

We can also obtain the p-values of the impacts by using the argument `R`. This argument indicates the number of simulations use to create distributions for the impact measures, provided that the fitted model object contains a coefficient covariance matrix.

Now with p-values:

```
# Compute standard errors of impacts
im_obj <- impacts(slm, listw = listw, R = 200)
summary(im_obj, zstats = TRUE, short = TRUE)
```

```
## Impact measures (lag, exact):
##           Direct   Indirect   Total
## INC    -1.1225156 -0.6783818 -1.8008973
## HOVAL  -0.2823163 -0.1706152 -0.4529315
## =====
## Simulation results ( variance matrix):
## =====
## Simulated standard errors
##           Direct   Indirect   Total
## INC    0.28508990 0.3418742 0.5254074
## HOVAL  0.09464442 0.1093738 0.1772125
##
## Simulated z-values:
##           Direct   Indirect   Total
## INC    -3.857740 -2.090573 -3.453541
## HOVAL  -3.114138 -1.761298 -2.750232
##
## Simulated p-values:
##           Direct   Indirect Total
## INC    0.00011444 0.036566 0.00055328
## HOVAL  0.00184483 0.078188 0.00595532
```

The results shows that the variable that exerts the largest negative direct impact is **INC**. That is, **INC** exert the largest reduction on own-crime rate. The indirect effects are presented in the second column. These effects help identify which variables produce the largest spatial spillovers. Negative effects could be considered spatial benefits, since these indicate variables that lead to a reduction in crime rate. Positive indirect effects would represent a negative externality, since this indicates that neighboring regions suffer from an increase in crime rate when these variables increase. From the results we observe that **INC** has the largest and significant negative indirect effects.

The indirect effect for **HOVAL** is not significant. The weakly significant effect in the SLM model can be explained by the fact that this model suffers from the problem that the ratio between the spillover effect and the direct effect is the same for every explanatory variable. Therefore, this model is too rigid to model spillover effects adequately.

Total effect takes into account both the direct and indirect effects, allowing us to draw an inference regarding what variables are important to reduce crime rate. We can observe that **INC** has the larges total effect.

Now we follow the example that converts the spatial weight matrix into “sparse” matrix, and power it up using the `trW` function.

```
# Impacts using traces.
W <- as(nb2listw(col.gal.nb, style = "W"), "CsparseMatrix")
trMC <- trW(W, type = "MC")
im <- impacts(slm, tr = trMC, R = 100)
summary(im, zstats = TRUE, short = TRUE)

## Impact measures (lag, trace):
```



```
##           Direct   Indirect     Total
## INC      -1.1220237 -0.6788736 -1.8008973
## HOVAL    -0.2821926 -0.1707389 -0.4529315
## =====
## Simulation results ( variance matrix):
## =====
## Simulated standard errors
##           Direct   Indirect     Total
## INC      0.30303387 0.2985437 0.4706174
## HOVAL    0.09635802 0.1161956 0.1879878
##
## Simulated z-values:
##           Direct   Indirect     Total
## INC      -3.821767 -2.198798 -3.855705
## HOVAL    -2.937362 -1.457964 -2.406792
##
## Simulated p-values:
##           Direct   Indirect Total
## INC      0.0001325 0.027892 0.0001154
## HOVAL    0.0033102 0.144850 0.0160933
```

We can also observe the cumulative impacts using the argument `Q`. When `Q` and `tr` are given in the `impacts` function the output will present the impact components for each step in the traces of powers of the weight matrix up to and including the Q th power.

```
# Cumulative impacts
im2  <- impacts(slm, tr = trMC, R = 100, Q = 5)
sums2 <- summary(im2, zstats = TRUE, reportQ = TRUE, short = TRUE)
sums2

## Impact measures (lag, trace):
##           Direct   Indirect     Total
## INC      -1.1220237 -0.6788736 -1.8008973
## HOVAL    -0.2821926 -0.1707389 -0.4529315
## =====
## Impact components
## $direct
##           INC           HOVAL
## Q1 -1.073533465 -0.2699971236
## Q2  0.000000000  0.0000000000
## Q3 -0.038985415 -0.0098049573
## Q4 -0.005035472 -0.0012664374
## Q5 -0.003072085 -0.0007726393
##
## $indirect
##           INC           HOVAL
## Q1  0.000000000  0.000000000
```

```

## Q2 -0.43358910 -0.109049054
## Q3 -0.13613675 -0.034238831
## Q4 -0.06569456 -0.016522394
## Q5 -0.02549505 -0.006412086
##
## $total
##          INC          HOVAL
## Q1 -1.07353347 -0.269997124
## Q2 -0.43358910 -0.109049054
## Q3 -0.17512216 -0.044043788
## Q4 -0.07073004 -0.017788832
## Q5 -0.02856713 -0.007184726
##
## =====
## Simulation results ( variance matrix):
## =====
## Simulated standard errors
##          Direct Indirect      Total
## INC    0.34631256 0.4029543 0.6401131
## HOVAL 0.08921795 0.1241116 0.1807874
##
## Simulated z-values:
##          Direct Indirect      Total
## INC    -3.233305 -1.853701 -2.916189
## HOVAL  -3.239503 -1.585160 -2.686904
##
## Simulated p-values:
##          Direct Indirect Total
## INC    0.0012237 0.063782 0.0035434
## HOVAL 0.0011974 0.112930 0.0072118
## =====
## Simulated impact components z-values:
## $Direct
##          INC          HOVAL
## Q1 -3.167835 -3.1796183
## Q2      NaN          NaN
## Q3 -1.703030 -1.5687862
## Q4 -1.272087 -1.0947684
## Q5 -1.002337 -0.8174607
##
## $Indirect
##          INC          HOVAL
## Q1      NaN          NaN
## Q2 -2.465562 -2.4657025
## Q3 -1.703030 -1.5687862
## Q4 -1.272087 -1.0947684

```

```
## Q5 -1.002337 -0.8174607
##
## $Total
##      INC      HOVAL
## Q1 -3.167835 -3.1796183
## Q2 -2.465562 -2.4657025
## Q3 -1.703030 -1.5687862
## Q4 -1.272087 -1.0947684
## Q5 -1.002337 -0.8174607
##
##
## Simulated impact components p-values:
## $Direct
##      INC      HOVAL
## Q1 0.0015358 0.0014747
## Q2 NA      NA
## Q3 0.0885624 0.1166978
## Q4 0.2033424 0.2736181
## Q5 0.3161810 0.4136652
##
## $Indirect
##      INC      HOVAL
## Q1 NA      NA
## Q2 0.013680 0.013674
## Q3 0.088562 0.116698
## Q4 0.203342 0.273618
## Q5 0.316181 0.413665
##
## $Total
##      INC      HOVAL
## Q1 0.0015358 0.0014747
## Q2 0.0136799 0.0136745
## Q3 0.0885624 0.1166978
## Q4 0.2033424 0.2736181
## Q5 0.3161810 0.4136652
```

4.7 Programing the SLM in R

In this Section, we show how to create our own function to estimate a SLM using ML estimation and two different approaches. First, we create a function to estimate the MLE using a constrained optimization procedure and the log-likelihood function (4.15). The second approach uses the algorithm outline in Algorithm (4.1).

4.7.1 First approach

To estimate the SLM using a maximum likelihood procedure, we first create a function that returns the log-likelihood, gradient and Hessian functions. Then, we optimize this function using `maxLik` function from **maxLik** package.

The following function returns the log-likelihood function:

```
# Create log-likelihood function for SLM ----
sml_ll <- function(theta, y, X, W, gradient = TRUE, hessian = TRUE){
  # Global
  K <- ncol(X)
  N <- nrow(X)

  # Extract parameters
  betas <- theta[1:K]
  rho <- theta[K + 1]
  sig.sq <- theta[K + 2]

  # Make residuals
  A <- diag(N) - rho * W
  Ay <- A %*% y
  Xb <- X %*% betas
  res <- Ay - Xb

  # Make log-likelihood
  detA <- det(A)
  ll <- -0.5 * N * log(2 * pi * sig.sq) - 0.5 * crossprod(res) / sig.sq + log(detA)

  # Gradient
  if (gradient){
    C <- W %*% solve(A)
    grad.betas <- (1 / sig.sq) * t(X) %*% res
    grad.rho <- - sum(diag(C)) + (1 / sig.sq) * t(res) %*% W %*% y
    grad.sig.sq <- (1 / (2 * sig.sq ^ 2)) * (t(res) %*% res - N * sig.sq)
    attr(ll, 'gradient') <- c(grad.betas, grad.rho, grad.sig.sq)
  }

  # Hessian
  if (hessian){
    H <- matrix(NA, nrow = (K + 2), ncol = (K + 2))
    h_bb <- - (1 / sig.sq) * t(X) %*% X
    h_bs <- - (1 / sig.sq ^ 2) * t(X) %*% res
    h_br <- - (1 / sig.sq) * t(X) %*% W %*% y
    h_ss <- (N / (2 * sig.sq ^ 2)) - (1 / sig.sq ^ 3) * t(res) %*% res
    h_sr <- - t(res) %*% W %*% y / sig.sq ^ 2
    h_rr <- - sum(diag(C %*% C)) - (1 / sig.sq) * (t(y) %*% t(W) %*% W %*% y)
    H[1:K, 1:K] <- h_bb
    H[1:K, K + 1] <- h_br
  }
}
```

```

H[1:K, K + 2] <- h_bs
H[K + 1, 1:K] <- t(h_br)
H[K + 1, K + 1] <- h_rr
H[K + 1, K + 2] <- h_sr
H[K + 2, 1:K] <- t(h_bs)
H[K + 2, K + 1] <- h_sr
H[K + 2, K + 2] <- h_ss
attr(11, 'hessian') <- H
}
return(11)
}

```

The function `sml_11` has the following arguments: `theta` is a $K + 2$ vector of parameters where the $K + 1$ and $K + 2$ elements are ρ and σ^2 , respectively; `y` is the $n \times 1$ vector of dependent variables; `X` is the $n \times k$ matrix of independent variables; `W` is the spatial weight matrix in `matrix` class; the arguments `gradient` and `hessian` indicate whether the analytical gradient and Hessian, respectively, should be use in the numerical optimization algorithm.

The log-likelihood function is given by object `11` in `sml_11` function. This object is based on Equation (4.15). The gradient is coded following Equation (4.32), whereas the Hessian is based on Equation (4.40).

The following function estimates the model by ML using a constrained optimization procedure. The optimization is made by `maxLik` function:

```

sml_ml <- function(formula, data, listw,
                    gradient = TRUE,
                    hessian = TRUE, ...){
  require("maxLik")
  require("spdep")
  # Model Frame: This part is standard in R to obtain
  #               the variables using formula and data argument.
  callT <- match.call(expand.dots = TRUE)
  mf <- callT
  m <- match(c("formula", "data"), names(mf), 0L)
  mf <- mf[c(1L, m)]
  mf[[1L]] <- as.name("model.frame")
  mf <- eval(mf, parent.frame()) # final model frame
  nframe <- length(sys.calls())

  # Get variables and globals
  y <- model.response(mf) # Get dependent variable from mf
  X <- model.matrix(formula, mf) # Get X from mf
  W <- listw2mat(listw) # listw to matrix
  K <- ncol(X)

  # Starting values
  b_hat <- coef(lm(y ~ X - 1))
  start <- c(b_hat, 0, 1)
}

```

```

names(start) <- c(colnames(X), "rho", "sig.sq")

# Restricted optimization: A %*% theta + B >= 0: Constraint rho and sigma2
sym      <- all(W == t(W))
omega    <- eigen(W, only.values = TRUE, symmetric = sym)
lambda_space <- if (is.complex(omega$values)) 1 / range(Re(omega$values)) else 1 / range(Im(omega$values))

A <- rbind(c(rep(0, K), 1, 0),
           c(rep(0, K), -1, 0),
           c(rep(0, K), 0, 1))
B <- c(-1L * (lambda_space[1] + sqrt(.Machine$double.eps)),
       lambda_space[2] - sqrt(.Machine$double.eps),
       -1L * sqrt(.Machine$double.eps))
callT$constraints <- list(ineqA = A, ineqB = B)

# Optimization default controls if not added by user
if (is.null(callT$method)) callT$method <- 'bfgs'
if (is.null(callT$iterlim)) callT$iterlim <- 100000
opt <- callT
m <- match(c('method', 'print.level', 'iterlim',
            'tol', 'ftol', 'stoptol', 'fixed', 'constraints',
            'control', 'finalHessian', 'reltol', 'rho', 'outer.iterations', 'outer.eps'),
          names(opt), 0L)
opt <- opt[c(1L, m)]
opt$start      <- start
opt[[1]]       <- as.name('maxLik')
opt$logLik     <- as.name('sml_ll')
opt$gradient   <- gradient
opt$hessian    <- hessian
opt[c('y', 'W', 'X')] <- list(as.name('y'),
                             as.name('W'),
                             as.name('X'))
out <- eval(opt, sys.frame(which = nframe))
return(out)
}

```

Now, we use our function:

```

# Load data
data(oldcol, package="spdep")
listw <- spdep::nb2listw(COL.nb, style = "W")

# Use our function
test1 <- slm_ml(CRIME ~ INC + HOVAL, data = COL.OLD, listw = listw)
summary(test1)

## -----

```

```
## Maximum Likelihood estimation
## BFGS maximization, 124 iterations
## Return code 0: successful convergence
## Log-Likelihood: -182.3918
## 5 free parameters
## Estimates:
##           Estimate Std. error t value Pr(> t)
## (Intercept) 44.87928    7.83040   5.731 9.96e-09 ***
## INC         -1.02620    0.32675  -3.141 0.001686 **
## HOVAL        -0.26550    0.08778  -3.025 0.002490 **
## rho          0.43394    0.12299   3.528 0.000418 ***
## sig.sq       94.54494   19.12486   4.944 7.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Warning: constrained likelihood estimation. Inference is probably wrong
## Constrained optimization based on constrOptim
## 1 outer iterations, barrier value -0.008952326
## -----

# Use lagsarlm from spatialreg
library("spatialreg")
sreg <- lagsarlm(CRIME ~ INC + HOVAL, data = COL.OLD, listw = listw)
summary(sreg)

##
## Call:lagsarlm(formula = CRIME ~ INC + HOVAL, data = COL.OLD, listw = listw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -37.68585  -5.35636   0.05421   6.02013  23.20555
##
## Type: lag
## Coefficients: (asymptotic standard errors)
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept) 45.079251    7.177347   6.2808 3.369e-10
## INC         -1.031616    0.305143  -3.3808 0.0007229
## HOVAL        -0.265926    0.088499  -3.0049 0.0026570
##
## Rho: 0.43102, LR test value: 9.9736, p-value: 0.001588
## Asymptotic standard error: 0.11768
##      z-value: 3.6626, p-value: 0.00024962
## Wald statistic: 13.415, p-value: 0.00024962
##
## Log likelihood: -182.3904 for lag model
## ML residual variance (sigma squared): 95.494, (sigma: 9.7721)
## Number of observations: 49
```

```
## Number of parameters estimated: 5
## AIC: 374.78, (AIC for lm: 382.75)
## LM test for residual autocorrelation
## test value: 0.31955, p-value: 0.57188
```

4.7.2 Second approach

Now, we create a function that estimates the parameters of the SLM using the steps in Algorithm (4.1).

```
logLik_sar <- function(rho, e_0, e_L, omega, n)
{
  # This function returns the concentrated log L for maximization

  #Generate determinant using Ord's approximation
  det    <- if (is.complex(omega)) Re(prod(1 - rho * omega)) else prod(1 - rho * omega)
  e_diff <- e_0 - rho * e_L
  sigma2 <- crossprod(e_diff) / n

  #Log-Likelihood function
  l_c    <- - (n / 2) - (n / 2) * log(2 * pi) - (n / 2) * log(sigma2) + log(det)
  return(l_c)
}
```

```
sarML <- function(formula, data, listw)
{
  require("spdep")
  # Model Frame: This part is standard in R to obtain
  # the variables using formula and data argument.
  callT <- match.call(expand.dots = TRUE)
  mf <- callT
  m    <- match(c("formula", "data"), names(mf), 0L)
  mf <- mf[c(1L, m)]
  mf[[1L]] <- as.name("model.frame")
  mf <- eval(mf, parent.frame()) # final model frame

  # Get variables and Globals
  y <- model.response(mf)           # Get dependent variable from mf
  X <- model.matrix(formula, mf)    # Get X from mf
  n <- nrow(X)                      # Number of spatial units
  k <- ncol(X)                     # Number of regressors
  Wy <- lag.listw(listw, y)        # Spatial lag
  W <- listw2mat(listw)            # listw to matrix

  # Generate auxiliary regressions
```



```

# See Algorithm 3.1
ols_0 <- lm(y ~ X - 1)
ols_L <- lm(Wy ~ X - 1)
e_0   <- residuals(ols_0)
e_L   <- residuals(ols_L)

# Get eigenvalues to constraint the optimization
omega <- eigenw(listw)

# Maximize concentrated log-likelihood
rho_space <- if (is.complex(omega)) 1 / range(Re(eig)) else 1 / range(omega)
opt_lc <- optimize(f = logLik_sar, # This function is below
                  lower = rho_space[1] + .Machine$double.eps,
                  upper = rho_space[2] - .Machine$double.eps,
                  maximum = TRUE,
                  e_0 = e_0, e_L = e_L, omega = omega, n = n)

# Obtain rho_hat from concentrated log-likelihood
rho_hat <- opt_lc$maximum

# Generate estimates
A      <- (diag(n) - rho_hat * W)
Ay     <- crossprod(t(A), y)
beta_hat <- solve(crossprod(X)) %*% crossprod(X, Ay) # See Equation (3.25)
error  <- Ay - crossprod(t(X), beta_hat)
sigma2_hat <- crossprod(error) / n # See Equation (3.26)

# Hessian
C      <- crossprod(t(W), solve(A)) #  $C = WA^{-1}$ 
alpha  <- sum(omega ^ 2 / ((1 - rho_hat * omega) ^ 2))
if (is.complex(alpha)) alpha <- Re(alpha)
b_b    <- drop(1 / sigma2_hat) * crossprod(X) #  $k \times k$ 
b_rho  <- drop(1 / sigma2_hat) * (t(X) %*% C %*% X %*% beta_hat) #  $k \times 1$ 
sig_sig <- n / (2 * sigma2_hat ^ 2) #  $1 \times 1$ 
sig_rho <- drop(1 / sigma2_hat) * sum(diag(C)) #  $1 \times 1$ 
rho_rho <- sum(diag(crossprod(C))) + alpha +
  drop(1 / sigma2_hat) * crossprod(C %*% X %*% beta_hat) #  $1 \times 1$ 
row_1  <- cbind(b_b, rep(0, k), b_rho)
row_2  <- cbind(t(rep(0, k)), sig_sig, sig_rho)
row_3  <- cbind(t(b_rho), sig_rho, rho_rho)
Hessian <- rbind(row_1, row_2, row_3)
std.err <- sqrt(diag(solve(Hessian)))

# Table of coefficients
all_names <- c(colnames(X), "sigma2", "rho")
all_coef  <- c(beta_hat, sigma2_hat, rho_hat)
z         <- all_coef / std.err

```

```

p                <- pnorm(abs(z), lower.tail = FALSE) * 2
sar_table        <- cbind(all_coef, std.err, z, p)
cat(paste("\nEstimates from SAR Model \n\n"))
colnames(sar_table) <- c("Estimate", "Std. Error", "z-value", "Pr(>|z|)")
rownames(sar_table) <- all_names
printCoefmat(sar_table)
}

```

```

test2 <- sarML(CRIME ~ INC + HOVAL, data = COL.OLD, listw = listw)

##
## Estimates from SAR Model
##
##              Estimate Std. Error z-value  Pr(>|z|)
## (Intercept) 45.079025   7.177334  6.2807 3.369e-10 ***
## INC         -1.031610   0.305143 -3.3807 0.0007229 ***
## HOVAL       -0.265926   0.088499 -3.0049 0.0026570 **
## sigma2      95.494400  19.487804  4.9002 9.573e-07 ***
## rho         0.431027   0.117680  3.6627 0.0002496 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

4.8 Exercises

Exercise 4.1 Consider the concentrated log-likelihood in Equation (4.30). Find the first and second derivative respect to ρ .

Exercise 4.2 Consider the Spatial Lag Model:

$$\begin{aligned} \mathbf{y} &= \rho \mathbf{W} \mathbf{y} + \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon} &\sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n) \end{aligned}$$

Let $\mathbf{z} = \mathbf{A} \mathbf{y}$. Show that $\hat{\sigma}_{ML}^2$ can be written as:

$$\hat{\sigma}_{ML}^2 = \frac{1}{N} \mathbf{z}^\top \mathbf{M} \mathbf{z}$$

where $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$.

Exercise 4.3 Consider the Spatial Error Model:

$$\begin{aligned} \mathbf{y} &= \mathbf{X} \boldsymbol{\beta} + \mathbf{u} \\ \mathbf{u} &= \lambda \mathbf{W} \mathbf{u} + \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon} &\sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n) \end{aligned}$$

- (a) Show that the OLS estimates $\hat{\beta}$ is unbiased, but inefficient.
- (b) Derived the ML estimates.
- (c) Derived the concentrated log-likelihood function.
- (d) Derive the asymptotic variance-covariance matrix of the estimates given in Equation (4.50).

Exercise 4.4 Consider the following SAC model with heteroskedastic errors:

$$\mathbf{y} = \rho \mathbf{W}_1 \mathbf{y} + \mathbf{X} \boldsymbol{\beta} + \mathbf{u} \quad (4.67)$$

$$\mathbf{u} = \lambda \mathbf{W}_2 \mathbf{u} + \boldsymbol{\varepsilon} \quad (4.68)$$

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Omega}) \quad (4.69)$$

The matrix $\boldsymbol{\Omega}$ is the variance-covariance matrix of the error terms, which is assumed to be known a priori. For example, we can assume that:

$$\mathbb{V}(\epsilon_i) = \sigma_i^2 = \mathbf{z}_i^\top \boldsymbol{\alpha} \quad (4.70)$$

or

$$\mathbb{V}(\epsilon_i) = \sigma_i^2 = \exp(\mathbf{z}_i^\top \boldsymbol{\alpha}) \quad (4.71)$$

or more general,

$$\mathbb{V}(\epsilon_i) = \sigma_i^2 = \mathbf{h}(\mathbf{z}_i^\top \boldsymbol{\alpha}) \quad (4.72)$$

where $\mathbf{h}(\cdot)$ is any function, \mathbf{z}_i is a vector of covariates for each spatial unit, and $\boldsymbol{\alpha}$ is a vector of parameters with element $\alpha_p, p = 0, 1, \dots, P$. Therefore, the diagonal elements of the error covariance matrix $\boldsymbol{\Omega}$ are:

$$\boldsymbol{\Omega}_{ii} = \sigma_i^2 = \mathbf{h}_i(\mathbf{z}_i^\top \boldsymbol{\alpha}), \quad \mathbf{h}_i > 0 \quad (4.73)$$

Note that the model has $2 + K + P$ unknown parameters:

$$\boldsymbol{\theta} = (\rho, \boldsymbol{\beta}^\top, \lambda, \boldsymbol{\alpha}^\top)^\top. \quad (4.74)$$

- (a) Find the Log-likelihood function.
- (b) Find the first order conditions

Exercise 4.5 Consider the model:

$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \rho_1 \mathbf{W}_1 \mathbf{y} + \rho_2 \mathbf{W}_2 \mathbf{y} + \mathbf{u}, \quad (4.75)$$

where \mathbf{u} has mean and VC matrix of $\mathbf{0}$ and $\sigma^2 \mathbf{I}_n$, respectively, and \mathbf{W}_1 and \mathbf{W}_2 , are observed exogenous weighting matrices.

- (a) Obtain the likelihood function, and then determine the first order conditions for $\boldsymbol{\beta}$.

- (b) Assume that \mathbf{W}_1 and \mathbf{W}_2 are row-normalized. Give a condition which is sufficient for the model to be solved for \mathbf{y} in terms of \mathbf{X} and $\boldsymbol{\varepsilon}$.

Exercise 4.6 Show that $\mathbf{I}_n + \rho_0 \mathbf{C}_n = \mathbf{A}_n^{-1}$.

Exercise 4.7 Consider the following DGP:

$$\begin{aligned} y_i &= \alpha + \beta x_i + u_i \\ u_i &= \lambda \sum_{j=1}^n w_{ij} u_j + \epsilon_i \\ \epsilon_i &\sim N(0, 1) \end{aligned} \tag{4.76}$$

where $\lambda = 0.8$, $\alpha = 0.5$, $\beta = 1$ and $x_i \sim N(0, 2^2)$. Using a Monte Carlo experiment, show that the $\hat{\beta}_{OLS}$ is unbiased, but inefficient. For experiment create 100 datasets with 225 spatial units. Set the seed at 123.

Appendix

4.A Consistency of SLM Model

Since $\hat{\beta}_n$ and $\hat{\sigma}_n^2$ are continuous functions of $\hat{\rho}_n$, it suffice to show consistency of $\hat{\rho}_n$. To prove that $\hat{\rho}_n$ is consistent, we will show that $\frac{1}{n}\ell_n(\rho) - \frac{1}{n}Q_n(\rho)$ converges in probability to zero uniformly on Γ (where $Q_n(\rho)$ is the expectation of the concentrated log-likelihood), and the identification-uniqueness condition holds.

Uniform convergence In this first part, we need to show that

$$\frac{1}{n}\ell_n(\rho) - \frac{1}{n}Q_n(\rho) \xrightarrow{p} 0,$$

uniformly on Γ , where $\ell_n(\rho)$ is the concentrated log-likelihood and $Q_n(\rho)$ is the expectation of the log-likelihood function evaluated in the optimal values of β and σ^2 , $Q_n(\rho) = \max_{\beta, \sigma^2} \mathbb{E}(\ell_n(\boldsymbol{\theta}))$.

For simplicity in the notation, let

$$\mathbf{A}_n = \mathbf{A}_n(\rho) = (\mathbf{I} - \rho \mathbf{W}) \quad \text{and} \quad \mathbf{A}_{n0} = \mathbf{A}_n(\rho_0) = (\mathbf{I} - \rho_0 \mathbf{W}).$$

For further reference, recall that the log-likelihood function is

$$\ell_n(\boldsymbol{\theta}) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \boldsymbol{\varepsilon}_n^\top \boldsymbol{\varepsilon}_n + \ln |\mathbf{A}_n|,$$

with $\boldsymbol{\varepsilon}_n = \mathbf{A}_n \mathbf{y}_n - \mathbf{X}_n \boldsymbol{\beta}$. The concentrated log-likelihood can be written as

$$\ell_n(\rho) = -\frac{n}{2} (\ln(2\pi) + 1) - \frac{n}{2} \ln(\tilde{\sigma}(\rho)) + \log |\mathbf{A}_n|,$$

where

$$\begin{aligned}\tilde{\sigma}(\rho) &= \frac{1}{n} \left[\mathbf{A}_n \mathbf{y}_n - \mathbf{X}_n \hat{\boldsymbol{\beta}}_n(\rho) \right]^\top \left[\mathbf{A}_n \mathbf{y}_n - \mathbf{X}_n \hat{\boldsymbol{\beta}}_n(\rho) \right], \\ &= \frac{1}{n} \mathbf{y}_n^\top \mathbf{A}_n^\top (\mathbf{I}_n - \mathbf{P}_{nX}) \mathbf{A}_n \mathbf{y}_n, \\ &= \frac{1}{n} \mathbf{y}_n^\top \mathbf{A}_n^\top \mathbf{M}_n \mathbf{A}_n \mathbf{y}_n,\end{aligned}\tag{4.77}$$

with $\hat{\boldsymbol{\beta}}_n(\rho)$ being the MLE of $\boldsymbol{\beta}_0$ with depends on ρ , $\mathbf{P}_{nX} = \mathbf{X}_n (\mathbf{X}_n^\top \mathbf{X}_n)^{-1} \mathbf{X}_n^\top$, and $\mathbf{M}_n = \mathbf{I}_n - \mathbf{P}_{nX}$ is the projection matrix.

The expectation of $\ell_n(\boldsymbol{\theta})$ is:

$$\mathbb{E}(\ell_n(\boldsymbol{\theta})) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \boldsymbol{\varepsilon}_n) + \ln |\mathbf{A}_n|. \tag{4.78}$$

We need to work on $\mathbb{E}(\boldsymbol{\varepsilon}_n^\top \boldsymbol{\varepsilon}_n)$. To do so, note that under the true DGP $\mathbf{y}_n = \mathbf{A}_{n0}^{-1} \mathbf{X}_n \boldsymbol{\beta}_0 + \mathbf{A}_{n0}^{-1} \boldsymbol{\varepsilon}_n$, then

$$\begin{aligned}\mathbb{E}(\mathbf{y}_n) &= \boldsymbol{\mu}_y = \mathbf{A}_{n0}^{-1} \mathbf{X}_n \boldsymbol{\beta}_0, \\ \mathbb{V}(\mathbf{y}_n) &= \boldsymbol{\Sigma}_y = \sigma_0^2 \mathbf{A}_{n0}^{-1} (\mathbf{A}_{n0}^{-1})^\top, \\ \text{tr}(\mathbf{A}_n^\top(\rho) \mathbf{A}_n(\rho) \boldsymbol{\Sigma}_y) &= \sigma_0^2 \text{tr}[\mathbf{A}_n^\top \mathbf{A}_n \mathbf{A}_{n0}^{-1} (\mathbf{A}_{n0}^{-1})^\top], \\ &= \sigma_0^2 \text{tr}[\mathbf{D}_n(\rho_0, \rho)], \\ \boldsymbol{\mu}_y^\top \mathbf{A}_n^\top(\rho) \mathbf{A}_n(\rho) \boldsymbol{\mu}_y &= \boldsymbol{\beta}_0^\top \mathbf{X}_n^\top (\mathbf{A}_{n0}^{-1})^\top \mathbf{A}_n^\top \mathbf{A}_n \mathbf{A}_{n0}^{-1} \mathbf{X}_n \boldsymbol{\beta}_0, \\ &= \boldsymbol{\beta}_0^\top \mathbf{X}_n^\top \mathbf{D}_n(\rho_0, \rho) \mathbf{X}_n \boldsymbol{\beta}_0,\end{aligned}\tag{4.79}$$

where $\mathbf{D}_n(\rho_0, \rho) = (\mathbf{A}_{n0}^{-1})^\top \mathbf{A}_n^\top \mathbf{A}_n \mathbf{A}_{n0}^{-1}$.

Then, using Lemma 3.25 for the expectation of quadratic forms and the results in Equation (4.79)

$$\begin{aligned}\mathbb{E}(\boldsymbol{\varepsilon}_n^\top \boldsymbol{\varepsilon}_n) &= \mathbb{E}[(\mathbf{A}_n \mathbf{y}_n - \mathbf{X}_n \boldsymbol{\beta})^\top (\mathbf{A}_n \mathbf{y}_n - \mathbf{X}_n \boldsymbol{\beta})], \\ &= \mathbb{E}[\mathbf{y}_n^\top \mathbf{A}_n^\top \mathbf{A}_n \mathbf{y}_n - 2\boldsymbol{\beta}^\top \mathbf{X}_n^\top \mathbf{A}_n \mathbf{y}_n + \boldsymbol{\beta}^\top \mathbf{X}_n^\top \mathbf{X}_n \boldsymbol{\beta}], \\ &= \mathbb{E}[\mathbf{y}_n^\top \mathbf{A}_n^\top \mathbf{A}_n \mathbf{y}_n] - \mathbb{E}[2\boldsymbol{\beta}^\top \mathbf{X}_n^\top \mathbf{A}_n \mathbf{y}_n] + \mathbb{E}[\boldsymbol{\beta}^\top \mathbf{X}_n^\top \mathbf{X}_n \boldsymbol{\beta}], \\ &= \text{tr}(\mathbf{A}_n^\top \mathbf{A}_n \boldsymbol{\Sigma}_y) + \boldsymbol{\mu}_y^\top \mathbf{A}_n^\top \mathbf{A}_n \boldsymbol{\mu}_y - 2\boldsymbol{\beta}^\top \mathbf{X}_n^\top \mathbf{A}_n \boldsymbol{\mu}_y + \boldsymbol{\beta}^\top \mathbf{X}_n^\top \mathbf{X}_n \boldsymbol{\beta}, \\ &= \sigma_0^2 \text{tr}[\mathbf{D}_n(\rho_0, \rho)] + \boldsymbol{\beta}_0^\top \mathbf{X}_n^\top \mathbf{D}_n(\rho_0, \rho) \mathbf{X}_n \boldsymbol{\beta}_0 - 2\boldsymbol{\beta}^\top \mathbf{X}_n^\top \mathbf{A}_n \mathbf{A}_{n0}^{-1} \mathbf{X}_n \boldsymbol{\beta}_0 + \boldsymbol{\beta}^\top \mathbf{X}_n^\top \mathbf{X}_n \boldsymbol{\beta}.\end{aligned}\tag{4.80}$$

Then, the FONC for the optimal solutions of $\mathbb{E}(\ell_n(\boldsymbol{\theta}))$ in Equation (4.78) are

$$\begin{aligned}\frac{\partial \mathbb{E}(\ell_n(\boldsymbol{\theta}))}{\partial \boldsymbol{\beta}} &= -2\mathbf{X}_n^\top \mathbf{A}_n \mathbf{A}_{n0}^{-1} \mathbf{X}_n \boldsymbol{\beta}_0 + 2\mathbf{X}_n^\top \mathbf{X}_n \boldsymbol{\beta}, \\ \frac{\partial \mathbb{E}(\ell_n(\boldsymbol{\theta}))}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \boldsymbol{\varepsilon}_n),\end{aligned}$$

such that

$$\boldsymbol{\beta}_n^* = (\mathbf{X}_n^\top \mathbf{X}_n)^{-1} \mathbf{X}_n^\top \mathbf{A}_n \mathbf{A}_{n0}^{-1} \mathbf{X}_n \boldsymbol{\beta}_0,$$

and

$$\begin{aligned}
\sigma_n^{2*} &= \frac{1}{n} \mathbb{E} \left(\boldsymbol{\varepsilon}_n^\top \boldsymbol{\varepsilon}_n \right), \\
&= \frac{1}{n} \mathbb{E} \left([\mathbf{A}_n \mathbf{y}_n - \mathbf{X}_n \boldsymbol{\beta}_n^*]^\top [\mathbf{A}_n \mathbf{y}_n - \mathbf{X}_n \boldsymbol{\beta}_n^*] \right), \\
&= \frac{1}{n} \left\{ (\rho_0 - \rho)^2 (\mathbf{C}_n \mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{M}_n (\mathbf{C}_n \mathbf{X}_n \boldsymbol{\beta}_0) + \sigma_0^2 \text{tr}(\mathbf{D}_n(\rho_0, \rho)) \right\}, \\
&= \frac{1}{n} \left\{ (\rho_0 - \rho)^2 (\mathbf{C}_n \mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{M}_n (\mathbf{C}_n \mathbf{X}_n \boldsymbol{\beta}_0) \right\} + \sigma_n^2(\rho),
\end{aligned} \tag{4.81}$$

where $\sigma_n^2(\lambda) = \frac{1}{n} \sigma_0^2 \text{tr}(\mathbf{D}_n(\rho_0, \lambda))$.

Therefore

$$Q_n(\rho) = -\frac{n}{2} (\ln(2\pi) + 1) - \frac{n}{2} \ln(\sigma_n^{2*}(\rho)) + \log |\mathbf{A}_n|.$$

Then,

$$\frac{1}{n} [\ell_n(\rho) - Q_n(\rho)] = -\frac{1}{2} [\ln(\tilde{\sigma}^2) - \ln(\sigma_n^{2*}(\rho))], \tag{4.82}$$

thus, we need to show that $\tilde{\sigma}^2 \xrightarrow{p} \sigma_n^{2*}$.

We can show that $\mathbf{A}_n \mathbf{A}_{n0}^{-1} = \mathbf{I}_n + (\rho_0 - \rho) \mathbf{C}_n(\rho_0)$, then we can also write

$$\begin{aligned}
(\mathbf{I}_n - \mathbf{P}_{nX}) \mathbf{A}_n \mathbf{y}_n &= \mathbf{M}_n \mathbf{A}_n \mathbf{y}_n, \\
&= \mathbf{M}_n \mathbf{A}_n \mathbf{A}_{n0}^{-1} \mathbf{X}_n \boldsymbol{\beta}_0 + \mathbf{M}_n \mathbf{A}_n \mathbf{A}_{n0}^{-1} \boldsymbol{\varepsilon}_n, \\
&= \mathbf{M}_n (\mathbf{I}_n + (\rho_0 - \rho) \mathbf{C}_n(\rho_0)) \mathbf{X}_n \boldsymbol{\beta}_0 + \mathbf{M}_n \mathbf{A}_n \mathbf{A}_{n0}^{-1} \boldsymbol{\varepsilon}_n, \\
&= (\rho_0 - \rho) \mathbf{M}_n \mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0 + \mathbf{M}_n \mathbf{A}_n \mathbf{A}_{n0}^{-1} \boldsymbol{\varepsilon}_n,
\end{aligned}$$

From (4.77), we can write

$$\begin{aligned}
\tilde{\sigma}(\rho) &= \frac{1}{n} \mathbf{y}_n^\top \mathbf{A}_n^\top (\mathbf{I}_n - \mathbf{P}_{nX}) \mathbf{A}_n \mathbf{y}_n, \\
&= \frac{1}{n} \left((\rho_0 - \rho) \mathbf{M}_n \mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0 + \mathbf{M}_n \mathbf{A}_n \mathbf{A}_{n0}^{-1} \boldsymbol{\varepsilon}_n \right)^\top \left((\rho_0 - \rho) \mathbf{M}_n \mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0 + \mathbf{M}_n \mathbf{A}_n \mathbf{A}_{n0}^{-1} \boldsymbol{\varepsilon}_n \right), \\
&= \frac{1}{n} (\rho_0 - \rho)^2 (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{M}_n (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0) + 2(\rho_0 - \rho) \underbrace{\frac{1}{n} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{M}_n \mathbf{A}_n \mathbf{A}_{n0}^{-1} \boldsymbol{\varepsilon}_n}_{B_{n1}} + \\
&\quad \underbrace{\frac{1}{n} \boldsymbol{\varepsilon}_n^\top (\mathbf{A}_{n0}^\top)^{-1} \mathbf{A}_n^\top \mathbf{M}_n \mathbf{A}_n \mathbf{A}_{n0}^{-1} \boldsymbol{\varepsilon}_n}_{B_{n2}} \\
&= \sigma_n^{2*} + B_{n1} + (B_{n2} - \sigma_n^2(\lambda)),
\end{aligned}$$

where the last equality comes from Equation (4.81).

It can be shown that

$$\begin{aligned}
\frac{1}{n} \mathbf{X}_n^\top \mathbf{C}_n^\top \boldsymbol{\varepsilon}_n &\xrightarrow{p} \mathbf{0} \\
\frac{1}{n} \mathbf{X}_n^\top \mathbf{C}_n^\top \mathbf{C}_n \boldsymbol{\varepsilon}_n &\xrightarrow{p} \mathbf{0}.
\end{aligned} \tag{4.83}$$

This implies that $\frac{1}{n} \mathbf{X}_n^\top \mathbf{C}_n^\top \boldsymbol{\varepsilon}_n = o_p(1)$ and $\frac{1}{n} \mathbf{X}_n^\top \mathbf{C}_n^\top \mathbf{C}_n \boldsymbol{\varepsilon}_n = o_p(1)$.

B_{n1} can be expanded as

$$\begin{aligned}
B_{n1} &= \frac{1}{n} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{M}_n \mathbf{A}_n \mathbf{A}_{n0}^{-1} \boldsymbol{\varepsilon}_n \\
&= \frac{1}{n} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{M}_n (\mathbf{I}_n + (\rho_0 - \rho) \mathbf{C}_n(\rho_0)) \boldsymbol{\varepsilon}_n \\
&= \frac{1}{n} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{M}_n \boldsymbol{\varepsilon}_n + (\rho_0 - \rho) \frac{1}{n} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{M}_n \mathbf{C}_n(\rho_0) \boldsymbol{\varepsilon}_n \\
&= o_p(1) + o_p(1) = o_p(1)
\end{aligned}$$

So that $B_{n1} = o_p(1)$ uniformly in $\rho \in \Gamma$.

B_{n2} can be expanded as

$$\begin{aligned}
B_{n2} - \sigma_n^2(\lambda) &= \frac{1}{n} \boldsymbol{\varepsilon}_n^\top (\mathbf{A}_{n0}^\top)^{-1} \mathbf{A}_n^\top \mathbf{M}_n \mathbf{A}_n \mathbf{A}_{n0}^{-1} \boldsymbol{\varepsilon}_n - \frac{1}{n} \sigma_0^2 \text{tr}(\mathbf{D}_n(\rho_0, \rho)) \\
&= \frac{1}{n} \boldsymbol{\varepsilon}_n^\top (\mathbf{A}_{n0}^\top)^{-1} \mathbf{A}_n^\top (\mathbf{I}_n - \mathbf{P}_{nX}) \mathbf{A}_n \mathbf{A}_{n0}^{-1} \boldsymbol{\varepsilon}_n - \frac{1}{n} \sigma_0^2 \text{tr}(\mathbf{D}_n(\rho_0, \rho)) \\
&= \left[\frac{1}{n} \boldsymbol{\varepsilon}_n^\top (\mathbf{A}_{n0}^\top)^{-1} \mathbf{A}_n^\top \mathbf{A}_n \mathbf{A}_{n0}^{-1} \boldsymbol{\varepsilon}_n - \frac{1}{n} \sigma_0^2 \text{tr}(\mathbf{D}_n(\rho_0, \rho)) \right] - \frac{1}{n} \boldsymbol{\varepsilon}_n^\top (\mathbf{A}_{n0}^\top)^{-1} \mathbf{A}_n^\top \mathbf{P}_{nX} \mathbf{A}_n \mathbf{A}_{n0}^{-1} \boldsymbol{\varepsilon}_n \\
&= o_p(1) + o_p(1)
\end{aligned}$$

So that $B_{n2} - \sigma_n^2(\lambda) = o_p(1)$ uniformly in $\rho \in \Gamma$. Then

$$\tilde{\sigma}(\rho) \xrightarrow{p} \sigma_n^{2*}$$

Consequently,

$$\begin{aligned}
\ln(\tilde{\sigma}^2) - \ln(\sigma_n^{2*}(\rho)) &= \ln \left(\frac{\tilde{\sigma}^2}{\sigma_n^{2*}(\rho)} \right) \\
&= \ln \left(\frac{\sigma_n^{2*} + B_{n1} + (B_{n2} - \sigma_n^2(\lambda))}{\sigma_n^{2*}(\rho)} \right) \\
&= \ln \left(1 + \frac{B_{n1}}{\sigma_n^{2*}(\rho)} + \frac{(B_{n2} - \sigma_n^2(\lambda))}{\sigma_n^{2*}(\rho)} \right) \\
&= o_p(1)
\end{aligned}$$

and

$$\sup_{\rho \in \Gamma} \left\{ \frac{1}{n} |\ell_n(\rho) - Q_n(\rho)| \right\} = o_p(1)$$

Identification Here, we need to show that for any $\epsilon > 0$, $\limsup_{n \rightarrow \infty} \left[\max_{\rho \in \bar{N}_\epsilon(\rho_0)} \frac{1}{n} Q_n(\rho) - \frac{1}{n} Q_n(\rho_0) \right] < 0$, where $\bar{N}_\epsilon(\rho_0)$ is the complement of an open neighborhood of ρ in Γ with radius ϵ .

Consider the log-likelihood function of a pure SLM process $\mathbf{y}_n = \rho \mathbf{W}_n \mathbf{y}_n + \boldsymbol{\varepsilon}_n$, where $\boldsymbol{\varepsilon}_n \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I}_n)$ is

$$\ell_{p,n}(\rho, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 + \ln |\mathbf{A}_n(\rho)| - \frac{1}{2\sigma^2} \mathbf{y}_n^\top \mathbf{A}_n(\rho)^\top \mathbf{A}_n(\rho) \mathbf{y}_n.$$

Consider $Q_{p,n}(\rho) = \max_{\sigma^2} \mathbb{E}_p [\ell_{p,n}(\rho, \sigma^2)]$. Then

$$\begin{aligned} Q_{p,n}(\rho) &= \max_{\sigma^2} \mathbb{E}_p [\ell_{p,n}(\rho, \sigma^2)] \\ &\leq \mathbb{E}_p \left[\max_{\sigma^2} \ell_{p,n}(\rho, \sigma^2) \right], \quad \text{by Jensen's inequality} \\ &= Q_{p,n}(\rho_0), \quad \text{for all } \rho. \end{aligned}$$

This implies that

$$\frac{1}{n} (Q_{p,n}(\rho) - Q_{p,n}(\rho_0)) \leq 0 \quad \text{for all } \rho.$$

It is also clear that:

$$\ln(\sigma_n^2(\rho)) \leq \ln \sigma_n^{2*}(\rho) \quad (4.84)$$

At $\rho_0, \sigma_n^{2*}(\rho) = \sigma_0^2$ (see Equation (4.81)). Then,

$$\begin{aligned} \frac{1}{n} Q_n(\rho) - \frac{1}{n} Q_n(\rho_0) &= -\frac{1}{2} (\ln \sigma_n(\rho) - \ln \sigma_0^2) + \frac{1}{n} (\ln |\mathbf{A}_n| - \ln |\mathbf{A}_0|) - \frac{1}{2} [\ln \sigma_n^{2*}(\rho) - \ln \sigma_n^2(\rho)] \\ &= \frac{1}{n} (Q_{p,n}(\rho) - Q_{p,n}(\rho_0)) - \frac{1}{2} [\ln \sigma_n^{2*}(\rho) - \ln \sigma_n^2(\rho)] \end{aligned}$$

It follows that

$$\frac{1}{n} Q_n(\rho) - \frac{1}{n} Q_n(\rho_0) \leq 0.$$

4.B Expected Value of Hessian for SLM

In this Section we drop the subindex n and use the results from Section 4.2.3. The following definitions and relations for the Spatial Lag Model are very useful:

Since $\boldsymbol{\varepsilon} = \mathbf{A}_0 \mathbf{y} - \mathbf{X} \boldsymbol{\beta}_0$, it follows that

$$\begin{aligned} \mathbb{E}(\boldsymbol{\varepsilon}) &= \mathbf{0} \\ \mathbb{E}(\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top) &= \sigma_0^2 \mathbf{I}_n, \\ \mathbb{E}(\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}) &= \mathbb{E}(\text{tr}(\boldsymbol{\varepsilon}^\top \mathbf{I}_n \boldsymbol{\varepsilon})) = n \mathbb{E}(\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top) \end{aligned}$$

Using Lemma 3.25 for the expectation of quadratic forms, yields

$$\begin{aligned} \mathbf{y} &= \mathbf{A}_0^{-1} \mathbf{X} \boldsymbol{\beta}_0 + \mathbf{A}_0^{-1} \boldsymbol{\varepsilon}, \\ \mathbb{E}(\mathbf{y}) &= \boldsymbol{\mu}_u = \mathbf{A}_0^{-1} \mathbf{X} \boldsymbol{\beta}_0, \\ \mathbb{E}(\mathbf{y}^\top \mathbf{y}) &= \boldsymbol{\Sigma}_y = \mathbf{A}_0^{-1} \sigma_0^2 \mathbf{I}_n (\mathbf{A}_0^{-1})^\top, \\ \mathbb{E}(\mathbf{y} \mathbf{y}^\top) &= (\mathbf{A}_0^{-1} \mathbf{X} \boldsymbol{\beta}_0) (\mathbf{A}_0^{-1} \mathbf{X} \boldsymbol{\beta}_0)^\top + \mathbf{A}_0^{-1} \sigma_0^2 \mathbf{I}_n (\mathbf{A}_0^{-1})^\top. \end{aligned} \quad (4.85)$$

We now derive the most difficult expectations. From (4.36) and letting $\mathbf{C}_0 = \mathbf{W} \mathbf{A}_0^{-1}$

$$\begin{aligned} \mathbb{E} \left(\frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta} \partial \rho} \right) &= -\frac{1}{\sigma_0^2} \mathbf{X}^\top \mathbb{E}(\mathbf{W} \mathbf{y}) \\ &= -\frac{1}{\sigma_0^2} \mathbf{X}^\top \mathbf{W} \mathbf{A}_0^{-1} \mathbf{X} \boldsymbol{\beta}_0 \\ &= -\frac{1}{\sigma_0^2} \mathbf{X}^\top (\mathbf{C}_0 \mathbf{X} \boldsymbol{\beta}_0) \end{aligned} \quad (4.86)$$

For (4.37) we obtain:

$$\begin{aligned}
 \mathbb{E} \left(\frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial (\sigma^2)^2} \right) &= \mathbb{E} \left[\frac{n}{2(\sigma_0^2)^2} - \frac{1}{(\sigma_0^2)^3} \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} \right] \\
 &= \frac{n}{2\sigma_0^4} - \frac{1}{\sigma_0^6} \mathbb{E} [\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}] \\
 &= \frac{n}{2\sigma_0^4} - \frac{n}{\sigma_0^6} \sigma_0^2 \mathbf{I}_n \\
 &= -\frac{n}{2\sigma_0^4}
 \end{aligned} \tag{4.87}$$

From (4.38):

$$\begin{aligned}
 \mathbb{E} \left[\frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \sigma^2 \partial \rho} \right] &= \mathbb{E} \left[-\frac{\boldsymbol{\varepsilon}^\top \mathbf{W} \mathbf{y}}{\sigma_0^4} \right] \\
 &= -\frac{1}{\sigma_0^4} \mathbb{E} [\boldsymbol{\varepsilon}^\top \mathbf{W} (\mathbf{A}_0^{-1} \mathbf{X} \boldsymbol{\beta}_0 + \mathbf{A}_0^{-1} \boldsymbol{\varepsilon})] \\
 &= -\frac{1}{\sigma_0^4} \mathbb{E} [\boldsymbol{\varepsilon}^\top \mathbf{C}_0 \mathbf{X} \boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}^\top \mathbf{C}_0 \boldsymbol{\varepsilon}] \\
 &= -\frac{1}{\sigma_0^4} \mathbb{E} [\boldsymbol{\varepsilon}^\top \mathbf{C}_0 \boldsymbol{\varepsilon}] \\
 &= -\frac{1}{\sigma_0^4} \text{tr}(\mathbf{C}_0) \mathbb{E} (\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top) \\
 &= -\text{tr}(\mathbf{C}_0) / \sigma_0^2
 \end{aligned} \tag{4.88}$$

From (4.39):

$$\begin{aligned}
 \mathbb{E} \left[\frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \rho^2} \right] &= \mathbb{E} \left[-\text{tr} [(\mathbf{W} \mathbf{A}_0^{-1})^2] - \frac{1}{\sigma_0^2} (\mathbf{y}^\top \mathbf{W}^\top \mathbf{W} \mathbf{y}) \right] \\
 &= -\text{tr} [(\mathbf{W} \mathbf{A}_0^{-1})^2] - \frac{1}{\sigma_0^2} \mathbb{E} [\mathbf{y}^\top \mathbf{W}^\top \mathbf{W} \mathbf{y}] \\
 &= -\text{tr}(\mathbf{C}_0^2) - \frac{1}{\sigma_0^2} (\text{tr}(\mathbf{W}^\top \mathbf{W} \boldsymbol{\Sigma}_y) + \boldsymbol{\mu}_y^\top \mathbf{W}^\top \mathbf{W} \boldsymbol{\mu}_y) \\
 &= -\text{tr}(\mathbf{C}_0^2) - \frac{1}{\sigma_0^2} (\sigma_0^2 \text{tr}(\mathbf{W}^\top \mathbf{W} \mathbf{A}_0^{-1} (\mathbf{A}_0^{-1})^\top) + (\mathbf{A}_0^{-1} \mathbf{X} \boldsymbol{\beta}_0)^\top \mathbf{W}^\top \mathbf{W} \mathbf{A}_0^{-1} \mathbf{X} \boldsymbol{\beta}_0) \\
 &= -\text{tr}(\mathbf{C}_0^2) - \frac{1}{\sigma_0^2} (\boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{C}_0^\top \mathbf{C}_0 \mathbf{X} \boldsymbol{\beta}_0 + \sigma_0^2 \text{tr}(\mathbf{C}_0^\top \mathbf{C}_0)) \\
 &= -\text{tr}(\mathbf{C}_0^2) - \frac{1}{\sigma_0^2} (\mathbf{C}_0 \mathbf{X} \boldsymbol{\beta}_0)^\top (\mathbf{C}_0 \mathbf{X} \boldsymbol{\beta}_0) - \text{tr}(\mathbf{C}_0^\top \mathbf{C}_0) \\
 &= -\text{tr}(\mathbf{C}_0^s \mathbf{C}_0) - \frac{1}{\sigma_0^2} (\mathbf{C}_0 \mathbf{X} \boldsymbol{\beta}_0)^\top (\mathbf{C}_0 \mathbf{X} \boldsymbol{\beta}_0)
 \end{aligned} \tag{4.89}$$

where $\mathbf{C}_0^s = \mathbf{C}_0 + \mathbf{C}_0^\top$.

Let $\mathbf{H}(\boldsymbol{\theta}) = \frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}$. Thus, minus the expected value of the Hessian evaluated at $\boldsymbol{\theta}_0$ is:

$$-\mathbb{E} [\mathbf{H}(\boldsymbol{\theta}_0)] = \begin{pmatrix} \frac{1}{n\sigma_0^2} (\mathbf{X}^\top \mathbf{X}) & \frac{1}{\sigma_0^2} \mathbf{X}^\top (\mathbf{C} \mathbf{X} \boldsymbol{\beta}_0) & \mathbf{0}^\top \\ \text{tr}(\mathbf{C}_0^s \mathbf{C}_0) + \frac{1}{\sigma_0^2} (\mathbf{C}_0 \mathbf{X} \boldsymbol{\beta}_0)^\top (\mathbf{C}_0 \mathbf{X} \boldsymbol{\beta}_0) & \frac{1}{\sigma_0^2} \text{tr}(\mathbf{C}_0) & \frac{n}{2\sigma_0^4} \end{pmatrix} \tag{4.90}$$

By multiplying $-\mathbb{E} [\mathbf{H}(\boldsymbol{\theta}_0)]$ by $(1/n)$, we obtain Equation (4.61).

4.C Variance of the Score Function

In this Appendix, we will derive the variance of the score function. That is

$$\mathbb{V} \left[\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right] = \mathbb{E} \left[\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \cdot \frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^\top} \right] = \mathbf{J}_{\boldsymbol{\theta},n} = \boldsymbol{\Sigma}_{\boldsymbol{\theta},n} + \boldsymbol{\Omega}_{\boldsymbol{\theta},n} \quad (4.91)$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\theta},n} = -\mathbb{E}[(1/n)\mathbf{H}(\boldsymbol{\theta}_0)]$ is given in Equation (4.61). The following results are important. For i.i.d $\boldsymbol{\varepsilon}_n = (\epsilon_1, \dots, \epsilon_n)$, it can be shown that

$$\begin{aligned} \mathbb{E}(\epsilon_i \epsilon_j) &= \begin{cases} \sigma_0^2 & \text{for } i = j, \\ 0 & \text{for } i \neq j \end{cases} \\ \mathbb{E}(\epsilon_i \epsilon_j \epsilon_s) &= \begin{cases} \mu_3 & \text{for } i = j = s, \\ 0 & \text{otherwise} \end{cases} \\ \mathbb{E}(\epsilon_i \epsilon_j \epsilon_s \epsilon_t) &= \begin{cases} \mu_4 & \text{for } i = j = s = t, \\ \sigma_0^4 & \text{for } i = j \neq s = t \text{ or } i = s \neq j = t \text{ or } i = t \neq s = t, \\ 0 & \text{otherwise} \end{cases} \end{aligned} \quad (4.92)$$

If $\epsilon_i \sim N(0, \sigma_0^2)$, we have $\mathbb{E}(\epsilon_i^3) = \mu_3 = 0$ and $\mathbb{E}(\epsilon_i^4) = \mu_4 = 3\sigma_0^4$. Thus

$$\begin{aligned} \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) &= \mathbb{E} \left[\sum_{i=1}^n \sum_{j=1}^n a_{ij} \epsilon_i \epsilon_j \right] = \sum_{i=1}^n a_{ii} \mathbb{E}(\epsilon_i \epsilon_i) = \sigma_0^2 \sum_{i=1}^n a_{ii} = \sigma_0^2 \text{tr}(\mathbf{A}_n), \\ \mathbb{E}(\boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) &= \sum_{s=1}^n \sum_{i=1}^n \sum_{j=1}^n a_{ij} \mathbb{E}(\epsilon_i \epsilon_j \epsilon_s) = \sum_i a_{ii} \mathbb{E}(\epsilon_i^3), \\ &= \mu_3 \text{tr}(\mathbf{A}_n) \\ \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n^\top \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^\top \mathbf{A}_n^\top \boldsymbol{\varepsilon}_n) &= \sum_i \sum_j \sum_s \sum_t a_{ij} a_{st} \mathbb{E}(\epsilon_i \epsilon_j \epsilon_s \epsilon_t), \\ &= (\mu_4 - 3\sigma_0^4) \sum_{i=1}^n a_{ii}^2 + \sigma_0^4 \left[\text{tr}^2(\mathbf{A}_n) + \text{tr}(\mathbf{A}_n \mathbf{A}_n^\top) + \text{tr}(\mathbf{A}_n^2) \right], \\ &= (\mu_4 - 3\sigma_0^4) \sum_{i=1}^n a_{ii}^2 + \sigma_0^4 \left[\text{tr}^2(\mathbf{A}_n) + \text{tr}(\mathbf{A}_n^s \mathbf{A}_n) \right] \end{aligned} \quad (4.93)$$

where $\mathbf{A}_n^s = (\mathbf{A}_n + \mathbf{A}_n^\top)$.

Then, the expectation for the elements of $\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \cdot \frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^\top}$ are

$$\begin{aligned} \mathbb{E} \left[\left(\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta}} \right) \left(\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta}} \right)^\top \right] &= \frac{1}{\sigma_0^4} \frac{1}{n} \mathbb{E}(\mathbf{X}_n^\top \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top \mathbf{X}_n), \\ &= \frac{1}{\sigma_0^2} \frac{1}{n} \mathbf{X}_n^\top \mathbf{X}_n. \end{aligned} \quad (4.94)$$

$$\begin{aligned}
\mathbb{E} \left[\left(\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta}} \right) \left(\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \rho} \right)^\top \right] &= \frac{1}{\sigma_0^4} \frac{1}{n} \left[\mathbb{E}(\mathbf{X}_n^\top \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^\top (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)) + \mathbb{E}(\mathbf{X}_n^\top \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^\top \mathbf{C}_{n0} \boldsymbol{\varepsilon}_n) - \right. \\
&\quad \left. \sigma_0^2 \mathbb{E}(\mathbf{X}_n^\top \boldsymbol{\varepsilon}_n \text{tr}(\mathbf{C}_{n0})) \right], \\
&= \frac{1}{\sigma_0^4} \frac{1}{n} \left[\mathbf{X}_n^\top \mathbb{E}(\boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^\top) (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0) + \mathbf{X}_n^\top \mathbb{E}(\boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^\top \mathbf{C}_{n0} \boldsymbol{\varepsilon}_n) \right], \\
&= \frac{1}{\sigma_0^2} \frac{1}{n} \mathbf{X}_n^\top (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0) + \frac{1}{\sigma_0^4} \frac{\mu_3}{n} \mathbf{X}_n^\top \text{diag}(\mathbf{C}_{n0}).
\end{aligned} \tag{4.95}$$

$$\begin{aligned}
\mathbb{E} \left[\left(\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta}} \right) \left(\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \sigma^2} \right)^\top \right] &= \frac{1}{\sigma_0^6} \frac{1}{n} \mathbb{E} \left[\frac{1}{2} \mathbf{X}_n^\top \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^\top \boldsymbol{\varepsilon}_n - \frac{n \sigma_0^2}{2} \mathbf{X}_n^\top \boldsymbol{\varepsilon}_n \right], \\
&= \frac{\mu_3}{\sigma_0^6} \frac{1}{2n} \mathbf{X}_n^\top \mathbf{z}_n.
\end{aligned} \tag{4.96}$$

$$\begin{aligned}
\mathbb{E} \left[\left(\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \sigma^2} \right) \left(\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \rho} \right)^\top \right] &= \frac{1}{n \sigma_0^2} \text{tr}(\mathbf{C}_{n0}) + \frac{1}{2n \sigma_0^6} \left[(\mu_4 - 3\sigma_0^4) \text{tr}(\mathbf{C}_{n0}) + \mu_3 \mathbf{z}^\top (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0) \right] \\
\mathbb{E} \left[\left(\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \sigma^2} \right) \left(\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \sigma^2} \right)^\top \right] &= \frac{1}{2\sigma_0^4} + \frac{(\mu_4 - 3\sigma_0^4)}{4\sigma_0^8}
\end{aligned} \tag{4.97}$$

Note that:

$$\begin{aligned}
\left(\frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \rho} \right) \left(\frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \rho} \right)^\top &= (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^\top (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0) + (\boldsymbol{\varepsilon}_n^\top \mathbf{C}_{n0} \boldsymbol{\varepsilon}_n)^2 + \sigma_0^4 \text{tr}^2(\mathbf{C}_{n0}) \\
&\quad + 2(\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^\top \mathbf{C}_{n0}^\top \boldsymbol{\varepsilon}_n - 2\sigma_0^2 \text{tr}(\mathbf{C}_{n0}) (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \boldsymbol{\varepsilon}_n \\
&\quad - 2\sigma_0^2 \text{tr}(\mathbf{C}_{n0}) \boldsymbol{\varepsilon}_n^\top \mathbf{C}_{n0} \boldsymbol{\varepsilon}_n
\end{aligned}$$

Taking the expectation for each element of this gives and given the results in Equation (4.93)

$$\begin{aligned}
\mathbb{E} \left[(\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^\top (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0) \right] &= \sigma_0^2 (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0) \\
\mathbb{E} \left[(\boldsymbol{\varepsilon}_n^\top \mathbf{C}_{n0} \boldsymbol{\varepsilon}_n)^2 \right] &= (\mu_4 - 3\sigma_0^4) \sum_{i=1}^n c_{n,ii}^2 + \sigma_0^4 \left[\text{tr}^2(\mathbf{C}_{n0}) + \text{tr}(\mathbf{C}_{n0}^s \mathbf{C}_{n0}) \right] \\
\mathbb{E} \left[\sigma_0^4 \text{tr}^2(\mathbf{C}_{n0}) \right] &= \sigma_0^4 \text{tr}^2(\mathbf{C}_{n0}) \\
\mathbb{E} \left[2(\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^\top \mathbf{C}_{n0}^\top \boldsymbol{\varepsilon}_n \right] &= 2\mu_3 (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \text{diag}(\mathbf{C}_{n0}) \\
\mathbb{E} \left[2\sigma_0^2 \text{tr}(\mathbf{C}_{n0}) (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \boldsymbol{\varepsilon}_n \right] &= \mathbf{0} \\
\mathbb{E} \left[2\sigma_0^2 \text{tr}(\mathbf{C}_{n0}) \boldsymbol{\varepsilon}_n^\top \mathbf{C}_{n0} \boldsymbol{\varepsilon}_n \right] &= 2\sigma_0^4 \text{tr}^2(\mathbf{C}_{n0})
\end{aligned}$$

Using these results, we obtain

$$\begin{aligned}
\mathbb{E} \left[\left(\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \rho} \right) \left(\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \rho} \right)^\top \right] &= \frac{1}{\sigma_0^4 n} \left[\sigma_0^2 (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0) - \sigma_0^4 \text{tr}^2(\mathbf{C}_{n0}) \right. \\
&\quad + (\mu_4 - 3\sigma_0^4) \sum_{i=1}^n c_{n,ii}^2 + \sigma_0^4 \left[\text{tr}^2(\mathbf{C}_{n0}) + \text{tr}(\mathbf{C}_{n0}^s \mathbf{C}_{n0}) \right] \\
&\quad + 2\mu_3 (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \text{diag}(\mathbf{C}_{n0}) \Big] \\
&= \frac{1}{\sigma_0^2} \frac{1}{n} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0) + \frac{1}{n} \text{tr}(\mathbf{C}_{n0}^s \mathbf{C}_{n0}) \\
&\quad + \frac{1}{\sigma_0^4} \frac{1}{n} (\mu_4 - 3\sigma_0^4) \sum_{i=1}^n c_{ii,0}^2 + \frac{1}{\sigma_0^4} \frac{2\mu_3}{n} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \text{diag}(\mathbf{C}_{n0})
\end{aligned} \tag{4.98}$$

Then,

$$\mathbf{J}_{\theta,n} = \begin{pmatrix} \frac{1}{\sigma_0^2} \frac{1}{n} \mathbf{X}_n^\top \mathbf{X}_n & \frac{1}{n\sigma_0^2} \mathbf{X}_n^\top (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0) + \frac{\mu_3}{n\sigma_0^4} \mathbf{X}_n^\top \text{diag}(\mathbf{C}_{n0}) & \frac{\mu_3}{2n\sigma_0^6} \mathbf{X}_n^\top \mathbf{z}_n \\ * & \frac{1}{n\sigma_0^2} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0) + \frac{1}{n} \text{tr}(\mathbf{C}_{n0}^s \mathbf{C}_{n0}) + J_{22,n} & \frac{1}{n\sigma_0^2} \text{tr}(\mathbf{C}_{n0}) + J_{23,n} \\ * & * & \frac{1}{2\sigma_0^4} + \frac{(\mu_4 - 3\sigma_0^4)}{4\sigma_0^8} \end{pmatrix} \tag{4.99}$$

where

$$\begin{aligned}
J_{22,n} &= \frac{1}{\sigma_0^4} \frac{1}{n} (\mu_4 - 3\sigma_0^4) \sum_{i=1}^n c_{ii,0}^2 + \frac{1}{\sigma_0^4} \frac{2\mu_3}{n} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \text{diag}(\mathbf{C}_{n0}), \\
J_{23,n} &= \frac{1}{2n\sigma_0^6} \left[(\mu_4 - 3\sigma_0^4) \text{tr}(\mathbf{C}_{n0}) + \mu_3 \mathbf{z}^\top (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0) \right].
\end{aligned}$$

Thus, we can write $\mathbf{J}_{\theta,n} = \boldsymbol{\Sigma}_{\theta,n} + \boldsymbol{\Omega}_{\theta,n}$ with

$$\boldsymbol{\Omega}_{\theta,n} = \begin{pmatrix} \mathbf{0} & \frac{\mu_3}{n\sigma_0^4} \mathbf{X}_n^\top \text{diag}(\mathbf{C}_{n0}) & \frac{\mu_3}{2n\sigma_0^6} \mathbf{X}_n^\top \mathbf{z}_n \\ * & \frac{1}{n\sigma_0^4} (\mu_4 - 3\sigma_0^4) \sum_{i=1}^n c_{ii,0}^2 + \frac{2\mu_3}{n\sigma_0^4} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \text{diag}(\mathbf{C}_{n0}) & \frac{1}{2n\sigma_0^6} \left[(\mu_4 - 3\sigma_0^4) \text{tr}(\mathbf{C}_{n0}) + \mu_3 \mathbf{z}^\top (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0) \right] \\ * & * & \frac{(\mu_4 - 3\sigma_0^4)}{4\sigma_0^8} \end{pmatrix}. \tag{4.100}$$

4.D Proof of Asymptotic Normality

From Equation (4.56), we know

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = - \left[\frac{1}{n} \frac{\partial^2 \ell_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right]^{-1} \frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}.$$

The sketch consists in the following steps:

(a) First, we need to show that:

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = - \lim_{n \rightarrow \infty} \mathbb{E} \left[\frac{1}{n} \frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right],$$

is non-singular. To show this is beyond the scope of this class notes. We will take this as given.

(b) Now we will show that

$$\frac{1}{n} \frac{\partial^2 \log L_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \xrightarrow{p} \frac{1}{n} \frac{\partial^2 \log L_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}$$

By Assumption 4.9 (No asymptotic multicollinearity), we know that $\lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{X}_n^\top \mathbf{X}_n$ exists, therefore $\mathbf{X}_n^\top \mathbf{X}_n = O(n)$ so that $\mathbf{X}_n^\top \mathbf{X}_n/n = O(1)$ (Lemma 3.23)⁷ and $\tilde{\sigma}_n^2 \xrightarrow{p} \sigma_0^2$ from consistency, then from Equation (4.34), we have:

$$\begin{aligned} \frac{1}{n} \frac{\partial^2 \log L_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} - \frac{1}{n} \frac{\partial^2 \log L_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} &= -\frac{1}{\tilde{\sigma}_n^2} \frac{(\mathbf{X}_n^\top \mathbf{X}_n)}{n} + \frac{1}{\sigma_0^2} \frac{(\mathbf{X}_n^\top \mathbf{X}_n)}{n} \\ &= \underbrace{\left(\frac{1}{\sigma_0^2} - \frac{1}{\tilde{\sigma}_n^2} \right)}_{o_p(1)} \underbrace{\frac{(\mathbf{X}_n^\top \mathbf{X}_n)}{n}}_{O(1)} \\ &= o_p(1) O(1) \\ &= o_p(1) \end{aligned}$$

It can also be shown that

$$\begin{aligned} \frac{1}{n} \mathbf{X}_n^\top \mathbf{W}_n^\top \mathbf{y}_n &= \frac{1}{n} \mathbf{X}_n^\top \mathbf{C}_n \mathbf{X}_n \boldsymbol{\beta}_0 + o_p(1) \\ \frac{1}{n} \mathbf{y}_n^\top \mathbf{W}_n^\top \boldsymbol{\varepsilon}_n &= \frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{C}_n^\top \boldsymbol{\varepsilon}_n + o_p(1) \\ \frac{1}{n} \mathbf{y}_n^\top \mathbf{W}_n^\top \mathbf{W}_n \mathbf{y}_n &= \frac{1}{n} (\mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{C}_n^\top \mathbf{C}_n \mathbf{X}_n \boldsymbol{\beta}_0 + \frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{C}_n^\top \mathbf{C}_n \boldsymbol{\varepsilon}_n + o_p(1) \end{aligned}$$

As $\mathbf{X}_n^\top \mathbf{W}_n \mathbf{y}_n/n = O_p(1)$, it follows from Equation (4.36):

$$\begin{aligned} \frac{1}{n} \frac{\partial^2 \log L_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\rho}} - \frac{1}{n} \frac{\partial^2 \log L_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\rho}} &= -\frac{1}{\tilde{\sigma}_n^2} \frac{(\mathbf{X}_n^\top \mathbf{W}_n \mathbf{y}_n)}{n} + \frac{1}{\sigma_0^2} \frac{(\mathbf{X}_n^\top \mathbf{W}_n \mathbf{y}_n)}{n} \\ &= \underbrace{\left(\frac{1}{\sigma_0^2} - \frac{1}{\tilde{\sigma}_n^2} \right)}_{o_p(1)} \underbrace{\frac{(\mathbf{X}_n^\top \mathbf{W}_n \mathbf{y}_n)}{n}}_{O_p(1)} \\ &= o_p(1) O_p(1) = o_p(1) \end{aligned}$$

The following identity will be useful:

$$\begin{aligned} \boldsymbol{\varepsilon}(\tilde{\boldsymbol{\delta}}_n) &= \mathbf{y}_n - \mathbf{X}_n \tilde{\boldsymbol{\beta}}_n - \tilde{\rho}_n \mathbf{W}_n \mathbf{y}_n + (\boldsymbol{\varepsilon}(\boldsymbol{\delta}_0) - \boldsymbol{\varepsilon}(\boldsymbol{\delta}_0)) \\ &= \mathbf{y}_n - \mathbf{X}_n \tilde{\boldsymbol{\beta}}_n - \tilde{\rho}_n \mathbf{W}_n \mathbf{y}_n - \mathbf{y}_n + \mathbf{X}_n \boldsymbol{\beta}_0 - \rho_0 \mathbf{W}_n \mathbf{y}_n + \boldsymbol{\varepsilon}_n(\boldsymbol{\delta}_0) \\ &= \mathbf{X}_n (\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}}_n) + (\rho_0 - \tilde{\rho}_n) \mathbf{W}_n \mathbf{y}_n + \boldsymbol{\varepsilon}_n(\boldsymbol{\delta}_0) \end{aligned} \tag{4.101}$$

⁷Since $\lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{X}_n^\top \mathbf{X}_n$ exists, then, each of its elements is $o(1)$ and hence $O(1)$. In other words, $\frac{1}{n} \mathbf{X}_n^\top \mathbf{X}_n$ is a bounded matrix.

Then, taking into account Equation (4.35) and using our result in Equation (4.101) yields:

$$\begin{aligned}
\frac{1}{n} \frac{\partial^2 \log L_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\beta} \partial \sigma^2} - \frac{1}{n} \frac{\partial^2 \log L_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta} \partial \sigma^2} &= -\frac{1}{\tilde{\sigma}_n^4} \frac{\mathbf{X}_n^\top \boldsymbol{\varepsilon}(\tilde{\boldsymbol{\delta}}_n)}{n} + \frac{1}{\sigma_0^4} \frac{\mathbf{X}_n^\top \boldsymbol{\varepsilon}_n(\boldsymbol{\delta}_0)}{n} \\
&= -\frac{1}{\tilde{\sigma}_n^4} \frac{1}{n} \mathbf{X}_n^\top \left(\mathbf{X}_n(\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}}_n) + (\rho_0 - \tilde{\rho}_n) \mathbf{W}_n \mathbf{y}_n + \boldsymbol{\varepsilon}_n(\boldsymbol{\delta}_0) \right) + \\
&\quad + \frac{1}{\sigma_0^4} \frac{\mathbf{X}_n^\top \boldsymbol{\varepsilon}_n(\boldsymbol{\delta}_0)}{n} \\
&= \frac{1}{\tilde{\sigma}_n^4} \frac{1}{n} \mathbf{X}_n^\top \mathbf{X}_n (\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}}_n) - \frac{1}{\tilde{\sigma}_n^4} \frac{1}{n} \mathbf{X}_n^\top \mathbf{W}_n \mathbf{y}_n (\rho_0 - \tilde{\rho}_n) \\
&\quad - \frac{1}{\tilde{\sigma}_n^4} \frac{1}{n} \mathbf{X}_n^\top \boldsymbol{\varepsilon}_n(\boldsymbol{\delta}_0) + \frac{1}{\sigma_0^4} \frac{\mathbf{X}_n^\top \boldsymbol{\varepsilon}_n(\boldsymbol{\delta}_0)}{n} \\
&= \left(\frac{1}{\sigma_0^4} - \frac{1}{\tilde{\sigma}_n^4} \right) \frac{\mathbf{X}_n^\top \boldsymbol{\varepsilon}_n(\boldsymbol{\delta}_0)}{n} + \frac{\mathbf{X}_n^\top \mathbf{X}_n}{n \tilde{\sigma}_n^4} (\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}}_n) \\
&\quad + \frac{\mathbf{W}_n^\top \mathbf{W}_n \mathbf{y}_n}{n \tilde{\sigma}_n^4} (\rho_0 - \tilde{\rho}_n) \\
&= o_p(1) O_p(1) + O(1) o_p(1) + O_p(1) o_p(1) \\
&= o_p(1)
\end{aligned}$$

From Equation (4.39), we know that:

$$\frac{\partial \log L_n(\boldsymbol{\theta})}{\partial \rho^2} = -\text{tr} \left[(\mathbf{C}_n(\rho))^2 \right] - \frac{1}{\sigma^2} (\mathbf{y}_n^\top \mathbf{W}_n^\top \mathbf{W}_n \mathbf{y}_n) \quad \text{where} \quad \mathbf{C}_n(\rho) = \mathbf{W}_n \mathbf{A}_n(\rho)^{-1}$$

From Mean Value Theorem around of $\text{tr} [(\mathbf{C}_n(\tilde{\rho}_n))^2]$ around ρ_0 :

$$\begin{aligned}
\text{tr} [(\mathbf{C}_n(\tilde{\rho}_n))^2] &= \text{tr} [(\mathbf{C}_n(\rho_0))^2] + 2 \text{tr} [(\mathbf{C}_n(\bar{\rho}))^3] (\rho_0 - \tilde{\rho}_n) \\
\text{tr} [(\mathbf{C}_n(\tilde{\rho}_n))^2] - \text{tr} [(\mathbf{C}_n(\rho_0))^2] &= 2 \text{tr} [(\mathbf{C}_n(\bar{\rho}))^3] (\rho_0 - \tilde{\rho}_n)
\end{aligned}$$

Then:

$$\begin{aligned}
\frac{1}{n} \frac{\partial^2 \log L_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \rho^2} - \frac{1}{n} \frac{\partial^2 \log L_n(\boldsymbol{\theta}_0)}{\partial \rho^2} &= \underbrace{2 \frac{1}{n} \text{tr} [(\mathbf{C}_n(\bar{\rho}))^3]}_{O(n)} \underbrace{(\rho_0 - \tilde{\rho}_n)}_{o_p(1)} + \underbrace{\left(\frac{1}{\sigma_0^2} - \frac{1}{\tilde{\sigma}_n^2} \right)}_{o_p(1)} \underbrace{\frac{\mathbf{y}_n^\top \mathbf{W}_n^\top \mathbf{W}_n \mathbf{y}_n}{n}}_{O_p(n)} \\
&= o_p(1)
\end{aligned}$$

Note that $\mathbf{C}_n(\bar{\rho})$ is uniformly bounded in row and column sums uniformly in a neighborhood of ρ_0 by Assumption 4.7 and 4.10. Note that $\text{tr} [(\mathbf{C}_n(\bar{\rho}))^3] = O(n)$.

Considering Equation (4.38):

$$\begin{aligned}
\frac{1}{n} \frac{\partial^2 \log L_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \sigma^2 \partial \rho} - \frac{1}{n} \frac{\partial^2 \log L_n(\boldsymbol{\theta}_0)}{\partial \sigma^2 \partial \rho} &= -\frac{1}{\tilde{\sigma}^4} \frac{1}{n} \mathbf{y}_n^\top \mathbf{W}_n^\top \boldsymbol{\varepsilon}_n(\tilde{\boldsymbol{\delta}}_n) + \frac{1}{\sigma^4} \frac{1}{n} \mathbf{y}_n^\top \mathbf{W}_n^\top \boldsymbol{\varepsilon}_n(\boldsymbol{\delta}_n) \\
&= -\frac{1}{\tilde{\sigma}^4} \frac{1}{n} \mathbf{y}_n^\top \mathbf{W}_n^\top \left[\mathbf{X}_n(\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}}_n) + (\rho_0 - \tilde{\rho}_n) \mathbf{W}_n \mathbf{y}_n + \boldsymbol{\varepsilon}_n(\boldsymbol{\delta}_0) \right] + \\
&\quad \frac{1}{\sigma^4} \frac{1}{n} \mathbf{y}_n^\top \mathbf{W}_n^\top \boldsymbol{\varepsilon}_n(\boldsymbol{\delta}_n) \\
&= -\frac{1}{\tilde{\sigma}^4} \frac{1}{n} \mathbf{y}_n^\top \mathbf{W}_n^\top \mathbf{X}_n(\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}}_n) + \frac{1}{\tilde{\sigma}^4} \frac{1}{n} \mathbf{y}_n^\top \mathbf{W}_n^\top \mathbf{W}_n \mathbf{y}_n (\tilde{\rho}_n - \rho_0) + \\
&\quad \left(\frac{1}{\sigma^4} - \frac{1}{\tilde{\sigma}^4} \right) \frac{1}{n} \mathbf{y}_n^\top \mathbf{W}_n^\top \boldsymbol{\varepsilon}_n \\
&= o_p(1)
\end{aligned}$$

Note the following:

$$\begin{aligned}
\frac{1}{n} \boldsymbol{\varepsilon}(\tilde{\boldsymbol{\delta}})^\top \boldsymbol{\varepsilon}(\tilde{\boldsymbol{\delta}}) &= (\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^\top \frac{\mathbf{X}_n^\top \mathbf{X}_n}{n} + (\tilde{\rho}_n - \rho_0)^2 \frac{\mathbf{y}_n^\top \mathbf{W}_n^\top \mathbf{W}_n \mathbf{y}_n}{n} + \frac{\boldsymbol{\varepsilon}_n^\top \boldsymbol{\varepsilon}_n}{n} + \\
&\quad 2(\tilde{\rho}_n - \rho_0) (\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^\top \frac{\mathbf{X}_n^\top \mathbf{W}_n \mathbf{y}_n}{n} + 2(\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}}_n)^\top \frac{\mathbf{X}_n^\top \boldsymbol{\varepsilon}_n}{n} + 2(\rho_0 - \tilde{\rho}_n) \frac{\mathbf{y}_n^\top \mathbf{W}_n^\top \boldsymbol{\varepsilon}_n}{n} \\
&= \frac{\boldsymbol{\varepsilon}_n^\top \boldsymbol{\varepsilon}_n}{n} + o_p(1)
\end{aligned}$$

Finally, considering the second derivative in Equation (4.37)

$$\begin{aligned}
\frac{1}{n} \frac{\partial^2 \log L(\tilde{\boldsymbol{\theta}})}{\partial (\sigma^2)^2} - \frac{1}{n} \frac{\partial^2 \log L(\boldsymbol{\theta}_0)}{\partial (\sigma^2)^2} &= \frac{1}{2(\tilde{\sigma}^2)^2} - \frac{1}{(\tilde{\sigma}^2)^3} \frac{1}{n} \boldsymbol{\varepsilon}(\tilde{\boldsymbol{\delta}})^\top \boldsymbol{\varepsilon}(\tilde{\boldsymbol{\delta}}) - \frac{1}{2(\sigma^2)^2} + \frac{1}{(\sigma^2)^3} \frac{1}{n} \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} \\
&= \frac{1}{2} \left(\frac{1}{(\tilde{\sigma}^2)^2} - \frac{1}{(\sigma^2)^2} \right) - \frac{1}{(\tilde{\sigma}^2)^3} \frac{1}{n} \boldsymbol{\varepsilon}(\tilde{\boldsymbol{\delta}})^\top \boldsymbol{\varepsilon}(\tilde{\boldsymbol{\delta}}) + \frac{1}{(\sigma^2)^3} \frac{1}{n} \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} \\
&= \frac{1}{2} \left(\frac{1}{(\tilde{\sigma}^2)^2} - \frac{1}{(\sigma^2)^2} \right) - \frac{1}{(\tilde{\sigma}^2)^3} \left(\frac{\boldsymbol{\varepsilon}_n^\top \boldsymbol{\varepsilon}_n}{n} + o_p(1) \right) + \frac{1}{(\sigma^2)^3} \frac{1}{n} \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} \\
&= \frac{1}{2} \left(\frac{1}{(\tilde{\sigma}^2)^2} - \frac{1}{(\sigma^2)^2} \right) + \left(\frac{1}{(\sigma^2)^3} - \frac{1}{(\tilde{\sigma}^2)^3} \right) \frac{\boldsymbol{\varepsilon}_n^\top \boldsymbol{\varepsilon}_n}{n} + o_p(1) \\
&= o_p(1)
\end{aligned}$$

(c) Now, we need to show that:

$$\frac{1}{n} \frac{\partial^2 \log L_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \xrightarrow{p} \mathbb{E} \left[\frac{1}{n} \frac{\partial^2 \log L_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right] \quad (4.102)$$

Form Appendix 4.B, we know that:

$$\begin{aligned}
\mathbb{E} \left[\frac{1}{n} \frac{\partial^2 \log L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \middle| \mathbf{W}, \mathbf{X} \right] &= -\frac{1}{\sigma_0^2} \frac{1}{n} (\mathbf{X}^\top \mathbf{X}) \\
\mathbb{E} \left[\frac{1}{n} \frac{\partial^2 \log L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta} \partial \sigma^2} \middle| \mathbf{W}, \mathbf{X} \right] &= \mathbf{0} \\
\mathbb{E} \left[\frac{1}{n} \frac{\partial^2 \log L(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta} \partial \rho} \middle| \mathbf{W}, \mathbf{X} \right] &= -\frac{1}{\sigma_0^2} \frac{1}{n} \mathbf{X}^\top \mathbf{C} \mathbf{X} \boldsymbol{\beta}_0 \\
\mathbb{E} \left[\frac{1}{n} \frac{\partial^2 \log L(\boldsymbol{\theta}_0)}{\partial (\sigma^2)^2} \middle| \mathbf{W}, \mathbf{X} \right] &= -\frac{1}{2\sigma_0^4} \\
\mathbb{E} \left[\frac{1}{n} \frac{\partial^2 \log L(\boldsymbol{\theta}_0)}{\partial \sigma^2 \partial \rho} \middle| \mathbf{W}, \mathbf{X} \right] &= -\frac{1}{n} \text{tr}(\mathbf{C}) / \sigma_0^2 \\
\mathbb{E} \left[\frac{1}{n} \frac{\partial^2 \log L(\boldsymbol{\theta}_0)}{\partial \rho^2} \middle| \mathbf{W}, \mathbf{X} \right] &= -\frac{1}{n} \text{tr}(\mathbf{C}^s \mathbf{C}) - \frac{1}{\sigma_0^2} \frac{1}{n} (\mathbf{C} \mathbf{X} \boldsymbol{\beta}_0)^\top (\mathbf{C} \mathbf{X} \boldsymbol{\beta}_0)
\end{aligned}$$

All these expectations exist in the limit by Assumption 4.11 and Lemma 3.25-3.26. Then, by nonsingularity of $\mathbb{E}[\mathbf{H}(\mathbf{w}_i; \boldsymbol{\theta})]$, we can say that

$$\left(\frac{1}{n} \mathbf{H}(\mathbf{w}; \hat{\boldsymbol{\theta}}) \right)^{-1} \xrightarrow{p} \mathbb{E}[\mathbf{H}(\mathbf{w}; \boldsymbol{\theta}_0)]^{-1}$$

(d) Recall that the first-order derivatives of the log-likelihood function at $\boldsymbol{\theta}_0$ are given by:

$$\frac{1}{\sqrt{n}} \frac{\partial \log L_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{1}{\sigma_0^2 \sqrt{n}} \mathbf{X}_n^\top \boldsymbol{\varepsilon}_n \\ \frac{1}{2\sigma_0^4 \sqrt{n}} (\boldsymbol{\varepsilon}_n' \boldsymbol{\varepsilon}_n - n\sigma_0^2) \\ \frac{1}{\sigma_0^2 \sqrt{n}} (\mathbf{C}_n \mathbf{X}_n \boldsymbol{\beta}_0)^\top \boldsymbol{\varepsilon}_n + \frac{1}{\sigma_0^2 \sqrt{n}} (\boldsymbol{\varepsilon}_n^\top \mathbf{C}_n \boldsymbol{\varepsilon}_n - \sigma_0^2 \text{tr}(\mathbf{C}_n)) \end{pmatrix}$$

As explained by Lee (2004, pag. 1905), these are linear and quadratic functions of $\boldsymbol{\varepsilon}_n$. In particular, the asymptotic distribution of $\frac{1}{\sqrt{n}} \frac{\partial \log L_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}$ may be derived from central limit theorem for linear-quadratic forms. The matrix \mathbf{C}_n is uniformly bounded in row sums. As the elements of \mathbf{X}_n are bounded, the elements of $\mathbf{C}_n \mathbf{X}_n \boldsymbol{\beta}_0$ for all n are uniformly bounded by Lemma 3.22. With the existence of high order moments of ϵ in Assumption 4.3, the central limit theorem for quadratic forms of double arrays of Kelejian and Prucha (2001) can be applied and the limit distribution of the score vector follows.

Hypothesis Testing

In the previous chapter we have presented the spatial autoregressive models, the intuition underlying their DGP, and their estimation by ML. At this stage the following question arises: which model is more convenient for empirical analysis? There exists two ways to proceed. The first way is to use a spatial model according to some theoretical considerations. The second approach suggests that a series of statistical test should be carried out on the different specifications of the spatial autocorrelation models to adopt the one that better control for spatial autocorrelation among residuals.

In this chapter we present some approaches to test whether the true spatial parameters are zero or not. In other words, we would like to assess the null $H_0 : \lambda = 0$ or $H_0 : \rho = 0$, under the alternative $H_1 : \lambda \neq 0$ or $H_1 : \rho \neq 0$.

We first start with the Moran's I statistic used to test whether there is some evidence of spatial autocorrelation in the error term. Then, we present several test based on the ML principle.

5.1 Test for Residual Spatial Autocorrelation Based on the Moran I Statistic

5.1.1 Cliff and Ord Derivation

Recall from Section 1.4.1 that the Moran's I test allows us assess whether the observed value of a variable at one location is independent of values of that variable at neighboring locations. One could also in principle apply the same test to the OLS residuals to assess whether some spatial autocorrelation remains. If the true DGP follows a spatial process, and we wrongly ignore it, then Moran's I on the OLS residuals should detect this misspecification.

A Moran I statistic for spatial autocorrelation can be applied to regression residuals in a straightforward way. Formally, this I statistic is:

$$I = \left(\frac{n}{S_0} \right) \frac{\hat{\epsilon}^\top \mathbf{W} \hat{\epsilon}}{\hat{\epsilon}^\top \hat{\epsilon}}$$

where $\hat{\epsilon}$ is a vector of OLS residuals, \mathbf{W} is a spatial weight matrix, n is the number of observations and S_0 is a standardization factor, equal to the sum of all elements in the

weight matrix. For a weight matrix that is normalized such that the row elements sum to one, expression (5.2) simplifies to:

$$I = \frac{\hat{\boldsymbol{\varepsilon}}^\top \mathbf{W} \hat{\boldsymbol{\varepsilon}}}{\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}} \quad (5.1)$$

The asymptotic distribution for the Moran statistic with regression residuals was developed by [Cliff and Ord \(1972, 1973\)](#). In particular, the following Theorem give us the moment of the Moran's I statistic and its distribution.

Theorem 5.1 — Moran's I . Consider H_0 : no spatial autocorrelation, and assume that $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$. Let the Moran's I statistic be:

$$I = \left(\frac{n}{S_0} \right) \frac{\hat{\boldsymbol{\varepsilon}}^\top \mathbf{W} \hat{\boldsymbol{\varepsilon}}}{\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}}} \quad (5.2)$$

where $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ is a vector of OLS residuals, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$, \mathbf{W} is a spatial weight matrix, n is the number of observations and S_0 is a standardization factor, equal to the sum of all elements in the weight matrix. Then, the moments under the null are:

$$\begin{aligned} \mathbb{E}(I) &= \frac{n}{S_0} \frac{\text{tr}(\mathbf{M}\mathbf{W})}{n - K} \\ \mathbb{E}(I^2) &= \frac{\left(\frac{n}{S_0} \right)^2 \text{tr}(\mathbf{M}\mathbf{W}\mathbf{M}\mathbf{W}^\top) + \text{tr}(\mathbf{M}\mathbf{W})^2 + [\text{tr}(\mathbf{M}\mathbf{W})]^2}{(n - K)(n - K + 2)} \end{aligned} \quad (5.3)$$

where $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$. Then:

$$z_I = \frac{I - \mathbb{E}(I)}{\sqrt{\mathbb{V}(I)}} \sim N(0, 1) \quad (5.4)$$

where $\mathbb{V}(I) = \mathbb{E}(I^2) - \mathbb{E}(I)^2$.

According to [Anselin \(1988, p. 102\)](#), the interpretation of this test is not always straightforward, even though it is by far the most widely used approach. While the null hypothesis is obviously the absence of spatial dependence, a precise expression for the alternative hypothesis does not exist. Intuitively, the spatial weight matrix is taken to represent the pattern of potential spatial interaction that causes dependence, but the nature of the underlying DGP is not specified. Usually it is assumed to be of a spatial autoregressive form. However, the coefficient 5.1 is mathematically equivalent to an OLS regression of $\mathbf{W}\hat{\boldsymbol{\varepsilon}}$ on $\hat{\boldsymbol{\varepsilon}}$, rather than for $\hat{\boldsymbol{\varepsilon}}$ on $\mathbf{W}\hat{\boldsymbol{\varepsilon}}$, which would correspond to an autoregressive process as in SEM model. In other words, Moran's I is a misspecification test that has power against a host of alternatives. This includes spatial error autocorrelation, but also residual correlation caused by a spatial lag alternative, and even heteroskedasticity! Thus, the rejection of the null hypothesis of no spatial autocorrelation does not imply the alternative of spatial error autocorrelation, which is typically how this result is incorrectly interpreted. Specifically, Moran's I also has considerable power against a spatial lag alternative, so rejection of the null does not provide any guidance in the choice of a spatial error vs. a spatial lag as the alternative spatial regression specification.

5.1.2 Kelijan and Prucha (2001) Derivation of Moran's I

More recently, [Kelejian and Prucha \(2001\)](#) have criticized Moran's I measure, arguing that the normalizing factor used by [Cliff and Ord \(1972\)](#) to derive its expected value and the variance under the null of no spatial correlation is not theoretically justified. In fact, the denominator of (5.2) represents the estimator of the standard deviation of the quadratic form appearing in the numerator and this can be proved to be inconsistent. With this motivation, [Kelejian and Prucha \(2001\)](#) proposed a different normalizing factor that removes this inconsistency and achieves the aim of normalizing the variance to unity. The Moran's I they proposed is the following:

$$\bar{I} = \frac{\hat{\boldsymbol{\varepsilon}}^\top \mathbf{W} \hat{\boldsymbol{\varepsilon}}}{\tilde{\sigma}^2}, \quad (5.5)$$

with $\tilde{\sigma}^2$ being normalizing factor that depends on the particular model chosen as an alternative hypothesis. In particular, if the alternative hypothesis is constituted by a SEM, the normalizing factor assumes the expression:

$$\tilde{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}} \left\{ \text{tr} \left[(\mathbf{W}^\top + \mathbf{W}) \mathbf{W} \right] \right\}^{-1/2}}{n}. \quad (5.6)$$

As a consequence the test statistic can be defined as:

$$\bar{I} = \frac{n \hat{\boldsymbol{\varepsilon}}^\top \mathbf{W} \hat{\boldsymbol{\varepsilon}}}{\hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}} \left\{ \text{tr} \left[(\mathbf{W}^\top + \mathbf{W}) \mathbf{W} \right] \right\}^{-1/2}}. \quad (5.7)$$

The two expressions reported in Equations (5.2) and (5.7) coincide if the weight matrix has dichotomous entries in which case $w_{ij} = w_{ij}^2$ and, therefore,

$$\sum_i \sum_j w_{ij} = \left\{ \text{tr} \left[(\mathbf{W}^\top + \mathbf{W}) \mathbf{W} \right] \right\}^{-1/2}.$$

In their paper, [Kelejian and Prucha \(2001\)](#) prove that the modified Moran test \bar{I} converges in distribution to a standardized normal distribution even when the priori assumption of the normality of the error is not satisfied. Even if in large samples $\bar{I} \sim N(0, 1)$, in small samples its expected value and variance may be different.

5.1.3 Example

We will continue here with [Anselin \(1988\)](#)'s example (see Section 4.6) and we analyze whether the regression residuals from a OLS model show evidence of some spatial autocorrelation.

To carry out the Moran's I test on the residuals in R we need to pass the regression object and spatial weight object (`listw`) to the `lm.morantest` function.

```
# Moran test for residuals
library("spdep")
# Load data
columbus <- readShapePoly(system.file("etc/shapes/columbus.shp",
                                     package = "spdep")[1])
col.gal.nb <- read.gal(system.file("etc/weights/columbus.gal",
```

```

                                package = "spdep")[1])
listw <- nb2listw(col.gal.nb, style = "W")
ols <- lm(CRIME ~ INC + HOVAL,
          data = columbus)
lm.morantest(ols, listw = listw, alternative = "two.sided")

##
## Global Moran I for regression residuals
##
## data:
## model: lm(formula = CRIME ~ INC + HOVAL, data = columbus)
## weights: listw
##
## Moran I statistic standard deviate = 2.681, p-value = 0.00734
## alternative hypothesis: two.sided
## sample estimates:
## Observed Moran I      Expectation      Variance
##      0.212374153      -0.033268284      0.008394853

```

The default setting in this function is to compute the p-value for one sided test. To get a two-sided test, the `alternative` argument must be specified explicitly.

The results show a Moran's I statistic of 0.212, which is highly significant and reject the null hypothesis of uncorrelated error terms.

Recall that the Moran's I statistic has high power against a range of alternatives. However, it does not provide much help in terms of which alternative model would be most appropriate.

5.2 Common Factor Hypothesis

The SEM model can be expanded and rewritten as follows:

$$\begin{aligned}
 \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + (\mathbf{I}_n - \lambda\mathbf{W})^{-1}\boldsymbol{\varepsilon} \\
 (\mathbf{I}_n - \lambda\mathbf{W})\mathbf{y} &= (\mathbf{I}_n - \lambda\mathbf{W})\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\
 \mathbf{y} - \lambda\mathbf{W}\mathbf{y} &= (\mathbf{X} - \lambda\mathbf{W}\mathbf{X})\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\
 \mathbf{y} &= \lambda\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} - \mathbf{W}\mathbf{X}(\lambda\boldsymbol{\beta}) + \boldsymbol{\varepsilon}
 \end{aligned} \tag{5.8}$$

resulting in a model including not only the spatially lagged dependent variable, $\mathbf{W}\mathbf{y}$, but also the spatially lagged explanatory variables ($\mathbf{W}\mathbf{X}$). Under some nonlinear restrictions we can see that (5.8) is equivalent to the SDM. The unconstrained form of the model—or the SDM model—is

$$\mathbf{y} = \gamma_1 \mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\gamma}_2 + \mathbf{W}\mathbf{X}\boldsymbol{\gamma}_3 + \boldsymbol{\varepsilon}, \tag{5.9}$$

where γ_1 is a scalar, $\boldsymbol{\gamma}_2$ is a $K \times 1$ vector (where K is the number of explanatory variables, including the constant), and $\boldsymbol{\gamma}_3$ is also a $K \times 1$ vector. Note that if $\boldsymbol{\gamma}_3 = -\gamma_1\boldsymbol{\gamma}_2$, then the SDM is equivalent to the SEM model. Note also that $\boldsymbol{\gamma}_3 = -\gamma_1\boldsymbol{\gamma}_2$ is a vector of $K \times 1$ nonlinear constraints of the form:

$$\gamma_{3,k} = -\gamma_1\gamma_{2,k}, \quad \text{for } k = 1, \dots, K. \quad (5.10)$$

These conditions are usually formulated as a null hypothesis, designated as the **Common Factor Hypothesis**, and written as:

$$H_0 : \gamma_3 + \gamma_1\gamma_2 = \mathbf{0}. \quad (5.11)$$

If the constraints hold it follows that the SDM is equivalent to the SEM model.

5.3 Hausman Test: OLS vs SEM

As we explained in Section 4.3, OLS estimates for the parameters β will be unbiased if the underlying DGP represents the SEM model, but standard errors from least-squares are biased. Since we are comparing two models that provide consistent estimates, but one is more efficient than the other, we can perform a Hausman test (Pace and LeSage, 2008).

The idea behind the Hausman test is to compare two set of estimators that are consistent, but one of them is more efficient. Let $\hat{\beta}_{OLS}$ and $\hat{\beta}_{SEM}$ the estimated parameters with OLS and for the SEM model estimated, for example, via MLE. Then a natural test is to consider the difference between the two estimators: $\hat{q} = \hat{\beta}_{OLS} - \hat{\beta}_{SEM}$. If the difference is ‘large’, then there exists evidence against the $H_0 : \hat{\beta}_{OLS} = \hat{\beta}_{SEM}$ suggesting misspecification and then the SEM model is more appropriate. If we cannot reject the null, it would be an indicator that spatially correlated omitted variables do not represent a problem or are not correlated with the explanatory variables.

The following definition provides the statistic and asymptotic distribution for the Hausman test.

Definition 5.3.1 — Hausman Test. Let $\hat{\beta}_{OLS}$ and $\hat{\beta}_{SEM}$ be OLS and SEM estimators. Define $\hat{q} = \hat{\beta}_{OLS} - \hat{\beta}_{SEM}$, and

$$\mathbb{V}(\hat{q}) = \mathbb{V}(\hat{\beta}_{OLS}) - \mathbb{V}(\hat{\beta}_{SEM}). \quad (5.12)$$

Then the Hausman statistic:

$$H = \hat{q}^\top (\mathbb{V}(\hat{q}))^{-1} \hat{q}, \quad (5.13)$$

is distributed asymptotically chi-square with $\#\beta$ degrees of freedom.

The estimated variance-covariance matrix $\hat{\beta}_{SEM}$ is given by (see Equation 4.50):

$$\mathbb{V}(\hat{\beta}_{SEM}) = \hat{\sigma}^2 \left[\mathbf{X}^\top (\mathbf{I}_n - \lambda \mathbf{W})^\top (\mathbf{I}_n - \lambda \mathbf{W}) \mathbf{X} \right]^{-1}. \quad (5.14)$$

However, as shown by Cordy and Griffith (1993), the usual OLS variance-covariance matrix $\sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$ is inconsistent under the null of a spatial error DGP. A consistent estimator of the OLS variance-covariance matrix under the spatial error DGP can be obtained as follows. Under the SEM model, the sampling error for the OLS estimator is:

$$\begin{aligned}
\hat{\beta}_{OLS} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \\
&= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top [\mathbf{X}\beta_0 + (\mathbf{I} - \lambda \mathbf{W}) \boldsymbol{\varepsilon}] \\
\hat{\beta}_{OLS} - \beta_0 &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{B} \boldsymbol{\varepsilon}
\end{aligned}$$

where $\mathbf{B} = (\mathbf{I} - \lambda \mathbf{W})$. Taking expectation, we get:

$$\begin{aligned}
\mathbb{E} [\hat{\beta}_{OLS} - \beta_0] &= \mathbb{E} [(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{B} \boldsymbol{\varepsilon}] \\
&= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{B} \mathbb{E}(\boldsymbol{\varepsilon}) \\
&= \mathbf{0}
\end{aligned}$$

So the OLS estimator is unbiased. For the variance, we obtain:

$$\begin{aligned}
\mathbb{V}(\hat{\beta}_{OLS}) &= \mathbb{E} [\hat{\beta} - \mathbb{E}(\hat{\beta})]^2 \\
&= \mathbb{E} [(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{B} \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top \mathbf{B}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}] \\
&= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{B} \mathbb{E} [\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top] \mathbf{B}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \\
&= \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{B} \mathbf{B}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}
\end{aligned} \tag{5.15}$$

Under the null of the spatial error process, the ML estimate $\hat{\sigma}^2$, based on the the variance of the residuals from the SEM provides a consistent estimate of σ^2 . The ML estimate $\hat{\lambda}$ provides a consistent estimate of λ . With these estimates, we can compute the variance of the OLS estimates as in Equation (5.15) (Pace and LeSage, 2008).

5.4 Tests Based on ML

In the previous section we shown how to perform a Moran's I test to assess whether the residuals present evidence of spatial autocorrelation. However, in this section we first estimate a spatial model and then we conduct **inference**. Thus, we will write the null hypothesis as a restriction on a subset of the parameter vector $\boldsymbol{\theta}$. Specifically, we would like to test whether $H_0 : \rho = 0$ or $H_0 : \lambda = 0$.

We begin our discussion of the hypothesis tests by describing the ML trinity: the Wald, Likelihood Ratio (LR), and Lagrange Multiplier (LM) test. These tests can be thought of as a comparison between the estimates obtained after the constraints implied by the hypothesis have been imposed to the estimates obtained without the constraints.

5.4.1 Likelihood Ratio Test

The likelihood ratio test is used to compare the difference between the value of the log-likelihood of a specification considered to be unconstrained and the value of log-likelihood obtained for a constrained model specification.

We define the constrained estimate as:

$$\tilde{\boldsymbol{\theta}}_{ML} = \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{ \frac{1}{n} \sum_{i=1}^n \log L(\boldsymbol{\theta}) \right\} \quad \text{s.t.} \quad \rho = 0 \quad (5.16)$$

or

$$\tilde{\boldsymbol{\theta}}_{ML} = \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{ \frac{1}{n} \sum_{i=1}^n \log L(\boldsymbol{\theta}) \right\} \quad \text{s.t.} \quad \lambda = 0 \quad (5.17)$$

and the unconstrained estimate as:

$$\hat{\boldsymbol{\theta}}_{ML} = \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{ \frac{1}{n} \sum_{i=1}^n \log L(\boldsymbol{\theta}) \right\} \quad (5.18)$$

Definition 5.4.1 — Likelihood Ratio Test. The Likelihood Ratio (LR) Test is formally defined as:

$$LR = 2 \cdot n \left(\frac{1}{n} \sum_{i=1}^n \log L(\hat{\boldsymbol{\theta}}) - \frac{1}{n} \sum_{i=1}^n \log L(\tilde{\boldsymbol{\theta}}) \right) \xrightarrow{d} \chi^2(r) \quad (5.19)$$

where r is the number of constraints.

The number of constraints imposed may vary depending on the specifications. In spatial models, the number of constraints is generally one or two, since we have the restriction $\rho = 0$, $\lambda = 0$, or $\lambda = \rho = 0$.

The likelihood ratio test is designed to evaluate the distance that separates the values of the two likelihoods: if the distance is small, then the constrained model is comparable to the unconstrained model. In this case, the constraint version is “acceptable” and do not reduce the performance of the model. It is thus statistically possible to not reject the null hypothesis (the postulated constraints prove to be credible). In other words, if the likelihood value of an unconstrained model strays too far from the constrained model, we cannot accept the null hypothesis: the gap is too large for the constraint to be consider realistic.

LR for the SLM

Note that the log-likelihood for the unconstrained model—that is the model for which $\rho \neq 0$ —is:

$$\log L(\boldsymbol{\theta}) = \log |\mathbf{A}| - \frac{n \log(2\pi)}{2} - \frac{n \log(\sigma^2)}{2} - \frac{1}{2\sigma^2} (\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (5.20)$$

The log-likelihood for the constrained model is found by setting $\rho = 0$ in Equation (5.20). Recall that if $\rho = 0$, then $\mathbf{A} = \mathbf{I} - \rho\mathbf{W} = \mathbf{I}$, then:

$$\log L(\boldsymbol{\theta}) = -\frac{n \log(2\pi)}{2} - \frac{n \log(\sigma^2)}{2} - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (5.21)$$

Therefore, following our definition in Equation (5.19):

$$\begin{aligned}
LR &= 2(\log L(\hat{\boldsymbol{\theta}}) - \log L(\tilde{\boldsymbol{\theta}})) \\
&= 2 \left[\log |\mathbf{A}| + \frac{1}{2\sigma^2} \left((\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - (\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right) \right] \\
&= 2 \log |\mathbf{A}| + \frac{1}{\sigma^2} \left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - (\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right]
\end{aligned} \tag{5.22}$$

with the coefficients respectively evaluated at their restricted and unrestricted estimates. The resulting test statistic is asymptotically distributed as χ^2 with 1 degree of freedom, or, alternatively, its square root is distributed as a standard normal variate.

LR for the SEM

Note that the log-likelihood for the unconstrained model—that is the model for which $\lambda \neq 0$ —is:

$$\log L(\boldsymbol{\theta}) = \log |\mathbf{B}| - \frac{n \log(2\pi)}{2} - \frac{n \log(\sigma^2)}{2} - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Omega}(\lambda) (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{5.23}$$

Then the LR for the SEM model is:

$$LR = 2 \log |\mathbf{B}| + \frac{1}{\sigma^2} \left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Omega}(\lambda) (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right] \tag{5.24}$$

which is also distributed as $\chi^2(1)$. We can use the formulae above or use the following algorithm:

Algorithm 5.2 — LR Test. To compute the test statistic LR,

- (a) compute the restricted MLE $\tilde{\boldsymbol{\theta}}$ and record the value of the log-likelihood function at convergence $\log L(\tilde{\boldsymbol{\theta}})$,
- (b) compute the unrestricted MLE $\hat{\boldsymbol{\theta}}$ and record the value of the log-likelihood function at convergence $\log L(\hat{\boldsymbol{\theta}})$,
- (c) and compute,

$$LR = 2 \left[\log L(\hat{\boldsymbol{\theta}}) - \log L(\tilde{\boldsymbol{\theta}}) \right]$$

This statistic is always positive because the unrestricted maximum value always exceeds the restricted one.

- (d) Compare LR with the critical value of chi-square distribution with 1 degrees of freedom.

5.4.2 Wald Test

This approach is based on the comparison of the distances between the estimated parameters in constrained and unconstrained form. Thus, this idea suggest that, if the distance between

the parameter estimates $\hat{\beta}$ and $\tilde{\theta}$ is too high, the data fail to support the null hypothesis. In such circumstances, the null hypothesis cannot be accepted.

Formally, the Wald test proposes to calculate the distance between unconstrained estimators and the constrained estimators. This distance can be expressed by $(\hat{\theta} - \tilde{\theta})^2$ and is influenced by the shape of the likelihood curve.

The Wald statistic is distributed asymptotically according to a χ_r^2 with r degrees of freedom, where r represents the number of constraints tested. A large value of W means that the null hypothesis should be rejected, and, conversely, a small value suggests non-rejection of the null hypothesis.

The Wald test commonly uses unconstrained model estimates for evaluating the statistical value of W . Thus, the researcher needs to estimate only the unconstrained model for hypothesis testing. This is different from the likelihood ratio test where both unconstrained and constrained models need to be estimated in order to compare their likelihoods.

Definition 5.4.2 — The Wald Test. Assume that we have r nonlinear restrictions (which includes linear restriction as special case):

$$\mathbf{r}(\boldsymbol{\theta}_0) = \mathbf{0}$$

Let also

$$\mathbf{R}(\boldsymbol{\theta}) = \frac{\partial \mathbf{r}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^\top}$$

The Wald test is given by:

$$W = n \cdot \mathbf{r}(\hat{\boldsymbol{\theta}})^\top [\mathbf{R}(\hat{\boldsymbol{\theta}}) \hat{\mathbf{V}} \mathbf{R}(\hat{\boldsymbol{\theta}})^\top]^{-1} \mathbf{r}(\hat{\boldsymbol{\theta}}) \xrightarrow{d} \chi^2(r) \quad (5.25)$$

where r is the number of constraints.

Wald Test for SLM

The W statistic is:

$$W_\rho = \frac{\hat{\rho}^2}{\hat{\mathbb{V}}(\rho)} \quad (5.26)$$

where $\hat{\mathbb{V}}(\rho)$ can be obtained from Equation ?? as:

$$\hat{\mathbb{V}}(\rho) = \left[\text{tr}(\mathbf{C}^s \mathbf{C}) + \frac{1}{\sigma^2} (\mathbf{C} \mathbf{X} \boldsymbol{\beta})^\top (\mathbf{C} \mathbf{X} \boldsymbol{\beta}) \right]^{-1} \quad (5.27)$$

Clearly,

$$\frac{\rho}{se(\rho)} \stackrel{a}{\sim} N(0, 1) \quad (5.28)$$

with $se(\rho)$ as the estimated standard deviation.

Extensions to hypotheses that consists of linear and nonlinear combinations of model parameters can be obtained in a straightforward way. Computationally, the W —and LR —is more demanding since they require ML estimation under the alternative, and the explicit forms of the tests are more complicated.

Wald Test for SEM

The W statistic is:

$$W_\lambda = \frac{\hat{\lambda}^2}{\hat{V}(\lambda)} \quad (5.29)$$

where $\hat{V}(\lambda)$ can be obtained from Equation 4.50 as:

$$\hat{V}(\lambda) = \left[-\frac{\text{tr}(\mathbf{W}_B)}{\sigma^2} + \text{tr}(\mathbf{W}_B)^2 + \text{tr}(\mathbf{W}_B^\top \mathbf{W}_B) \right]^{-1} \quad (5.30)$$

Algorithm 5.3 — Wald Test. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top)^\top$. In general, to compute the Wald test statistic for $H_0 : \boldsymbol{\theta}_{02} = \mathbf{0}$,

- (a) compute the unrestricted MLE $\hat{\boldsymbol{\theta}}$,
- (b) compute an estimator of the variance matrix of the asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$, for example, the information $\mathbf{I}(\hat{\boldsymbol{\theta}})^{-1}$,
- (c) and finally compute the quadratic form:

$$W = n \cdot \hat{\boldsymbol{\theta}}^\top \hat{\mathbf{V}}_w^{-1} \hat{\boldsymbol{\theta}} \quad (5.31)$$

where $\hat{\mathbf{V}}_w$ is the $(2, 2)$ block of $\mathbf{I}(\hat{\boldsymbol{\theta}})^{-1}$ partitioned conformably with $\boldsymbol{\theta}$: that is

$$\hat{\mathbf{V}}_w = \left\{ \mathbf{I}_{22}(\hat{\boldsymbol{\theta}}) - \mathbf{I}_{21}(\hat{\boldsymbol{\theta}}) [\mathbf{I}_{11}(\hat{\boldsymbol{\theta}})]^{-1} \mathbf{I}_{12}(\hat{\boldsymbol{\theta}}) \right\}^{-1} \quad (5.32)$$

which is a **consistent estimator** of the asymptotic variance of $\hat{\boldsymbol{\theta}}_2$.

- (d) Compare W with the critical value of chi-square distribution with $K - r$ degrees of freedom.

5.4.3 Lagrange Multiplier Test

This approach is also based on the log-likelihood function curve, with the slope of the likelihood function being evaluated by the constraint type. The idea is that when the constraints are verified, the value of the estimated parameters $\boldsymbol{\theta}_0$ is such that the likelihood function slope at this point is zero. The goal is to compare, whether the slope evaluated using the constrained model is zero or strays too far from 0. In the last case, the null hypothesis must be rejected.

The Lagrange Multiplier test (or just score test) is based on the restricted model instead of the unrestricted model. Suppose that we maximize the log-likelihood subject to the set of constraints

Theorem 5.4 — Lagrange Multiplier Test. The Lagrange multiplier test statistic is:

$$LM = \left(\frac{\partial \log L(\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}} \right)^\top [\mathbf{I}(\tilde{\boldsymbol{\theta}})]^{-1} \left(\frac{\partial \log L(\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}} \right) \xrightarrow{d} \chi(r) \quad (5.33)$$

Under the null hypothesis, LM has a limiting chi-square distribution with degrees of freedom equal to the number of restrictions. All terms are computed at the restricted estimator.

The main advantage of the LM statistic is that it only requires the constrained model to be estimated, and it is very often less complex since it mainly lies on the OLS. This is one of the reasons that has lead to the widespread use of this approach.

LM statistical test construction depends on the postulated specification of the spatial autoregressive DGP: SEM or SLM. The usual practice is to initially use a general test for detecting residual spatial autocorrelation (Moran's I test for example) in order to then be able to carry out the statistical LM test to identify the specific type of the autoregressive process.

Test for SEM

This test, proposed by Burridge assumes the omission of a spatial autoregressive process of the error term u_i , where $u_i = \lambda \sum_j w_{ij} u_j + \epsilon_i$. The null hypothesis is $H_0 : \lambda = 0$. The constrained version of the SEM model can be reduced to a standard linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

For the SEM model we need to find the score function of the log-likelihood for the constrained model. Note that

$$\begin{aligned} \frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} &= \frac{1}{\sigma^2} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top \mathbf{B}(\lambda)^\top \mathbf{B}(\lambda) \mathbf{X} \\ \frac{\partial \log L(\boldsymbol{\theta})}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top \mathbf{B}(\lambda)^\top \mathbf{B}(\lambda) (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ \frac{\partial \log L(\boldsymbol{\theta})}{\partial \lambda} &= -\text{tr}(\mathbf{B}^{-1} \mathbf{W}) + \frac{1}{\sigma^2} [\boldsymbol{\epsilon}^\top \mathbf{W} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})] \end{aligned} \quad (5.34)$$

Under the null hypothesis $H_0 : \lambda = 0$, we get:

$$\begin{aligned} \left. \frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} \right|_{\lambda=0} &= \frac{1}{\sigma^2} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top \mathbf{I}_n^\top \mathbf{I}_n \mathbf{X} = \frac{1}{\sigma^2} \hat{\boldsymbol{\epsilon}}_{OLS}^\top \mathbf{X} \\ \left. \frac{\partial \log L(\boldsymbol{\theta})}{\partial \sigma^2} \right|_{\lambda=0} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \hat{\boldsymbol{\epsilon}}_{OLS}^\top \hat{\boldsymbol{\epsilon}}_{OLS} \\ \left. \frac{\partial \log L(\boldsymbol{\theta})}{\partial \lambda} \right|_{\lambda=0} &= \frac{\boldsymbol{\epsilon}^\top \mathbf{W} \boldsymbol{\epsilon}}{\sigma^2} \end{aligned} \quad (5.35)$$

The test is essentially based on the score with respect to λ , i.e., on

$$s_\lambda = \left. \frac{\partial \log L(\boldsymbol{\theta})}{\partial \lambda} \right|_{\lambda=0} = \frac{\boldsymbol{\epsilon}^\top \mathbf{W} \boldsymbol{\epsilon}}{\sigma^2} \quad (5.36)$$

Recall that:

$$\text{AsyVar}(\boldsymbol{\beta}, \sigma^2, \lambda) = \begin{pmatrix} \frac{\mathbf{X}(\lambda)^\top \mathbf{X}(\lambda)}{\sigma^2} & 0 & 0 \\ k \times k & & \\ 0 & \frac{n}{2\sigma^4} & \frac{\text{tr}(\mathbf{W}_B)}{\sigma^2} \\ 0 & \frac{\text{tr}(\mathbf{W}_B)}{\sigma^2} & \text{tr}(\mathbf{W}_B)^2 + \text{tr}(\mathbf{W}_B^\top \mathbf{W}_B) \end{pmatrix}^{-1} \quad (5.37)$$

where $\mathbf{W}_B = \mathbf{W}(\mathbf{I} - \lambda \mathbf{W})^{-1}$. Under the null, $\mathbb{E}_{H_0}(\partial^2 \ln L / \partial \boldsymbol{\beta} \partial \lambda) = \mathbf{0}$, and $\mathbb{E}_{H_0}(\partial^2 \ln L / \partial \sigma \partial \lambda) = \mathbf{0}$ because $\mathbb{E}(\boldsymbol{\varepsilon}^\top \mathbf{W} \boldsymbol{\varepsilon}) = \sigma^2 \text{tr}(\mathbf{W}) = \mathbf{0}$ as \mathbf{W} has a zero diagonal. Furthermore,

$$\mathbb{E}_{H_0} \left(\frac{\partial \log L(\boldsymbol{\theta})}{\partial \lambda^2} \right) = -\text{tr}(\mathbf{W}^2 + \mathbf{W}^\top \mathbf{W}) \quad (5.38)$$

Then the expression for the LM test for a SEM specification is:

$$LM_{ERR} = \frac{1}{C} \left(\frac{\hat{\boldsymbol{\varepsilon}}^\top \mathbf{W} \hat{\boldsymbol{\varepsilon}}}{\hat{\sigma}^2} \right)^2 \quad (5.39)$$

where $C = \text{tr}[(\mathbf{W} + \mathbf{W}^\top) \mathbf{W}]$. Therefore, the test requires only OLS estimates. Under the null hypothesis, this statistic converges asymptotically to a $\chi^2(1)$. For example, if we use a significance level of 95%, the critical value is 3.84. Thus, we reject the null hypothesis, if the value of the statistical test LM_{ERR} is greater than 3.84. We can conclude in this case that spatial autocorrelation is present in the standard linear model residuals and we must proceed to estimate the SEM specification.

Note also that it is similar in expression to Moran's I : except for the scaling factor T , this statistic is essentially the square of Moran's I .

Test for SLM

The LM test can also be used to detect whether the detected spatial autocorrelation among the residuals of the multiple regression does not rise from the omission of spatially lagged dependent variable regressors.

The null hypothesis of this test is based on the significance of the autoregressive parameter, $H_0 : \rho = 0$.

In this case:

$$s_{\rho=0} = \left. \frac{\partial \log L(\boldsymbol{\theta})}{\partial \rho} \right|_{\rho=0} = \frac{1}{\sigma^2} \boldsymbol{\varepsilon}^\top \mathbf{W} \mathbf{y} \quad (5.40)$$

The inverse of the information matrix is given in (??). The complicating feature of this matrix is that even under $\rho = 0$, it is not block diagonal; the $(\rho, \boldsymbol{\beta})$ term is equal to $(\mathbf{X}^\top \mathbf{W} \mathbf{X} \boldsymbol{\beta}) / \sigma^2$, obtained by inserting $\rho = 0$; i.e., $\mathbf{C} = \mathbf{W}$. The main problem of this is that even under $\rho = 0$, we cannot ignore one of the off-diagonal terms. This is not the case for $s_{\lambda=0}$. Asymptotic variance of $s_{\lambda=0}$ was obtained just using the (2,2) element of ?. For the spatial lag model, asymptotic variance of $s_{\rho=0}$ is obtained from the reciprocal of the last element of: ¹

¹This is obtained using partitioned Inversion.

$$\mathbb{V}(\boldsymbol{\beta}, \sigma^2, \rho) \Big|_{\rho=0} = \begin{pmatrix} \frac{1}{\sigma^2}(\mathbf{X}^\top \mathbf{X}) & \mathbf{0}' & \frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{W} \mathbf{X} \boldsymbol{\beta} \\ \cdot & \frac{n}{2\sigma^4} & \mathbf{0} \\ \cdot & \cdot & \text{tr}(\mathbf{W}^2 + \mathbf{W}^\top \mathbf{W}) + \frac{1}{\sigma^2} (\mathbf{W} \mathbf{X} \boldsymbol{\beta})^\top (\mathbf{W} \mathbf{X} \boldsymbol{\beta}) \end{pmatrix}^{-1}$$

Since under $\rho = 0$, $\mathbf{C} = \mathbf{W}$ and $\text{tr}(\mathbf{W}) = 0$. Recall that $T = \text{tr}[(\mathbf{W}^\top + \mathbf{W}) \mathbf{W}]$, then we can write:

$$LM_{SAR} = \frac{1}{T_1} \left(\frac{\hat{\boldsymbol{\varepsilon}} \mathbf{W} \mathbf{y}}{\hat{\sigma}^2} \right)^2 \quad (5.41)$$

where $T_1 = [(\mathbf{W} \mathbf{X} \hat{\boldsymbol{\varepsilon}})^\top \mathbf{M} (\mathbf{W} \mathbf{X} \hat{\boldsymbol{\varepsilon}}) + T \hat{\sigma}^2] / \hat{\sigma}^2$ with $\mathbf{M} = \mathbf{I} - \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$. Under the null hypothesis, the test asymptotically converges according to the χ^2 distribution to 1 degree of freedom.

5.4.4 Anselin and Florax Recipe

How to decide? For the simple case of choosing between a SLM or SEM alternative, there is evidence that the proper model is most likely the one with the largest significant LM test value (Anselin and Rey, 1991).

- When the test LM_{LAG} value is significant and the LM_{ERR} is insignificant, the most appropriate model is the SLM model;
- in the same vein, when the test LM_{ERR} is significant and the LM_{LAG} value is insignificant, the most appropriate model is the SEM model.

As you can guess, sometimes it is possible to find that both statistical test are significant. In this case, one decision rule can be as follows:

- when the test LM_{LAG} value is higher than the test LM_{ERR} value, it would be best to consider the SLM model;
- when the test LM_{ERR} value is higher than the test LM_{LAG} value, it would be best to consider the SEM model.

Of course, if both statistics are significant, it could also well be appropriate to estimate a general autoregressive model (SAC).

5.4.5 Lagrange Multiplier Test Statistics in R

Lagrange Multiplier tests, as well as their robust forms are included in the `lm.LMtests` function. An OLS regression object and a spatial `listw` object must be passed as arguments. In addition, the tests must be specified as a character vector as illustrated below.

```
# LM test
lm.LMtests(ols, listw,
            test = c("LMerr", "RLMerr", "LMlag", "RLMlag"))
```

```
##
##  Lagrange multiplier diagnostics for spatial dependence
##
## data:
## model: lm(formula = CRIME ~ INC + HOVAL, data = columbus)
## weights: listw
##
## LMerr = 4.6111, df = 1, p-value = 0.03177
##
##
##  Lagrange multiplier diagnostics for spatial dependence
##
## data:
## model: lm(formula = CRIME ~ INC + HOVAL, data = columbus)
## weights: listw
##
## RLMerr = 0.033514, df = 1, p-value = 0.8547
##
##
##  Lagrange multiplier diagnostics for spatial dependence
##
## data:
## model: lm(formula = CRIME ~ INC + HOVAL, data = columbus)
## weights: listw
##
## LMLag = 7.8557, df = 1, p-value = 0.005066
##
##
##  Lagrange multiplier diagnostics for spatial dependence
##
## data:
## model: lm(formula = CRIME ~ INC + HOVAL, data = columbus)
## weights: listw
##
## RLMlag = 3.2781, df = 1, p-value = 0.07021
```

Note that both `LMerr` and `LMLag` are significant. However, the robust statistics point to the lag model as the proper alternative. With this information in hand, we can select the spatial lag model as the proper model.

5.5 Exercises

Exercise 5.1 Example 1.

Appendix

5.A Asymptotic Properties of Moran's I

Let the model be:

$$\begin{aligned} \mathbf{y}_n &= \rho_0 \mathbf{M}_n \mathbf{y}_n + \mathbf{X}_n \boldsymbol{\beta}_0 + \mathbf{u}_n = \mathbf{D}_n \boldsymbol{\theta}_0 + \mathbf{u}_n \\ \mathbf{u}_n &= \lambda_0 \mathbf{M}_n \mathbf{u}_n + \boldsymbol{\varepsilon}_n \end{aligned} \quad (5.42)$$

Let $Q_n^* = \hat{\mathbf{u}}_n^\top \mathbf{W}_n \hat{\mathbf{u}}_n$. Since we need to let the model as a function of the true error \mathbf{u} , note that:

$$\begin{aligned} \hat{\mathbf{u}}_n &= \mathbf{y}_n - \mathbf{D}_n \hat{\boldsymbol{\theta}}_n \\ &= \mathbf{D}_n \boldsymbol{\theta}_0 + \mathbf{u}_n - \mathbf{D}_n \hat{\boldsymbol{\theta}}_n \quad \text{using (5.42)} \\ &= \mathbf{u}_n - \mathbf{D}_n (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \end{aligned} \quad (5.43)$$

Then:

$$\begin{aligned} \hat{\mathbf{u}}_n^\top \mathbf{W}_n \hat{\mathbf{u}}_n &= [\mathbf{u}_n - \mathbf{D}_n (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)]^\top \mathbf{W}_n [\mathbf{u}_n - \mathbf{D}_n (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)] \\ &= \mathbf{u}_n^\top \mathbf{W}_n \mathbf{u}_n - \mathbf{u}_n^\top \mathbf{W}_n \mathbf{D}_n (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) - (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top \mathbf{D}_n^\top \mathbf{W}_n \mathbf{u}_n + (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top \mathbf{D}_n^\top \mathbf{W}_n \mathbf{D}_n (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \\ &= \mathbf{u}_n^\top \mathbf{W}_n \mathbf{u}_n - \mathbf{u}_n^\top (\mathbf{W}_n + \mathbf{W}_n^\top) \mathbf{D}_n (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top \mathbf{D}_n^\top \mathbf{W}_n \mathbf{D}_n (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \end{aligned}$$

where in the last line we use the fact that $(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top \mathbf{D}_n^\top \mathbf{W}_n \mathbf{u}_n = \mathbf{u}_n^\top \mathbf{W}_n^\top \mathbf{D}_n (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ and \mathbf{W}_n is not necessarily symmetric. Multiplying the previous equation by $1/\sqrt{n}$, we obtain:

$$\frac{1}{\sqrt{n}} \hat{\mathbf{u}}_n^\top \mathbf{W}_n \hat{\mathbf{u}}_n = \frac{1}{\sqrt{n}} \mathbf{u}_n^\top \mathbf{W}_n \mathbf{u}_n - \frac{1}{n} \mathbf{u}_n^\top (\mathbf{W}_n + \mathbf{W}_n^\top) \mathbf{D}_n \sqrt{n} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top \left[\frac{1}{n} \mathbf{D}_n^\top \mathbf{W}_n \mathbf{D}_n \right] \sqrt{n} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \quad (5.44)$$

First, we will show that the last element of (5.44) converges to 0 as $n \rightarrow \infty$. Since $\mathbf{D}_n = [\mathbf{M}_n \mathbf{y}_n, \mathbf{X}_n]$ has bounded elements in absolute value, then

$$\frac{1}{n} \mathbf{D}_n^\top \mathbf{W}_n \mathbf{D}_n = O_p(1)$$

that is, it is stochastically bounded. Furthermore in Section 6.4.4, we show that the 2SLS estimator $\hat{\boldsymbol{\theta}}_n$ is consistent, so that $(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = o_p(1)$ and has a limiting distribution, that is:

$$\sqrt{n} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = O_p(1)$$

Thus using these two results, we can say that:

$$(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top \left[\frac{1}{n} \mathbf{D}_n^\top \mathbf{W}_n \mathbf{D}_n \right] \sqrt{n} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = o_p(1) \cdot O_p(1) \cdot O_p(1) = o_p(1)$$

R Let \mathbf{D} be any $n \times n$ matrix. Then, since $\mathbf{v}^\top \mathbf{D} \mathbf{v} = \mathbf{v}^\top \mathbf{D}^\top \mathbf{v}$, then we can write:

$$\mathbf{v}^\top \mathbf{D} \mathbf{v} = \mathbf{v}^\top \left[\frac{\mathbf{D} + \mathbf{D}^\top}{2} \right] \mathbf{v} \quad (5.45)$$

Using the previous remark, and noting that $\mathbf{u}_n = (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^{-1} \boldsymbol{\varepsilon}_n$, we can write the first element of Equation (5.44) as:

$$\begin{aligned} \mathbf{u}_n^\top \mathbf{W} \mathbf{u}_n &= \boldsymbol{\varepsilon}_n^\top (\mathbf{I}_n - \lambda_0 \mathbf{M}_n^\top)^{-1} \left[\frac{\mathbf{W}_n + \mathbf{W}_n^\top}{2} \right] (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^{-1} \boldsymbol{\varepsilon}_n \\ &= \boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n \end{aligned} \quad (5.46)$$

where $\mathbf{A}_n = (\mathbf{I}_n - \lambda_0 \mathbf{M}_n^\top)^{-1} \left[\frac{\mathbf{W}_n + \mathbf{W}_n^\top}{2} \right] (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^{-1}$ which is uniformly bounded in absolute value. According to Lemma 3.25 and assuming normality (which is much simpler) we can say that:

$$\begin{aligned} \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) &= \sigma_0^2 \text{tr}(\mathbf{A}_n) \\ \mathbb{V}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) &= \sigma_0^4 [\text{tr}(\mathbf{A}_n \mathbf{A}_n^\top) + \text{tr}(\mathbf{A}_n^2)] \end{aligned} \quad (5.47)$$

The the second part of (5.44), note that $\frac{1}{n} \mathbf{u}_n^\top (\mathbf{W}_n + \mathbf{W}_n^\top) \mathbf{D}_n = \boldsymbol{\varepsilon}_n^\top (\mathbf{I}_n - \lambda_0 \mathbf{M}_n^\top)^{-1} (\mathbf{W}_n + \mathbf{W}_n^\top) \mathbf{D}_n$ where $\mathbf{D}_n = [\mathbf{M}_n \mathbf{y}_n, \mathbf{X} \beta_0]$, and $\mathbf{y}_n = (\mathbf{I}_n - \rho_0 \mathbf{M}_n^\top)^{-1} \mathbf{X}_n \beta_0 + (\mathbf{I}_n - \rho_0 \mathbf{M}_n^\top)^{-1} (\mathbf{I}_n - \lambda_0 \mathbf{M}_n^\top)^{-1} \boldsymbol{\varepsilon}_n$, so that:

$$\begin{aligned} \boldsymbol{\varepsilon}_n^\top (\mathbf{I}_n - \lambda_0 \mathbf{M}_n^\top)^{-1} (\mathbf{W}_n + \mathbf{W}_n^\top) \mathbf{M} \mathbf{y}_n &= \boldsymbol{\varepsilon}_n^\top (\mathbf{I}_n - \lambda_0 \mathbf{M}_n^\top)^{-1} (\mathbf{W}_n + \mathbf{W}_n^\top) \mathbf{M} \left[(\mathbf{I}_n - \rho_0 \mathbf{M}_n^\top)^{-1} \mathbf{X}_n \beta_0 \right. \\ &\quad \left. + (\mathbf{I}_n - \rho_0 \mathbf{M}_n^\top)^{-1} (\mathbf{I}_n - \lambda_0 \mathbf{M}_n^\top)^{-1} \boldsymbol{\varepsilon}_n \right] \\ &= \boldsymbol{\varepsilon}_n^\top (\mathbf{I}_n - \lambda_0 \mathbf{M}_n^\top)^{-1} (\mathbf{W}_n + \mathbf{W}_n^\top) \mathbf{M} (\mathbf{I}_n - \rho_0 \mathbf{M}_n^\top)^{-1} \mathbf{X}_n \beta_0 \\ &\quad + \boldsymbol{\varepsilon}_n^\top (\mathbf{I}_n - \lambda_0 \mathbf{M}_n^\top)^{-1} (\mathbf{W}_n + \mathbf{W}_n^\top) \mathbf{M} (\mathbf{I}_n - \rho_0 \mathbf{M}_n^\top)^{-1} (\mathbf{I}_n - \lambda_0 \mathbf{M}_n^\top)^{-1} \boldsymbol{\varepsilon}_n \\ &= \boldsymbol{\varepsilon}_n^\top \mathbf{B}_n^* + \boldsymbol{\varepsilon}_n^\top \mathbf{C}_n^* \boldsymbol{\varepsilon}_n \end{aligned}$$

where \mathbf{B}_n^* is a nonstochastic vector whose elements are uniformly bounded in absolute value, and where \mathbf{C}_n^* is a nonstochastic matrix whose row and columns sums are uniformly bounded in absolute value. Note that

$$\mathbf{d}_n^\top = \mathbb{E} \left[\frac{1}{n} \mathbf{u}_n^\top (\mathbf{W}_n + \mathbf{W}_n^\top) \mathbf{D}_n \right] = O(1) \quad (5.48)$$

since:

$$\begin{aligned} \mathbb{E} \left(\frac{1}{n} \mathbf{B}_n^{*\top} \boldsymbol{\varepsilon}_n \right) &= 0 \\ \mathbb{V} \left(\frac{1}{n} \mathbf{B}_n^{*\top} \boldsymbol{\varepsilon}_n \right) &= \sigma^2 \frac{1}{n^2} \mathbf{B}_n^{*\top} \mathbf{B}_n^* = o(1) \\ \mathbb{E} \left(\frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{C}_n^* \boldsymbol{\varepsilon}_n \right) &= \sigma^2 \frac{1}{n} \text{tr}(\mathbf{C}_n^*) = O(1) \\ \mathbb{V} \left(\frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{C}_n^* \boldsymbol{\varepsilon}_n \right) &= \sigma_0^4 [\text{tr}(\mathbf{A}_n \mathbf{A}_n^\top) + \text{tr}(\mathbf{A}_n^2)] = o(1) \end{aligned}$$

Then:

$$\mathbb{V} \left(\frac{1}{n} \mathbf{u}_n^\top (\mathbf{W}_n + \mathbf{W}_n^\top) \mathbf{D}_n \right) = o(1) \quad (5.49)$$

and the Claim in Equation (5.48) follows by Chebychev's inequality. Thus, we can write:

$$\begin{aligned} \frac{1}{n} \mathbf{u}_n^\top (\mathbf{W}_n + \mathbf{W}_n^\top) \mathbf{D}_n \sqrt{n} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) &= \frac{1}{n} \mathbf{u}_n^\top (\mathbf{W}_n + \mathbf{W}_n^\top) \mathbf{D}_n \sqrt{n} \left[\frac{1}{\sqrt{n}} \mathbf{P}_n \mathbf{F}_n^\top \boldsymbol{\varepsilon}_n + o_p(1) \right] \\ &= \frac{1}{\sqrt{n}} = \mathbf{b}_n^\top \boldsymbol{\varepsilon}_n + o_p(1) \end{aligned} \quad (5.50)$$

where $\mathbf{b}_n^\top = -\mathbf{d}_n^\top \mathbf{P} \mathbf{F}_n^\top$. Finally, we can write Equation (5.44) as

$$\frac{1}{\sqrt{n}} \hat{\mathbf{u}}_n^\top \mathbf{W}_n \hat{\mathbf{u}}_n = \frac{1}{\sqrt{n}} \left[\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n + \mathbf{b}_n^\top \boldsymbol{\varepsilon}_n \right] + o_p(1) \quad (5.51)$$

Thus, the asymptotic distribution of the Moran's I statistic is based on estimated disturbances involves the large sample distributioon of a linear-quadratic form in innovations.

Instrumental Variables and GMM

One of the main disadvantages of the MLE is that it may be computational intensive when the number of spatial units is large. This procedure requires the manipulation of $n \times n$ matrices, such as the matrix multiplication, matrix inversion, the computation of characteristics roots and so on.

In this chapter, we will study the instrumental variables and the generalized method of moments method (IV/GMM). One of the reason for developing IV/GMM estimators was a response to perceived computational difficulties of the ML method (Kelejian and Prucha, 1998, 1999). Unlike ML, the IV/GMM procedure does not require the computation of the Jacobian, and does not rely on the normality assumption.

6.1 A Review of GMM

Before explaining the estimation procedure for the SLM, SEM and SAC model, we review some aspects of the GMM procedure in the spatial context. This section is heavily based on Prucha (2014).

6.1.1 Model Specification

Suppose the data are generated from a model

$$f(y_{in}, \mathbf{x}_{in}, \boldsymbol{\theta}_0) = u_{in} \quad i = 1, \dots, n.,$$

where $f(y_{in}, \mathbf{x}_{in}, \boldsymbol{\theta}_0)$ might represent a system of spatial equations, y_{in} is the dependent variable corresponding to unit i , \mathbf{x}_{in} is a vector of explanatory variables, u_{in} is a disturbance term, $\boldsymbol{\theta}_0$ is the $k \times 1$ unknown parameter vector, and $f(\cdot)$ is a known function.

Also, assume that there exists a $1 \times s$ vector of instruments \mathbf{h}_{in} and let w_{in} be the vector of all observables variables, including instruments, pertaining to the i th unit. For simplicity, assume that the disturbances are i.i.d. $(0, \sigma^2)$ and that the instruments are non-stochastic (those assumptions can be relaxed). Note that we are considering a triangular array since the variables are indexed by n . In particular the explanatory variables could be of the form $\mathbf{x}_{in} = [\mathbf{x}_i, \bar{\mathbf{x}}_{in}, \bar{y}_{in}]$ where \mathbf{x}_i is some exogenous explanatory variable, and $\bar{\mathbf{x}}_{in} = \sum_j w_{ij} \mathbf{x}_j$ and $\bar{y}_{in} = \sum_j w_{ij} y_{jn}$ are spatial lags, where w_{ij} denote spatial weights with $w_{ii} = 0$.

Suppose that there exists a vector $s \times 1$ of sample moments

$$\mathbf{g}_n(\boldsymbol{\theta}) = \mathbf{g}_n(w_1, \dots, w_n, \boldsymbol{\theta}) = \begin{pmatrix} g_{1,n}(w_1, \dots, w_n, \boldsymbol{\theta}) \\ \vdots \\ g_{s,n}(w_1, \dots, w_n, \boldsymbol{\theta}) \end{pmatrix},$$

with $s \geq k$ (for identification), and suppose that

$$\mathbb{E}[\mathbf{g}_n(w_1, \dots, w_n, \boldsymbol{\theta})] = \mathbf{0} \iff \boldsymbol{\theta} = \boldsymbol{\theta}_0,$$

that is, the model is identified. Let $\boldsymbol{\mathcal{R}}$ be some $s \times s$ symmetric positive semidefinite weighting matrix, then the corresponding GMM estimator is defined as:

$$\hat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \mathbf{g}_n(w_1, \dots, w_n, \boldsymbol{\theta})_{(1 \times s)}^\top \boldsymbol{\mathcal{R}}_{(s \times s)} \mathbf{g}_n(w_1, \dots, w_n, \boldsymbol{\theta})_{(s \times 1)}. \quad (6.1)$$

If $s = k$ the weighting matrix is irrelevant and $\hat{\boldsymbol{\theta}}_n$ can be found as a solution to the moment condition:

$$\mathbf{g}_n(w_1, \dots, w_n, \hat{\boldsymbol{\theta}}) = \mathbf{0}. \quad (6.2)$$

The classical GMM literature exploits **linear moment conditions** of the form:

$$\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n \mathbf{h}_i^\top u_i \right] = \mathbf{0},$$

which holds since $\mathbb{E}[\mathbf{h}_i^\top u_i] = \mathbf{h}_i^\top \mathbb{E}[u_i] = \mathbf{0}$ under the maintained assumptions. The spatial literature frequently considers **quadratic** moment conditions. Let \mathbf{A}_q , with element $[a_{ijq}]$ be some $n \times n$ matrix with $\operatorname{tr}(\mathbf{A}_q) = 0$, and assume for ease of exposition that \mathbf{A}_q is non-stochastic. Then the quadratic moment conditions considered in the spatial literature are of the form:

$$\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n a_{ijq} u_i u_j \right] = \mathbf{0}, \quad (6.3)$$

which clearly holds under the maintained assumptions. To see this, let $\mathbf{u} = [u_1, \dots, u_n]^\top$, then the moment conditions in (6.3) can be rewritten as:

$$\mathbb{E} \left[\frac{\mathbf{u}^\top \mathbf{A}_q \mathbf{u}}{n} \right] = \operatorname{tr} \left[\frac{\mathbf{A}_q \mathbb{E}(\mathbf{u} \mathbf{u}^\top)}{n} \right] = \sigma^2 \frac{\operatorname{tr}(\mathbf{A}_q)}{n} = \mathbf{0},$$

since under the maintained assumptions $\mathbb{E}[\mathbf{u} \mathbf{u}^\top] = \sigma^2 \mathbf{I}_n$ and $\operatorname{tr}(\mathbf{A}_q) = 0$.

Now let $\boldsymbol{\theta}_0 = [\lambda_0, \boldsymbol{\delta}_0]^\top$ and suppose the sample moment vector in (6.2) can be decomposed into:

$$\mathbf{g}_n(w_1, \dots, w_n, \boldsymbol{\theta}) = \begin{pmatrix} \mathbf{g}_n^\lambda(w_1, \dots, w_n, \lambda, \boldsymbol{\delta}) \\ \mathbf{g}_n^\delta(w_1, \dots, w_n, \lambda, \boldsymbol{\delta}) \end{pmatrix},$$

where λ is, for example, the spatial autoregressive parameter and $\boldsymbol{\delta}$ is the rest of parameters in the model, such that:

$$\begin{aligned}\mathbb{E} \left[\mathbf{g}_n^\lambda(w_1, \dots, w_n, \lambda, \boldsymbol{\delta}) \right] &= \mathbf{0} \iff \lambda = \lambda_0, \\ \mathbb{E} \left[\mathbf{g}_n^\delta(w_1, \dots, w_n, \lambda, \boldsymbol{\delta}) \right] &= \mathbf{0} \iff \boldsymbol{\delta} = \boldsymbol{\delta}_0,\end{aligned}$$

and that some easily (and consistent) computable initial estimator, say $\widehat{\boldsymbol{\delta}}_n$, for $\boldsymbol{\delta}_0$ is available. In this case we may consider the following GMM estimator for λ_0 corresponding to some weighting matrix $\boldsymbol{\Upsilon}_n^{\lambda\lambda}$:

$$\widehat{\lambda}_n = \underset{\lambda}{\operatorname{argmin}} \quad \mathbf{g}_n^\lambda(w_1, \dots, w_n, \lambda, \widehat{\boldsymbol{\delta}})^\top \boldsymbol{\Upsilon}_n^{\lambda\lambda}(w_1, \dots, w_n, \lambda, \widehat{\boldsymbol{\delta}}). \quad (6.4)$$

Utilizing $\widehat{\lambda}_n$ we may further consider the following estimator for $\boldsymbol{\delta}_0$ corresponding to some weight matrix $\boldsymbol{\Upsilon}_n^{\delta\delta}$:

$$\widehat{\boldsymbol{\delta}}_n = \underset{\boldsymbol{\delta}}{\operatorname{argmin}} \quad \mathbf{g}_n^\delta(w_1, \dots, w_n, \widehat{\lambda}_n, \boldsymbol{\delta})^\top \boldsymbol{\Upsilon}_n^{\delta\delta}(w_1, \dots, w_n, \widehat{\lambda}_n, \boldsymbol{\delta}). \quad (6.5)$$

GMM estimator like $\widehat{\boldsymbol{\theta}}$ in Equation (6.1) are often referred to as **one-step estimators**. Estimators like $\widehat{\lambda}_n$ and $\widehat{\boldsymbol{\delta}}_n$ in Equations (6.4) and (6.5) above, where the sample moments depend on some initial estimator, are often referred to as **two-step estimators**.

If the model conditions are valid, we would expect the most efficient one-step estimator to be more efficient than the most efficient two-step estimators. However, as usual, there are trade-offs. One trade-off is in terms of computations. Recall that for small sample sizes ML is available as an alternative to GMM. For large sample size, statistical efficiency may be less important than computational efficiency and feasibility, and thus the use of two-step GMM estimators may be attractive. Also, Monte Carlo studies suggest that in many situations, the loss of efficiency may be relatively small. Another trade-off is that the misspecification of one moment condition will typically result in inconsistent estimates of all model parameters.

6.1.2 One-Step GMM Estimation

Assuming that $\widehat{\boldsymbol{\theta}}_n$ is an interior point, the first-order condition for maximization of the objective function is:

$$\underset{(k \times 1)}{\mathbf{0}} = \underset{(k \times 1)}{\frac{\partial Q_n(\widehat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}}} = -\underset{(k \times s)}{\mathbf{G}_n(\widehat{\boldsymbol{\theta}})}^\top \underset{(s \times s)}{\boldsymbol{\Upsilon}} \underset{(s \times 1)}{\mathbf{g}_n(\widehat{\boldsymbol{\theta}})}, \quad (6.6)$$

where $\mathbf{G}_n(\boldsymbol{\theta})$ is the Jacobian of $\mathbf{g}_n(\boldsymbol{\theta})$:

$$\mathbf{G}_n(\boldsymbol{\theta}) \equiv \frac{\partial \mathbf{g}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}.$$

Now using Taylor expansion to $\mathbf{g}_n(\boldsymbol{\theta})$, yields:

$$\mathbf{g}_n(\widehat{\boldsymbol{\theta}}) = \mathbf{g}_n(\boldsymbol{\theta}_0) + \mathbf{G}_n(\bar{\boldsymbol{\theta}}) (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0). \quad (6.7)$$

Substituting (6.7) into the first-order condition (6.6), we obtain:

$$\underset{(k \times 1)}{\mathbf{0}} = \underset{(k \times 1)}{\frac{\partial Q_n(\widehat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}}} = -\underset{(k \times s)}{\mathbf{G}_n(\widehat{\boldsymbol{\theta}})}^\top \underset{(s \times s)}{\boldsymbol{\Upsilon}} \underset{(s \times 1)}{\mathbf{g}_n(\widehat{\boldsymbol{\theta}})} - \underset{(k \times s)}{\mathbf{G}_n(\widehat{\boldsymbol{\theta}})}^\top \underset{(s \times s)}{\boldsymbol{\Upsilon}} \underset{(s \times k)}{\mathbf{G}_n(\bar{\boldsymbol{\theta}})} \underset{(k \times 1)}{(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)}.$$

Solving this for $(\hat{\theta}_n - \theta_0)$ and multiplying by \sqrt{n} yield:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = - \left[G(\hat{\theta})_n^\top \mathbf{r} G_n(\bar{\theta}) \right]^{-1} G(\hat{\theta})_n^\top \mathbf{r} \left[\sqrt{n} g_n(\mathbf{w}_i, \theta_0) \right] + o_p(1).$$

For easy exposition, assume that

$$G_n(\hat{\theta}) \xrightarrow{p} G \quad \text{by some LLN} \tag{6.8}$$

$$\mathbf{r}_n \xrightarrow{p} \mathbf{r} \quad \text{by some LLN} \tag{6.9}$$

$$\sqrt{n} g_n(\theta_0) \xrightarrow{d} N(\mathbf{0}, \Psi) \quad \text{by some CLT} \tag{6.10}$$

where Ψ is some positive definite matrix. Then applying traditional asymptotic rules:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(\mathbf{0}, \Phi),$$

where:

$$\Phi = \left[G^\top \mathbf{r} G \right]^{-1} G^\top \mathbf{r} \Psi \mathbf{r} G \left[G^\top \mathbf{r} G \right]^{-1}.$$

It can be seen that if we choose $\mathbf{r} = \hat{\Psi}_n^{-1}$ (weights are given by the variance-covariance matrix of the moment conditions), where $\hat{\Psi} \xrightarrow{p} \Psi$, the variance-covariance simplifies to

$$\mathbb{V}(\hat{\theta}_n) = \Phi = \left[G^\top \Psi^{-1} G \right]^{-1}.$$

Since $\left[G^\top \mathbf{r} G \right]^{-1} G^\top \mathbf{r} \Psi \mathbf{r} G \left[G^\top \mathbf{r} G \right]^{-1} - \left[G^\top \Psi^{-1} G \right]^{-1}$ is positive semidefinite it follows that $\mathbf{r} = \hat{\Phi}_n^{-1}$ gives the optimal GMM estimator (less asymptotic variance).

6.1.3 Two-Step GMM Estimation

The usual approach to deriving the limiting distribution of two-step GMM estimators is to manipulate the score of the objective function by expanding the sample moment vector around the true parameter, using a Taylor expansion.¹

Consider the two-step GMM estimators for λ_0 defined in Equation (6.4). Applying this approach, and assuming typical regularity conditions, we get:

$$\sqrt{n}(\hat{\lambda}_n - \lambda_0) = - \left[(G_n^{\lambda\lambda})^\top \mathbf{r}_n^{\lambda\lambda} G_n^{\lambda\lambda} \right]^{-1} (G_n^{\lambda\lambda})^\top \mathbf{r}_n^{\lambda\lambda} \left[\sqrt{n} g_n^\lambda(\lambda_0, \delta_0) + G_n^{\lambda\delta} \sqrt{n}(\hat{\delta}_n - \delta_0) \right] + o_p(1), \tag{6.11}$$

where

$$\begin{aligned} \frac{\partial g_n^\lambda(\lambda_0, \delta_0)}{\partial \lambda} &\xrightarrow{p} G^{\lambda\lambda}, \\ \frac{\partial g_n^\lambda(\lambda_0, \delta_0)}{\partial \delta} &\xrightarrow{p} G^{\lambda\delta}, \\ \mathbf{r}_n^{\lambda\lambda} &\xrightarrow{p} \mathbf{r}^{\lambda\lambda}. \end{aligned}$$

¹For more on two-step estimation see Newey and McFadden (1994, section 6)

In many cases the estimator $\hat{\boldsymbol{\delta}}_n$ will be asymptotically linear in the sense that

$$\sqrt{n}(\hat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0) = \frac{1}{\sqrt{n}} \mathbf{T}_n^\top \mathbf{u}_n + o_p(1),$$

where \mathbf{T}_n is a non-stochastic $n \times k_\delta$ matrix, where k_δ is the dimension of $\boldsymbol{\delta}_0$, and where $\mathbf{u}_n = (u_1, \dots, u_n)^\top$. Now define:

$$\mathbf{g}_{*n}^\lambda(\boldsymbol{\lambda}_0, \boldsymbol{\delta}_0) = \mathbf{g}_n^\lambda(\boldsymbol{\lambda}_0, \boldsymbol{\delta}_0) + \frac{1}{n} \mathbf{G}^{\lambda\delta} \mathbf{T}_n^\top \mathbf{u}_n.$$

Then Equation (6.11) can be rewritten as:

$$\sqrt{n}(\hat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda}_0) = - \left[(\mathbf{G}^{\lambda\lambda})^\top \boldsymbol{\Upsilon}^{\lambda\lambda} \mathbf{G}^{\lambda\lambda} \right]^{-1} (\mathbf{G}^{\lambda\lambda})^\top \boldsymbol{\Upsilon}^{\lambda\lambda} \left[\sqrt{n} \mathbf{g}_{*n}^\lambda(\boldsymbol{\lambda}_0, \boldsymbol{\delta}_0) \right] + o_p(1). \quad (6.12)$$

Now suppose that

$$\sqrt{n} \mathbf{g}_{*n}^\lambda(\boldsymbol{\lambda}_0, \boldsymbol{\delta}_0) \xrightarrow{d} N(0, \boldsymbol{\Psi}_*^{\lambda\lambda})$$

where $\boldsymbol{\Psi}_*^{\lambda\lambda}$ is some positive definite matrix. Then

$$\sqrt{n}(\hat{\boldsymbol{\lambda}}_n - \boldsymbol{\lambda}_0) \xrightarrow{d} N(0, \boldsymbol{\Phi}_*^{\lambda\lambda})$$

with:

$$\boldsymbol{\Phi}_*^{\lambda\lambda} = \left[(\mathbf{G}^{\lambda\lambda})^\top \boldsymbol{\Upsilon}^{\lambda\lambda} \mathbf{G}^{\lambda\lambda} \right]^{-1} (\mathbf{G}^{\lambda\lambda})^\top \boldsymbol{\Upsilon}^{\lambda\lambda} \boldsymbol{\Phi}_*^{\lambda\lambda} \boldsymbol{\Upsilon}^{\lambda\lambda} \mathbf{G}^{\lambda\lambda} \left[(\mathbf{G}^{\lambda\lambda})^\top \boldsymbol{\Upsilon}^{\lambda\lambda} \mathbf{G}^{\lambda\lambda} \right]^{-1}$$

From this it is seen that if we choose $\boldsymbol{\Upsilon}_n^{\lambda\lambda} = (\boldsymbol{\Phi}_{*n}^{\lambda\lambda})^{-1}$ where $\boldsymbol{\Phi}_{*n}^{\lambda\lambda} \xrightarrow{p} \boldsymbol{\Phi}_*^{\lambda\lambda}$, then variance-covariance simplifies to

$$\boldsymbol{\Phi}_*^{\lambda\lambda} = \left[(\mathbf{G}^{\lambda\lambda})^\top (\boldsymbol{\Psi}_*^{\lambda\lambda})^{-1} \mathbf{G}^{\lambda\lambda} \right]^{-1}.$$

So, using the weighting matrix $\boldsymbol{\Upsilon}_n^{\lambda\lambda}$, a consistent estimator for the inverse of the limiting variance-covariance matrix $\boldsymbol{\Psi}_*^{\lambda\lambda}$ yields the efficient two-step GMM estimator.

Suppose that Equation (6.10) holds and:

$$\boldsymbol{\Psi} = \begin{pmatrix} \boldsymbol{\Psi}^{\lambda\lambda} & \boldsymbol{\Psi}^{\lambda\delta} \\ \boldsymbol{\Psi}^{\delta\lambda} & \boldsymbol{\Psi}^{\delta\delta} \end{pmatrix},$$

then the limiting distribution of the sample moment vector \mathbf{g}_n^λ evaluated at the true parameter is given by

$$\sqrt{n} \mathbf{g}_n^\lambda(\boldsymbol{\lambda}_0, \boldsymbol{\delta}_0) \xrightarrow{d} N(0, \boldsymbol{\Psi}^{\lambda\lambda})$$

Note that in general $\boldsymbol{\Psi}_*^{\lambda\lambda} \neq \boldsymbol{\Psi}^{\lambda\lambda}$, unless $\mathbf{G}^{\lambda\delta} = \mathbf{0}$, and that in general $\boldsymbol{\Psi}_*^{\lambda\lambda}$ will depend on \mathbf{T}_n , which in turn will depend on the employed estimator $\hat{\boldsymbol{\delta}}_n$. In other words, unless $\mathbf{G}^{\lambda\delta} = \mathbf{0}$, for a two-step GMM estimator, we cannot simply use the variance-covariance matrix $\boldsymbol{\Psi}^{\lambda\lambda}$ of the sample moment vector $\mathbf{m}^\lambda(\boldsymbol{\lambda}_0, \boldsymbol{\delta}_0)$, rather we need to work with the variance-covariance matrix $\boldsymbol{\Psi}_*^{\lambda\lambda}$.

Prucha (2014) illustrate the difference between $\boldsymbol{\Psi}^{\lambda\lambda}$, with elements $\Psi_{rs}^{\lambda\lambda}$, and $\boldsymbol{\Psi}_*^{\lambda\lambda}$, with elements $\Psi_{*rs}^{\lambda\lambda}$, for the important special case where the moment conditions are quadratic and u_i is i.i.d $N(0, \sigma^2)$. For simplicity assume that

$$\mathbf{g}_n^\lambda(\lambda_0, \boldsymbol{\delta}_0) = \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n a_{ij1} u_i u_j \\ \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n a_{ij2} u_i u_j \end{pmatrix}.$$

Now, for $r = 1, 2$, let a_{ir} denote the (i, r) th element of $\mathbf{G}^{\lambda\delta} \mathbf{T}_n^\top$, then by Equation (6.10):

$$\mathbf{g}_{*n}^\lambda(\lambda_0, \boldsymbol{\delta}_0) = \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n a_{ij1} u_i u_j + \frac{1}{n} \sum_{i=1}^n a_{i1} u_i \\ \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n a_{ij2} u_i u_j + \frac{1}{n} \sum_{i=1}^n a_{i2} u_i \end{pmatrix}$$

It then follows from Limiting Distribution for linear-quadratic forms 3.29 that

$$\Psi_{rs}^{\lambda\lambda} = 2\sigma^4 \sum_{i=1}^n \sum_{j=1}^n a_{ijr} a_{ijs}$$

but

$$\Psi_{*rs}^{\lambda\lambda} = 2\sigma^4 \sum_{i=1}^n \sum_{j=1}^n a_{ijr} a_{ijs} + \sigma^2 \sum_{i=1}^n a_{ir} a_{is}$$

Note that a_{ir} and a_{is} in the last sum of the RHS for the expression for $\Psi_{*rs}^{\lambda\lambda}$ depend on what estimator $\hat{\boldsymbol{\delta}}_n$ is employed in the sample moment vector $\mathbf{g}_n^\lambda(\lambda_0, \hat{\boldsymbol{\delta}})$ used to form the objective function for the two-step GMM estimator $\hat{\lambda}_n$ defined in Equation (6.4). It is for this reason that in the literature on two-step GMM estimation, users are often advised to follow a specific sequence of steps, to ensure the proper estimation of respective variance-covariance matrices.

6.2 Spatial Two Stage Estimation of SLM

In this section we will derive the Spatial Two Stage Least Square (S2SLS) procedure for estimating the SLM model. The asymptotic properties of this model was first derived by Kelejian and Prucha (1998).² To get some insights about this procedure recall that Spatial Lag Model (SLM) is given by:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \rho \mathbf{W}\mathbf{y} + \boldsymbol{\varepsilon}.$$

A more concise way to express the model is:

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\varepsilon},$$

where $\mathbf{Z} = [\mathbf{X}, \mathbf{W}\mathbf{y}]$ and the $(k+1) \times 1$ coefficient column vector is rearranged as $\boldsymbol{\delta} = (\boldsymbol{\beta}^\top, \rho)^\top$. As we have previously shown in Section 4.1, the presence of the spatially lagged dependent variable on the right hand side of the equation induces endogeneity or simultaneous equation bias. Therefore the OLS estimates are inconsistent.

Instead of applying QML or ML estimation procedure, we might rely on the instrumental variable approach in order to deal with the endogeneity caused by the spatial lag variable. The principle of instrumental variables estimation is based on the existence of a matrix of instruments, say \mathbf{H} , that are strongly correlated with \mathbf{Z} but asymptotically uncorrelated with $\boldsymbol{\varepsilon}$.

²In particular, Kelejian and Prucha (1998) derived this model as the first step in their Generalized S2SLS.

At this point is important to stress that the only endogenous variable in this model is the spatial lagged variable. Therefore, matrix \mathbf{H} should contain all the predetermined variables, that is, \mathbf{X} and the instrument(s) for $\mathbf{W}\mathbf{y}$. As we will see later, an important feature of this estimation procedure is that it does not require to compute the Jacobian term. Another important feature is that it does not make the strong assumption of normality of the error terms.

6.2.1 Instruments in the Spatial Context

What is the best instrument(s) for $\mathbf{W}\mathbf{y}$? To obtain the ideal matrix of instruments \mathbf{H} we should understand the literature of optimal instrumental variables. Roughly, it states that the ‘best instruments’ for the r.h.s variables are the conditional means. Thus, the ideal instruments are:

$$\begin{aligned}\mathbb{E}(\mathbf{Z}|\mathbf{X}) &= [\mathbb{E}(\mathbf{X}|\mathbf{X}), \mathbb{E}(\mathbf{W}\mathbf{y}|\mathbf{X})] \\ &= [\mathbf{X}, \mathbf{W}\mathbb{E}(\mathbf{y}|\mathbf{X})] \quad \text{since } \mathbf{W} \text{ is non-stochastic.}\end{aligned}$$

Since \mathbf{X} is non-stochastic, \mathbf{X} is its own best instrument, whereas the best instruments for $\mathbf{W}\mathbf{y}$ are given by $\mathbf{W}\mathbb{E}(\mathbf{y}|\mathbf{X})$. Noting that the reduced-form equation is $\mathbf{y} = (\mathbf{I}_n - \rho\mathbf{W})^{-1}(\mathbf{X}\beta + \varepsilon)$, and using Leontief Expansion (Lemma 2.3), the expected value of the reduced form is:

$$\mathbb{E}(\mathbf{y}|\mathbf{X}) = (\mathbf{I}_n - \rho\mathbf{W})^{-1}\mathbf{X}\beta = \left[\mathbf{I}_n + \rho\mathbf{W} + \rho^2\mathbf{W}^2 + \dots \right] \mathbf{X}\beta = \left[\sum_{l=1}^{\infty} \rho^l \mathbf{W}^l \right] \mathbf{X}\beta \quad (6.13)$$

In principle, the problem is to approximate $\mathbb{E}(\mathbf{y}|\mathbf{X})$ as closely as possible without incurring in the inversion of $(\mathbf{I}_n - \rho\mathbf{W})$. Thus, note that (6.13) can be expressed as a linear function of $\mathbf{X}, \mathbf{W}\mathbf{X}, \mathbf{W}^2\mathbf{X}, \dots$. As a result, and given that the roots of $\rho\mathbf{W}_n$ are less than one in absolute value, the conditional expectation can also be written as:

$$\begin{aligned}\mathbb{E}(\mathbf{W}\mathbf{y}|\mathbf{X}) &= \mathbf{W}\mathbb{E}(\mathbf{y}|\mathbf{X}) \\ &= \mathbf{W}(\mathbf{I}_n - \rho\mathbf{W})^{-1}\mathbf{X}\beta \\ &= \mathbf{W} \left[\mathbf{I}_n + \rho\mathbf{W} + \rho^2\mathbf{W}^2 + \rho^3\mathbf{W}^3 + \dots \right] \mathbf{X}\beta \\ &= \mathbf{W} \left[\sum_{l=1}^{\infty} \rho^l \mathbf{W}^l \right] \mathbf{X}\beta \\ &= \mathbf{W}\mathbf{X}\beta + \mathbf{W}^2\mathbf{X}(\rho\beta) + \mathbf{W}^3\mathbf{X}(\rho^2\beta) + \mathbf{W}^4\mathbf{X}(\rho^3\beta) + \dots\end{aligned}$$

To avoid issues associated with the computation of the inverse of the $n \times n$ matrix $(\mathbf{I}_n - \rho\mathbf{W})$, Kelejian and Prucha (1998, 1999) suggest the use of an **approximation** of the best instruments. Specifically, since $\mathbb{E}(\mathbf{y}|\mathbf{X})$ is linear in $\mathbf{X}, \mathbf{W}\mathbf{X}, \mathbf{W}^2\mathbf{X}, \dots$, they suggest using a set of instruments \mathbf{H} which consists of the linearly independent (LI) columns of $\mathbf{X}, \mathbf{W}\mathbf{X}, \mathbf{W}^2\mathbf{X}, \dots, \mathbf{W}^l\mathbf{X}$ where l is a pre-selected finite constant and is generally set to 2 in applied studies. Thus, if $l = 2$ we can write the instruments as:

$$\mathbf{H} = (\mathbf{X}, \mathbf{W}\mathbf{X}, \mathbf{W}^2\mathbf{X}).$$

- R** The intuition behind the instruments is the following: Since \mathbf{X} determines \mathbf{y} , then it must be true that $\mathbf{W}\mathbf{X}, \mathbf{W}^2\mathbf{X}, \dots$ determines $\mathbf{W}\mathbf{y}$. Furthermore, since \mathbf{X} is uncorrelated with $\boldsymbol{\varepsilon}$, then $\mathbf{W}\mathbf{X}$ must be also uncorrelated with $\boldsymbol{\varepsilon}$.

The theoretical literature have also suggested the so-called optimal instruments matrix. For example, using the conditional expectation in (6.13), Lee (2003) suggested the instrument matrix:

$$\mathbf{H}^* = [\mathbf{X}, \mathbf{W}(\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta}],$$

which requires the use of consistent first round estimates for ρ and $\boldsymbol{\beta}$. In Kelejian et al. (2004), a similar approach is outlined where the matrix inverse is replaced by the power expansion. This yield an instruments matrix as:

$$\mathbf{H} = \left[\mathbf{X}, \mathbf{W} \left(\sum_{l=1}^{\infty} \rho_0^l \mathbf{W}^l \right) \mathbf{X}\boldsymbol{\beta} \right].$$

In any practical implementation, the power expansion must be truncated at some point.

6.2.2 Defining the S2SLS Estimator

Now that we have defined the matrix of instruments \mathbf{H} , we can apply the standard two-stage procedure with the exception that the assumptions must also consider the asymptotic behavior of \mathbf{W} and $(\mathbf{I}_n - \rho\mathbf{W})$. Given the inclusion of the weight matrices, this procedure is called spatial two stage least squares (S2SLS) (Kelejian and Prucha, 1998).

As usual, we start with some assumptions about the error term. Specifically, we will assume that the errors form triangular arrays and are heterokedastic. Note that Kelejian and Prucha (1998) derived the asymptotic properties assuming that the errors are homokedastic. Kelejian and Prucha (2010) extend the model by assuming heteroskedasticity. We will keep Kelejian and Prucha (2010)'s assumption since homoskedasticity can be viewed as a particular case. A key difference with the ML approach is that we do not need to assume the whole distribution of the error term.

Assumption 6.1 — Heterokedastic Errors (Kelejian and Prucha, 2010). The errors $\{\epsilon_{i,n}, 1 \leq i \leq n, n \geq 1\}$ satisfy $\mathbb{E}(\epsilon_{i,n}) = 0$, $\mathbb{E}(\epsilon_{i,n}^2) = \sigma_{i,n}^2$, with $0 < \underline{\sigma} \leq \sigma_{i,n}^2 \leq \bar{\sigma} < \infty$. Additionally the errors are assumed to possess fourth moments, that is $\sup_{1 \leq i \leq n, n \geq 1} \mathbb{E}|\epsilon_{i,n}|^{4+\eta} < \infty$ for some $\eta > 0$. Furthermore, for each $n \geq 1$ the random variables $\epsilon_{1,n}, \dots, \epsilon_{n,n}$ are totally independent.

Assumption 6.1 (Heterokedastic Errors) states the first two moments of the error terms, but it says nothing about its distribution. It also assumes that the error terms are heterokedastic, i.e., the unobserved variables have different variance for all spatial units. Finally, this assumption also allows for the innovations to depend on the sample size n , i.e., to form a **triangular arrays**. See our discussion in Section 3.7 about triangular arrays.

Now, we state some assumptions about the behavior of the spatial weight matrix \mathbf{W} .

Assumption 6.2 — Diagonal elements of \mathbf{W}_n (Kelejian and Prucha, 1998). All diagonal elements of the spatial weighting matrix \mathbf{W}_n are zero

Assumption 6.2 (Diagonal elements of \mathbf{W}_n) is a normalization of the model and it also implies that no spatial unit is viewed as its own neighbor.

Assumption 6.3 — Nonsingularity (Kelejian and Prucha, 1998). The matrix $(\mathbf{I}_n - \rho_0 \mathbf{W}_n)$ is nonsingular with $|\rho_0| < 1$.

Under Nonsingularity Assumption 6.3, we can write the reduced form of the true model as:

$$\mathbf{y}_n = (\mathbf{I}_n - \rho_0 \mathbf{W}_n)^{-1} \mathbf{X}_n \beta_0 + (\mathbf{I}_n - \rho_0 \mathbf{W}_n)^{-1} \boldsymbol{\varepsilon}_n.$$

That is, Assumption 6.3 implies that the model is complete in that it determines \mathbf{y}_n . Furthermore, Kelejian and Prucha (1998) note that the elements of $(\mathbf{I}_n - \rho_0 \mathbf{W}_n)^{-1}$ will depend on the sample size n , even if the elements of \mathbf{W}_n does not depend on n . Therefore, in general, the elements of \mathbf{y}_n will also depend on n and thus form a triangular array, even in the case where the errors $\epsilon_{i,n}$ do not depend on n .

Assumption 6.1 (Heterokedastic Errors) implies further that the population variance-covariance matrix of \mathbf{y}_n is equal to

$$\mathbb{E}(\mathbf{y}_n \mathbf{y}_n^\top) = \boldsymbol{\Omega}_{y_n} = (\mathbf{I}_n - \rho_0 \mathbf{W}_n)^{-1} \boldsymbol{\Sigma}_n (\mathbf{I}_n - \rho_0 \mathbf{W}_n^\top)^{-1}, \quad (6.14)$$

where $\boldsymbol{\Sigma} = \text{diag}(\sigma_{i,n}^2)$. If we assume homokedasticity, then the variance-covariance matrix of \mathbf{y} reduces to:

$$\mathbb{E}(\mathbf{y}_n \mathbf{y}_n^\top) = \boldsymbol{\Omega}_{y_n} = \sigma_\epsilon^2 (\mathbf{I}_n - \rho_0 \mathbf{W}_n)^{-1} (\mathbf{I}_n - \rho_0 \mathbf{W}_n^\top)^{-1}.$$

Assumption 6.4 — Bounded matrices (Kelejian and Prucha, 1998). The row and column sums of the matrices \mathbf{W}_n and $(\mathbf{I}_n - \rho_0 \mathbf{W}_n)$ are bounded uniformly in absolute value.

This assumption guarantees that the variance of \mathbf{y}_n in Equation (6.14), which depend on \mathbf{W}_n and $(\mathbf{I}_n - \rho_0 \mathbf{W}_n)$, are uniformly bounded in absolute value as n goes to infinity, thus limiting the degree of correlation between, respectively, the elements of $\boldsymbol{\varepsilon}_n$ and \mathbf{y}_n . This assumption is technical and will be used in the large-sample derivations of the regression parameters estimator.

R Applied to \mathbf{W}_n Assumption 6.4 (Bounded matrices) means that each cross-sectional unit can only have a limited number of neighbors. Applied to $(\mathbf{I}_n - \rho_0 \mathbf{W}_n)$ limits the degree of correlation.

Assumption 6.5 — No Perfect Multicollinearity (Kelejian and Prucha, 1998). The regressor matrices \mathbf{X}_n have full column rank (for n large enough). Furthermore, the elements of the matrices \mathbf{X}_n are uniformly bounded in absolute value.

Now we state some assumptions about the instruments.

Assumption 6.6 — Rank Instruments, (Kelejian and Prucha, 1998). The instrument matrices \mathbf{H}_n have full column rank $p \geq k + 1$ for all n large enough. Furthermore, the elements of the matrices \mathbf{H}_n are uniformly bounded in absolute value. They are composed of a subset of the linearly independent columns of $(\mathbf{X}_n, \mathbf{W}_n \mathbf{X}_n, \mathbf{W}_n^2 \mathbf{X}_n, \dots)$.

Assumption 6.7 — Limits of Instruments (Kelejian and Prucha, 1998). Let \mathbf{H}_n be a matrix of instruments, then:

- (a) $\lim_{n \rightarrow \infty} n^{-1} \mathbf{H}_n^\top \mathbf{H}_n = \mathbf{Q}_{HH}$ where \mathbf{Q}_{HH} is finite and nonsingular.
- (b) $\text{plim}_{n \rightarrow \infty} n^{-1} \mathbf{H}_n^\top \mathbf{Z}_n = \mathbf{Q}_{HZ}$ where \mathbf{Q}_{HZ} is finite and has full column rank.

Since the instrument matrix \mathbf{H}_n contains the spatially lagged explanatory variables, the first condition in Assumption 6.7 (Limits of Instruments) $\lim_{n \rightarrow \infty} n^{-1} \mathbf{H}_n^\top \mathbf{H}_n = \mathbf{Q}_{HH}$ implies that $\mathbf{W}_n \mathbf{X}_n$ and \mathbf{X}_n cannot be linearly dependent. This condition would be violated if for example $\mathbf{W}_n \mathbf{X}_n$ include a spatial lag for the constant term or the model is the pure SLM. The second condition in Assumption 6.7 (Limits of Instruments) requires a non-null correlation between the instruments and the original variables.

Given all this assumptions we can define the S2SLS estimator as follows.

Definition 6.2.1 — Spatial Two Stage Least Square Estimator. Let \mathbf{H}_n be the matrix ($n \times p$) of instruments. Then the S2SLS is given by:

$$\hat{\delta}_{S2SLS} = \left(\widehat{\mathbf{Z}}_n^\top \mathbf{Z}_n \right)^{-1} \widehat{\mathbf{Z}}_n^\top \mathbf{y}_n, \quad (6.15)$$

where:

$$\widehat{\mathbf{Z}}_n = \mathbf{H}_n \hat{\boldsymbol{\theta}}_n = \mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{Z}_n = \mathbf{P}_{H,n} \mathbf{Z}_n \quad (6.16)$$

Note that the S2SLS estimator in (6.15) is similar to the standard 2SLS. We first need the predicted values for \mathbf{Z}_n based on the OLS regression of \mathbf{Z}_n on \mathbf{H}_n in the first stage. Consider this first stage as the regression $\mathbf{Z}_n = \mathbf{H}_n \boldsymbol{\theta} + \boldsymbol{\xi}_n$, so that $\hat{\boldsymbol{\theta}}_n = (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{Z}_n$. Then the predicted values $\widehat{\mathbf{Z}}_n$ is obtained using Equation (6.16) where $\mathbf{P}_{H,n}$ is the projection matrix, which symmetric and idempotent, and hence singular. Note also that \mathbf{H}_n is a $n \times p$ matrix, which also includes the exogenous variables \mathbf{X}_n . It is also important to note that the projection matrix does not affect \mathbf{X}_n , but it does affect the endogenous variable $\mathbf{W}_n \mathbf{y}_n$:

$$\mathbf{P}_{H,n} \mathbf{Z}_n = [\mathbf{X}_n, \mathbf{P}_{H,n} \mathbf{W}_n \mathbf{y}_n] = [\mathbf{X}_n, \widehat{\mathbf{W}_n \mathbf{y}_n}] \quad (6.17)$$

Note that this approach is in the same spirit as the traditional treatment in simultaneous equation setting, where each endogenous variable (including the spatial lag) is regressed on the complete set of exogenous variables to form its instrument.

6.2.3 S2SLS Estimator as GMM

Maybe you remember from your econometric class that the 2SLS procedure is a sub-model of the one-step GMM estimator. Recall that the GMM estimator is defined as the solution of the minimization problem as in Equation (6.1):

$$\hat{\boldsymbol{\delta}}_{GMM} = \arg \min_{\boldsymbol{\beta}} \left\{ \underbrace{\mathbf{g}_n(\boldsymbol{\beta})^\top}_{1 \times p} \underbrace{\boldsymbol{\Upsilon}_n^{-1}}_{p \times p} \underbrace{\mathbf{g}_n(\boldsymbol{\beta})}_{p \times 1} \right\},$$

where

$$\mathbf{g}_n = \frac{1}{n} \mathbf{H}_n^\top \boldsymbol{\varepsilon}_n = \frac{1}{n} \mathbf{H}_n^\top (\mathbf{y}_n - \mathbf{Z}_n \boldsymbol{\delta}).$$

The matrix $\boldsymbol{\Upsilon}_n^{-1}$ is the optimal weight matrix, which correspond to the inverse of the covariance matrix of the sample moments:

$$\boldsymbol{\Upsilon}_n = \frac{1}{n} \hat{\sigma}_\epsilon^2 \mathbf{H}_n^\top \mathbf{H}_n$$

Then, the function to minimize is:

$$Q = \frac{1}{n \hat{\sigma}^2} \left\{ \left[\mathbf{H}_n^\top \mathbf{y}_n - \mathbf{H}_n^\top \mathbf{Z}_n \boldsymbol{\delta} \right]^\top \left(\mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \left[\mathbf{H}_n^\top \mathbf{y}_n - \mathbf{H}_n^\top \mathbf{Z}_n \boldsymbol{\delta} \right] \right\}$$

Obtaining the first order conditions and solving for $\boldsymbol{\delta}$, we obtain:

$$\hat{\boldsymbol{\delta}}_{GMM,n} = \left(\mathbf{Z}_n^\top \mathbf{P}_{H,n} \mathbf{Z}_n \right)^{-1} \mathbf{Z}_n^\top \mathbf{P}_{H,n} \mathbf{y}_n \quad (6.18)$$

6.2.4 Additional Endogenous Variables

In the specification considered so far, the only endogenous variable is the spatially lagged dependent variable $\mathbf{W}\mathbf{y}$. However, in practice, some of other explanatory variables may be endogenous as well, requiring instruments in addition to the spatially lagged exogenous variables that were necessary for the spatially lagged dependent variable.

For example [Anselin and Lozano-Gracia \(2008\)](#) were interested in the effect of improved air quality on house prices. Since the air-quality variables were obtained using interpolated air pollution measures, they argued that these measure may suffer from “error in variable” problem which lead to an additional endogeneity problem to that of spatially lagged variable. In particular they consider the following model:

$$y_i = \rho \sum_{j=1}^n w_{ij} y_j + \mathbf{x}_i' \boldsymbol{\beta} + \gamma_1 \text{pol}_i^1 + \gamma_2 \text{pol}_i^2 + \epsilon_i,$$

where y_i is the house price, \mathbf{x}_i is a vector of controls, pol_i^1 and pol_i^2 are the air quality variables and ϵ_i is the error term. Since the actual pollution is not observed at locations i of the house transaction, it is replaced by a spatially interpolated value, such as the result of a **kriging prediction**. This interpolated value measures the true pollution with error causing simultaneous equation bias, so they needed proper instruments for these variables. They instrumentalize these endogenous variables using the latitude, longitude and their product as the instruments.

In particular, we can write the general model with additional endogenous variables

$$\mathbf{y} = \rho \mathbf{W}\mathbf{y} + \mathbf{X}_1 \boldsymbol{\beta} + \mathbf{Y} \boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

where \mathbf{Y} is a $n \times q$ matrix the endogenous explanatory variables and \mathbf{X}_1 is a $n \times k_1$ matrix of exogenous variables. In a spatial lag model, an additional question is whether these instruments (for the endogenous explanatory variables) should be included in spatially lagged form as well, similar to what is done for the exogenous variables. As before, the rationale for this comes from the structure of the reduced form. In this case the reduced form is given by:

$$\mathbb{E}[\mathbf{W}\mathbf{y} | \mathbf{Z}] = \mathbf{W} (\mathbf{I} - \rho \mathbf{W})^{-1} \mathbf{X}_1 \boldsymbol{\beta} + \mathbf{W} (\mathbf{I} - \rho \mathbf{W})^{-1} \mathbf{Y} \boldsymbol{\gamma},$$

where $\mathbf{Z} = [\mathbf{X}_1, \mathbf{Y}]$. The problem here is that the \mathbf{Y} are endogenous, and thus they do not belong on the right hand side of the reduced form! If they are replaced by their instruments, then the presence of the term $\mathbf{W}(\mathbf{I} - \rho\mathbf{W})^{-1}$ would suggest the need for spatial lags to be included as well. In other words, since the system determining \mathbf{y} and \mathbf{Y} is not completely specified, the optimal instruments are not known (Bivand and Piras, 2015). If there exists a matrix $n \times k_1$ of additional pre-determined variables, say \mathbf{X}_2 , the instruments should be:

$$\mathbf{H} = (\mathbf{X}_1, \mathbf{W}\mathbf{X}_1, \dots, \mathbf{W}^l\mathbf{X}_1, \mathbf{X}_2, \mathbf{W}\mathbf{X}_2, \dots, \mathbf{W}^l\mathbf{X}_2)_{LI} \quad (6.19)$$

6.2.5 Consistency of S2SLS Estimator

In this section, we will sketch the proof the consistency of the S2SLS Estimator. First, note that $\mathbf{H}_n(\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top$ is symmetric and idempotent and so $\widehat{\mathbf{Z}}_n^\top \mathbf{Z}_n = \widehat{\mathbf{Z}}_n^\top \widehat{\mathbf{Z}}_n$. As usual, we first write the estimator in terms of the population error term:

$$\begin{aligned} \widehat{\delta}_n &= \delta_0 + (\widehat{\mathbf{Z}}_n^\top \widehat{\mathbf{Z}}_n)^{-1} \widehat{\mathbf{Z}}_n^\top \boldsymbol{\varepsilon}_n, \\ &= \delta_0 + \left[(\mathbf{H}_n(\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{Z}_n)^\top (\mathbf{H}_n(\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{Z}_n) \right]^{-1} (\mathbf{H}_n(\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{Z}_n)^\top \boldsymbol{\varepsilon}_n, \\ &= \delta_0 + [\mathbf{Z}_n^\top \mathbf{H}_n(\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{Z}_n]^{-1} \mathbf{Z}_n^\top \mathbf{H}_n(\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \boldsymbol{\varepsilon}_n, \end{aligned} \quad (6.20)$$

where we used Assumption 6.6 (Rank of Instruments). Solving for $\widehat{\delta}_n - \delta_0$ we obtain:

$$\begin{aligned} (\widehat{\delta}_n - \delta_0) &= \left[\left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right)^\top \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right) \right]^{-1} \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right)^\top \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \left(\frac{1}{n} \mathbf{H}_n^\top \boldsymbol{\varepsilon}_n \right), \\ &= \widetilde{\mathbf{P}}_n^\top \left(\frac{1}{n} \mathbf{H}_n^\top \boldsymbol{\varepsilon}_n \right), \end{aligned} \quad (6.21)$$

where:

$$\widetilde{\mathbf{P}}_n = \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right) \left[\left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right)^\top \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right) \right]^{-1}.$$

From Assumption 6.7 (Limits of Instruments), we know that:

$$\begin{aligned} \lim_{n \rightarrow \infty} n^{-1} \mathbf{H}_n^\top \mathbf{H}_n &= \mathbf{Q}_{HH} \\ \text{plim}_{n \rightarrow \infty} n^{-1} \mathbf{H}_n^\top \mathbf{Z}_n &= \mathbf{Q}_{HZ}. \end{aligned}$$

Therefore, $\widetilde{\mathbf{P}}_n \xrightarrow{p} \mathbf{P}_n$, where $\mathbf{P}_n = \mathbf{Q}_{HH}^{-1} \mathbf{Q}_{HZ} (\mathbf{Q}_{HZ}^\top \mathbf{Q}_{HH}^{-1} \mathbf{Q}_{HZ})^{-1} = O_p(1)$ is a finite matrix. Thus,

$$\widetilde{\mathbf{P}}_n - \mathbf{P}_n = o_p(1) \implies \widetilde{\mathbf{P}}_n = \mathbf{P}_n + o_p(1). \quad (6.22)$$

By Assumption 6.6 (Rank of Instruments) \mathbf{H}_n is uniformly bounded in absolute value. Assumption 6.1 (Heterokedastic Errors) implies that $\epsilon_{i,n}$ forms a triangular array of identically

distributed random variables. Furthermore, we know from that assumption that $\mathbb{E}(\boldsymbol{\varepsilon}_n) = \mathbf{0}$ and $\mathbb{V}(\boldsymbol{\varepsilon}_n) = \boldsymbol{\Sigma}_n = \text{diag}(\sigma_{i,n}^2)$. Thus,

$$\begin{aligned}\mathbb{E}\left(\frac{1}{n}\mathbf{H}_n^\top \boldsymbol{\varepsilon}_n\right) &= \mathbf{0} \\ \mathbb{V}\left(\frac{1}{n}\mathbf{H}_n^\top \boldsymbol{\varepsilon}_n\right) &= \frac{1}{n^2}\mathbf{H}_n^\top \boldsymbol{\Sigma}_n \mathbf{H}_n\end{aligned}$$

Since $\mathbb{V}\left(\frac{1}{n}\mathbf{H}_n^\top \boldsymbol{\varepsilon}_n\right) \rightarrow 0$ as $n \rightarrow \infty$, by Chebyshev's Theorem 3.5, $n^{-1}\mathbf{H}_n^\top \boldsymbol{\varepsilon}_n \xrightarrow{p} \mathbf{0}$ and $\hat{\boldsymbol{\delta}}_n \xrightarrow{p} \boldsymbol{\delta}_0$

6.2.6 Asymptotic Distribution of S2SLS Estimator

Multiplying Equation (6.21) by \sqrt{n} we obtain:

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0) &= \left[\left(\frac{1}{n}\mathbf{H}_n^\top \mathbf{Z}_n \right)^\top \left(\frac{1}{n}\mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \left(\frac{1}{n}\mathbf{H}_n^\top \mathbf{Z}_n \right) \right]^{-1} \left(\frac{1}{n}\mathbf{H}_n^\top \mathbf{Z}_n \right)^\top \left(\frac{1}{n}\mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \frac{1}{\sqrt{n}}\mathbf{H}_n^\top \boldsymbol{\varepsilon}_n, \\ &= \widetilde{\mathbf{P}}_n^\top \left(\frac{1}{\sqrt{n}}\mathbf{H}_n^\top \boldsymbol{\varepsilon}_n \right).\end{aligned}\tag{6.23}$$

Inserting (6.22) into (6.23), we get:

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0) &= \frac{1}{\sqrt{n}} [\mathbf{P}_n + o_p(1)]^\top \mathbf{H}_n^\top \boldsymbol{\varepsilon}_n, \\ &= \mathbf{P}_n^\top \frac{1}{\sqrt{n}} \mathbf{H}_n^\top \boldsymbol{\varepsilon}_n + o_p(1)\end{aligned}$$

Note that we can also write: $\sqrt{n}(\hat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0) = \frac{1}{\sqrt{n}}\mathbf{T}_n^\top \boldsymbol{\varepsilon}_n + o_p(1)$ with $\mathbf{T}_n = \mathbf{H}_n \mathbf{P}_n$. Thus, by Chebyshev's inequality $n^{-1/2}\mathbf{P}_n \mathbf{H}_n^\top \boldsymbol{\varepsilon}_n = O_p(1)$ and consequently:

$$\sqrt{n}(\hat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0) = \mathbf{P}_n^\top \frac{1}{\sqrt{n}} \mathbf{H}_n \boldsymbol{\varepsilon}_n + o_p(1) = O_p(1) + o_p(1) = O_p(1)$$

Therefore using Theorem 3.28 (CLT for Linear Forms),

$$\frac{1}{\sqrt{n}}\mathbf{H}_n^\top \boldsymbol{\varepsilon}_n \xrightarrow{d} \mathbf{N}(\mathbf{0}, \mathbf{H}_n^\top \boldsymbol{\Sigma} \mathbf{H}_n)$$

Finally :

$$\sqrt{n}(\hat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0) \xrightarrow{d} \mathbf{N}(\mathbf{0}, \boldsymbol{\Omega}_n)$$

where

$$\boldsymbol{\Omega}_n = \mathbf{P}_n^\top \mathbf{H}_n^\top \boldsymbol{\Sigma}_n \mathbf{H}_n \mathbf{P}_n\tag{6.24}$$

Now, we present a formal Theorem for the asymptotic properties of the 2SLS Estimator for SLM.

Theorem 6.8 — Spatial 2SLS Estimator for SLM. Suppose that Assumptions 6.1 to 6.7 hold. Then the S2SLS estimator defined as

$$\hat{\delta}_n = (\widehat{\mathbf{Z}}_n^\top \widehat{\mathbf{Z}}_n)^{-1} \widehat{\mathbf{Z}}_n^\top \mathbf{y}_n \quad (6.25)$$

is consistent, and its asymptotic distribution is:

$$\sqrt{n}(\hat{\delta}_n - \delta_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Omega}_n) \quad (6.26)$$

where

$$\boldsymbol{\Omega}_n = \mathbf{P}_n^\top \mathbf{H}_n^\top \boldsymbol{\Sigma} \mathbf{H}_n \mathbf{P}_n \quad (6.27)$$

Inference on δ is then based on the asymptotic variance-covariance matrix:

$$\begin{aligned} \mathbb{V}(\hat{\delta}_{2SLS}) &= \left[\mathbf{Z}^\top \mathbf{H} (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{Z} \right]^{-1} \\ &\quad \times \left[\mathbf{Z}^\top \mathbf{H} (\mathbf{H}^\top \mathbf{H})^{-1} (\mathbf{H}^\top \boldsymbol{\Sigma} \mathbf{H}) (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{Z} \right] \\ &\quad \times \left[\mathbf{Z}^\top \mathbf{H} (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{Z} \right]^{-1} \\ &= (\widehat{\mathbf{Z}}^\top \widehat{\mathbf{Z}})^{-1} (\widehat{\mathbf{Z}}^\top \boldsymbol{\Sigma} \widehat{\mathbf{Z}}) (\mathbf{Z}^\top \widehat{\mathbf{Z}})^{-1} \end{aligned} \quad (6.28)$$

Theorem 6.8 gives us a very general asymptotic distribution for the S2SLS estimator. The estimator of $\boldsymbol{\Sigma}$ will be based on HAC estimators. However, under certain conditions the asymptotic variance-covariance matrix of the estimator can be reduced. For example, under homokedasticity the asymptotic variance-covariance matrix reduced to :

$$\mathbb{V}(\hat{\delta}_{2SLS}) = \sigma_\epsilon^2 (\mathbf{Q}_{HZ}^\top \mathbf{Q}_{HH}^{-1} \mathbf{Q}_{HZ})^{-1} \quad (6.29)$$

A good estimator for the asymptotic variance will be:

$$\widehat{\mathbb{V}}(\hat{\delta}_{2SLS}) = \hat{\sigma}_\epsilon^2 \left[\mathbf{Z}^\top \mathbf{H} (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{Z} \right]^{-1} \quad (6.30)$$

where:

$$\hat{\sigma}^2 = \frac{\widehat{\boldsymbol{\epsilon}}^\top \widehat{\boldsymbol{\epsilon}}}{n}, \quad \widehat{\boldsymbol{\epsilon}} = \mathbf{y} - \widehat{\mathbf{y}}. \quad (6.31)$$

6.2.7 S2SLS Estimation in R

In this section we continue our example from Section 4.6. In particular, we will estimate the following SLM model:

$$\mathbf{y} = \rho \mathbf{W} \mathbf{y} + \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where \mathbf{y} is our crime variable and \mathbf{X} contains a vector of ones and the variables INC and HOVAL. We will estimate this model again by ML procedure and then compare it with the S2SLS procedure. In R there exists two functions in order to compute the S2SLS procedure. The first one is the `stsls` from `spdep` and `stslshac` from `sphet` package (Piras, 2010). The latter allows estimating also S2SLS with heterokedasticity using HAC estimators.

We first load the required packages and dataset:

```
# Load packages and data
library("memisc")
library("spdep")
library("spatialreg")
library("sphet")

##
## Attaching package: 'sphet'
## The following object is masked from 'package:spatialreg':
##
##      impacts

data("columbus")
listw <- nb2listw(col.gal.nb)
source("getSummary.sarlm.R")
```

Now we estimate the SLM model by ML using Ord's eigen approximation of the determinant and S2SLS with homokedastic and robust standard errors.

```
# Estimate models
slm      <- lagsarlm(CRIME ~ INC + HOVAL,
                    data = columbus,
                    listw,
                    method = "eigen")
s2sls    <- stspls(CRIME ~ HOVAL + INC,
                  data = columbus,
                  listw = listw,
                  robust = FALSE,
                  W2X = TRUE)
s2sls_rob <- stspls(CRIME ~ HOVAL + INC,
                  data = columbus,
                  listw = listw,
                  robust = TRUE,
                  W2X = TRUE)
s2sls_pir <- stslshac(CRIME ~ INC + HOVAL,
                    data = columbus,
                    listw = listw,
                    HAC = FALSE)
```

stspls function fits SLM model by S2SLS, with the option of adjusting the results for heteroskedasticity. Note that the arguments are similar to **lagsarlm** from **spdep**. The **robust** option of **stspls** is set **FALSE** as default. If **TRUE** the function applies a heteroskedasticity correction to the coefficient covariances. Note that the third model **s2sls_rob** uses this option. The argument **W2X** controls the number of instruments. When **W2X = FALSE** only **WX** are used as instruments, however when **W2X = TRUE** **WX** and **W²X** are used as instruments

for $\mathbf{W}\mathbf{y}$. The function `stslshac` from **sphet** with the argument `HAC = FALSE` estimate the S2SLS estimates with homokedastic standard errors without adjusting for heteroskedasticity.

Some caution should be expressed regarding the standard errors. When the argument `robust = FALSE` is used, the variance-covariance matrix is computed as:

$$\hat{\mathbf{V}}(\hat{\delta}_{2SLS}) = \hat{\sigma}_\epsilon^2 [\mathbf{Z}^\top \mathbf{Z}]^{-1}$$

where:

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\epsilon}}^\top \hat{\boldsymbol{\epsilon}}}{n - K}, \quad \hat{\boldsymbol{\epsilon}} = \mathbf{y} - \hat{\mathbf{y}}$$

Note that the error variance is calculated with a degrees of freedom correction (i.e., dividing by $n - K$). When `robust = TRUE` the variance-covariance matrix is computed as we have previously stated. That is:

$$\hat{\mathbf{V}}(\hat{\delta}_{2SLS}) = \hat{\sigma}_\epsilon^2 [\mathbf{Z}^\top \mathbf{H}(\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \mathbf{Z}]^{-1}$$

The results are presented in Table 6.1.

LeSage (2014, pag. 24) points out that researcher should consider performance of estimation procedures, not simply point estimates. That is, when comparing models we should also focus on the scalar summaries of the partial derivatives (direct/indirect effects estimates) and their standard errors. That is, methods that seems superior in terms of bias of the parameters might performance worse in terms of partial effects.

Now we compare the direct and indirect effects:

```
im_ml      <- impacts(slm, listw = listw, R = 200)

## Error in UseMethod("impacts", obj): no applicable method for 'impacts' applied
## to an object of class "Sarlm"

im_s2sls <- impacts(s2sls_rob, listw = listw, R = 200)

## Error in UseMethod("impacts", obj): no applicable method for 'impacts' applied
## to an object of class "Stsls"

summary(im_ml, zstats = TRUE, short = TRUE)

## Error in h(simpleError(msg, call)): error in evaluating the argument 'object'
## in selecting a method for function 'summary': object 'im_ml' not found

summary(im_s2sls, zstats = TRUE, short = TRUE)

## Error in h(simpleError(msg, call)): error in evaluating the argument 'object'
## in selecting a method for function 'summary': object 'im_s2sls' not found
```

6.3 Generalized Moment Estimation of SEM Model

Kelejian and Prucha (1999) derive a Method of Moments (MOM) estimator for λ in order to use it later in a FGLS estimator. The main Kelejian and Prucha (1999)'s motivation

Table 6.1: Spatial Models for Crime in Columbus: ML vs S2SLS

	SLM	S2SLS	S2SLSR
<i>Constant</i>	46.851*** (7.315)	44.116*** (11.172)	44.116*** (7.632)
INC	-1.074*** (0.311)	-1.008** (0.391)	-1.008* (0.458)
HOVAL	-0.270** (0.090)	-0.270** (0.093)	-0.270 (0.174)
ρ	0.404*** (0.121)	0.455* (0.191)	0.455** (0.141)
N	49	49	49
Significance: *** $\equiv p < 0.001$; ** $\equiv p < 0.01$; * $\equiv p < 0.05$			

to derive this new estimator is that the (quasi) maximum likelihood estimator may not be computationally feasible in many cases involving moderate- or large-sized samples. As they state, the MOM estimator is computationally simple irrespective of the sample size, which makes it very attractive if we have a very large spatial data base. Since the IV/GMM estimators ignore the Jacobian term, many of the problems related with matrix inversion, the computation of characteristic roots and/or Cholesky decomposition could be avoided. Another motivation was that at the time there were no formal results available regarding the consistency and asymptotic normality of the ML estimator (Prucha, 2014, pag. 1608). Recall that Lee formally derived the asymptotic properties of the ML in 2004 for the SLM.³

Recall that the SEM model is given by:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \\ \mathbf{u} &= \lambda \mathbf{M}\mathbf{u} + \boldsymbol{\varepsilon}. \end{aligned} \tag{6.32}$$

In brief, Kelejian and Prucha (1999) suggest the use of **nonlinear least square** to obtain a consistent generalized moment estimator for λ , which can be used to obtain consistent estimators for $\boldsymbol{\beta}$ in a FGLS approach. The main difference between the MOM estimation discussed here and the Generalized Method of Moment (GMM) estimation discussed later is that in the former there is no inference for the spatial autoregressive coefficient. In other words, λ is viewed purely as a nuisance parameter, whose only function is to aid in obtaining consistent estimates for $\boldsymbol{\beta}$.

R The MOM procedure proposed by Kelejian and Prucha (1999) was originally motivated by the computational difficulties of the ML.

R Kelejian and Prucha (1999) does not provide an asymptotic variance for λ . Thus, some software just provide the estimate $\hat{\lambda}$, but not its standard error.

³The consistency and asymptotic normality of the ML estimator for the SEM and SAC model remain to be derived.

One advantage of the MOM estimator (and of QML) is that they do not rely on the assumption of normality of the disturbances ϵ . Nonetheless, both estimators assume that ϵ_i are independently and identically distributed for all i with zero mean and variance σ^2 . To begin with, we state the same assumption about the error terms as in [Kelejian and Prucha \(1999\)](#).

Assumption 6.9 — Homokedastic Errors ([Kelejian and Prucha, 1999](#)). The innovations $\{\epsilon_{i,n}, 1 \leq i \leq n, n \geq 1\}$ are independently and identically distributed for all n with zero mean and variance σ^2 , where $0 < \sigma^2 < b$, with $b < \infty$. Additionally, the innovations are assumed to possess finite fourth moments.

Now we state the following assumptions:

Assumption 6.10 — Weight Matrix M_n ([Kelejian and Prucha, 1999](#)). Assume the following:

- (a) All diagonal elements of the spatial weighting matrix M_n are zero.
- (b) The matrix $(I_n - \lambda_0 M_n)$ is nonsingular with $|\lambda_0| < 1$.

Given Equation (6.32), and Assumption 6.10 (Weight Matrix M_n), we can write $\mathbf{u}_n = (I_n - \lambda_0 M_n)^{-1} \epsilon_n$. Therefore, the expectation and variance of \mathbf{u}_n are $\mathbb{E}(\mathbf{u}_n) = 0$ and $\mathbb{E}(\mathbf{u}_n \mathbf{u}_n^\top) = \Omega_n(\lambda_0)$, respectively, where:

$$\Omega_n(\lambda_0) = \sigma_{\epsilon,n}^2 (I_n - \lambda_0 M_n)^{-1} (I_n - \lambda_0 M_n^\top)^{-1}.$$

Note that a row-standardized spatial weight matrix is typically not symmetric, such that $M_n \neq M_n^\top$ and thus $(I_n - \lambda_0 M_n)^{-1} \neq (I_n - \lambda_0 M_n^\top)^{-1}$.

6.3.1 Spatially Weighted Least Squares

The key issue in [Kelejian and Prucha \(1999\)](#) is to find a **consistent estimator** of λ so that the consistency of the resulting spatially weighted estimator is assured. Under this approach, [Kelejian and Prucha \(1999\)](#) were not necessarily interested in inference about λ per se, but only interested in its estimate as a way to obtain estimates for β . This implies that λ is considered a **nuisance parameter**.

The spatially weighted least squares (SWLS) boils down to:

$$\hat{\beta}_{SWLS} = (\mathbf{X}_s^\top \mathbf{X}_s)^{-1} \mathbf{X}_s^\top \mathbf{y}_s, \quad (6.33)$$

where $\mathbf{X}_s = \mathbf{X} - \hat{\lambda} \mathbf{W} \mathbf{X}$ and $\mathbf{y}_s = \mathbf{y} - \hat{\lambda} \mathbf{M} \mathbf{y}$, using a consistent estimate $\hat{\lambda}$ for the autoregressive parameter.

Note that this model is basically OLS applied to spatially filtered variables. Furthermore, it should be noted that the SWLS are nothing but a special case of feasible generalized least squares (FGLS). To note this consider the homoskedastic case, with $\mathbb{E}[\epsilon^\top \epsilon] = \sigma^2 I_n$. Consequently:

$$\mathbb{E}[\mathbf{u} \mathbf{u}^\top] = \Omega = \sigma^2 [(I_n - \lambda \mathbf{M})^\top (I_n - \lambda \mathbf{M})]^{-1}, \quad (6.34)$$

and the corresponding generalized least squares (GLS) estimator—assuming we know λ_0 —for β is:

$$\hat{\beta}_{GLS} = [\mathbf{X}^\top \boldsymbol{\Omega}^{-1} \mathbf{X}]^{-1} \mathbf{X}^\top \boldsymbol{\Omega}^{-1} \mathbf{y}.$$

From Equation (6.34), it can be observed that $\boldsymbol{\Omega}$ contains a matrix inverse, so that its inverse, $\boldsymbol{\Omega}^{-1}$ is simple the product of the two spatial filters scaled by σ^2 . Thus, the expression for the GLS estimator simplifies to:

$$\begin{aligned} \hat{\beta}_{GLS} &= \left[\mathbf{X}^\top \frac{1}{\sigma^2} (\mathbf{I}_n - \lambda \mathbf{M})^\top (\mathbf{I}_n - \lambda \mathbf{M}) \mathbf{X} \right]^{-1} \mathbf{X}^\top \frac{1}{\sigma^2} (\mathbf{I}_n - \lambda \mathbf{M})^\top (\mathbf{I}_n - \lambda \mathbf{M}) \mathbf{y}, \\ &= \left[\mathbf{X}^\top (\mathbf{I}_n - \lambda \mathbf{M})^\top (\mathbf{I}_n - \lambda \mathbf{M}) \mathbf{X} \right]^{-1} \mathbf{X}^\top (\mathbf{I}_n - \lambda \mathbf{M})^\top (\mathbf{I}_n - \lambda \mathbf{M}) \mathbf{y}. \end{aligned}$$

The FGLS estimator substitutes a consistent estimate for λ into this expression, as:

$$\hat{\beta}_{FGLS} = \left[\mathbf{X}^\top (\mathbf{I}_n - \hat{\lambda} \mathbf{M})^\top (\mathbf{I}_n - \hat{\lambda} \mathbf{M}) \mathbf{X} \right]^{-1} \mathbf{X}^\top (\mathbf{I}_n - \hat{\lambda} \mathbf{M})^\top (\mathbf{I}_n - \hat{\lambda} \mathbf{M}) \mathbf{y},$$

which is the same as Equation (6.33).

6.3.2 Moment Conditions

The basic idea behind a method of moments estimator is to find a set of population moments equations that provide a relationship between population moments and parameters. Then, we replace the population moments using sample moments to obtain a consistent estimate of λ to plug into Equation (6.33).

Given the DGP in Equation (6.32), we can write:

$$\boldsymbol{\varepsilon} = \mathbf{u} - \lambda \mathbf{M} \mathbf{u},$$

where $\boldsymbol{\varepsilon}$ is the idiosyncratic error and \mathbf{u} is the regression error. The GM estimation approach employs the following simple **quadratic moment conditions**:⁴

$$\begin{aligned} \mathbb{E} [n^{-1} \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}] &= \sigma^2, \\ \mathbb{E} [n^{-1} \boldsymbol{\varepsilon}^\top \mathbf{M} \mathbf{M} \boldsymbol{\varepsilon}] &= \frac{\sigma^2}{n} \mathbb{E} [\text{tr}(\mathbf{M}^\top \mathbf{M} \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top)], \\ \mathbb{E} [n^{-1} \boldsymbol{\varepsilon}^\top \mathbf{M} \boldsymbol{\varepsilon}] &= 0. \end{aligned}$$

The Kelejian and Prucha (1999)'s GM estimator of λ is based on these three moments. The final value of $\mathbb{E} [n^{-1} \boldsymbol{\varepsilon}^\top \mathbf{M} \mathbf{M} \boldsymbol{\varepsilon}]$ will depend on the assumption about the variance of $\boldsymbol{\varepsilon}$. If we assume heterokedasticity then:

$$\begin{aligned} \mathbb{E} [n^{-1} \boldsymbol{\varepsilon}^\top \mathbf{M} \mathbf{M} \boldsymbol{\varepsilon}] &= \mathbb{E} [n^{-1} \text{tr}(\boldsymbol{\varepsilon}^\top \mathbf{M} \mathbf{M} \boldsymbol{\varepsilon})] \\ &= n^{-1} \text{tr} [\mathbf{W} \text{diag} [\mathbb{E}(\epsilon_i^2)] \mathbf{W}^\top] \end{aligned}$$

where we use the fact that $\text{tr}(\mathbf{X}^\top \mathbf{A} \mathbf{X}) = \mathbf{X}^\top \mathbf{A} \mathbf{X} = \text{tr}(\mathbf{A} \mathbf{X} \mathbf{X}^\top)$. Furthermore, note that under homokedasticity as in Kelejian and Prucha (1999) we obtain:

⁴Please, derive these moment conditions.

$$\mathbb{E} \left[n^{-1} \boldsymbol{\varepsilon}^\top \mathbf{M} \mathbf{M} \boldsymbol{\varepsilon} \right] = \frac{\sigma^2}{n} \text{tr} \left(\mathbf{M}^\top \mathbf{M} \right)$$

Definition 6.3.1 — Moment Conditions. Under homoskedasticity (Kelejian and Prucha, 1999) the moment conditions are:

$$\begin{aligned} \mathbb{E} \left[n^{-1} \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} \right] &= \sigma^2, \\ \mathbb{E} \left[n^{-1} \boldsymbol{\varepsilon}^\top \mathbf{M} \mathbf{M} \boldsymbol{\varepsilon} \right] &= \frac{\sigma^2}{n} \text{tr} \left(\mathbf{M}^\top \mathbf{M} \right), \\ \mathbb{E} \left[n^{-1} \boldsymbol{\varepsilon}^\top \mathbf{M} \boldsymbol{\varepsilon} \right] &= 0. \end{aligned}$$

Under heterokedasticity (Kelejian and Prucha, 2010) the moment conditions are:

$$\begin{aligned} \mathbb{E} \left[n^{-1} \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} \right] &= \sigma^2, \\ \mathbb{E} \left[n^{-1} \boldsymbol{\varepsilon}^\top \mathbf{M} \mathbf{M} \boldsymbol{\varepsilon} \right] &= n^{-1} \text{tr} \left[\mathbf{W} \text{diag} \left[\mathbb{E}(\epsilon_i^2) \right] \mathbf{W}^\top \right], \\ \mathbb{E} \left[n^{-1} \boldsymbol{\varepsilon}^\top \mathbf{M} \boldsymbol{\varepsilon} \right] &= 0. \end{aligned}$$

In order to operationalize the moment conditions, we need to convert conditions on $\boldsymbol{\varepsilon}$ into conditions on \mathbf{u} (since $\boldsymbol{\varepsilon}$ is not observed). Since $\mathbf{u} = \lambda \mathbf{M} \mathbf{u} + \boldsymbol{\varepsilon}$ it follows that $\boldsymbol{\varepsilon} = \mathbf{u} - \lambda \mathbf{M} \mathbf{u}$, i.e., the spatially filtered regression error terms.

$$\begin{aligned} \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} &= (\mathbf{u} - \lambda \mathbf{M} \mathbf{u})^\top (\mathbf{u} - \lambda \mathbf{M} \mathbf{u}) \\ &= \mathbf{u}^\top \mathbf{u} - 2\lambda \mathbf{u}^\top \mathbf{M} \mathbf{u} + \lambda^2 \mathbf{u}^\top \mathbf{M}^\top \mathbf{M} \mathbf{u} \end{aligned} \quad (6.35)$$

$$\begin{aligned} \boldsymbol{\varepsilon}^\top \mathbf{M}^\top \mathbf{M} \boldsymbol{\varepsilon} &= (\mathbf{u} - \lambda \mathbf{M} \mathbf{u})^\top \mathbf{M}^\top \mathbf{M} (\mathbf{u} - \lambda \mathbf{M} \mathbf{u}) \\ &= \mathbf{u}^\top \mathbf{M}^\top \mathbf{M} \mathbf{u} - 2\lambda \mathbf{u}^\top \mathbf{M}^\top \mathbf{M} \mathbf{M} \mathbf{u} + \lambda^2 \mathbf{u}^\top \mathbf{M}^\top \mathbf{M} \mathbf{M}^\top \mathbf{M} \mathbf{u} \end{aligned} \quad (6.36)$$

$$\begin{aligned} \boldsymbol{\varepsilon}^\top \mathbf{M} \boldsymbol{\varepsilon} &= (\mathbf{u} - \lambda \mathbf{M} \mathbf{u})^\top \mathbf{M} (\mathbf{u} - \lambda \mathbf{M} \mathbf{u}) \\ &= \mathbf{u}^\top \mathbf{M} \mathbf{u} - 2\lambda \mathbf{u}^\top \mathbf{M} \mathbf{M} \mathbf{u} + \lambda^2 \mathbf{u}^\top \mathbf{M}^\top \mathbf{M} \mathbf{M} \mathbf{u} \end{aligned} \quad (6.37)$$

Let $\mathbf{u}_L = \mathbf{M} \mathbf{u}$, $\mathbf{u}_{LL} = \mathbf{M} \mathbf{M} \mathbf{u}$.⁵ Taking the expectation over (6.35) and assuming Homokedasticity by Assumption 6.9, we get:

$$\begin{aligned} \mathbb{E} \left[\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} \right] &= \mathbb{E} \left[\mathbf{u}^\top \mathbf{u} \right] - 2\lambda \mathbb{E} \left[\mathbf{u}^\top \mathbf{M} \mathbf{u} \right] + \lambda^2 \mathbb{E} \left[\mathbf{u}^\top \mathbf{M}^\top \mathbf{M} \mathbf{u} \right] \\ \sigma^2 &= \frac{1}{n} \mathbb{E} \left[\mathbf{u}^\top \mathbf{u} \right] - \lambda \frac{2}{n} \mathbb{E} \left[\mathbf{u}^\top \mathbf{u}_L \right] + \lambda^2 \frac{1}{n} \mathbb{E} \left[\mathbf{u}_L^\top \mathbf{u}_L \right] \quad \text{since } \mathbb{E} \left[n^{-1} \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} \right] = \sigma^2 \\ 0 &= \sigma^2 - \frac{1}{n} \mathbb{E} \left[\mathbf{u}^\top \mathbf{u} \right] + \lambda \frac{2}{n} \mathbb{E} \left[\mathbf{u}^\top \mathbf{u}_L \right] - \lambda^2 \frac{1}{n} \mathbb{E} \left[\mathbf{u}_L^\top \mathbf{u}_L \right] \\ 0 &= \lambda \frac{2}{n} \mathbb{E} \left[\mathbf{u}^\top \mathbf{u}_L \right] - \lambda^2 \frac{1}{n} \mathbb{E} \left[\mathbf{u}_L^\top \mathbf{u}_L \right] + \frac{1}{n} \sigma^2 - \frac{1}{n} \mathbb{E} \left[\mathbf{u}^\top \mathbf{u} \right] \\ 0 &= \begin{pmatrix} \frac{2}{n} \mathbb{E} \left[\mathbf{u}^\top \mathbf{u}_L \right] & -\frac{1}{n} \mathbb{E} \left[\mathbf{u}_L^\top \mathbf{u}_L \right] & 1 \end{pmatrix} \begin{pmatrix} \lambda \\ \lambda^2 \\ \sigma^2 \end{pmatrix} - \frac{1}{n} \mathbb{E} \left[\mathbf{u}^\top \mathbf{u} \right] \end{aligned} \quad (6.38)$$

⁵Spatially lagged variables are denoted by bar superscripts in the articles. Instead, we will use the L subscript throughout. That is, a first order spatial lag of \mathbf{y} , $\mathbf{W} \mathbf{y}$, is denoted by \mathbf{y}_L . Higher order spatial lags are symbolized by adding additional L subscripts.

In similar fashion,

$$0 = \begin{pmatrix} \frac{2}{n}\mathbb{E}[\mathbf{u}_{LL}^\top \mathbf{u}_L] & -\frac{1}{n}\mathbb{E}[\mathbf{u}_{LL}^\top \mathbf{u}_{LL}] & \frac{1}{n}\text{tr}(\mathbf{M}^\top \mathbf{M}) \end{pmatrix} \begin{pmatrix} \lambda \\ \lambda^2 \\ \sigma^2 \end{pmatrix} - \frac{1}{n}\mathbb{E}[\mathbf{u}_L^\top \mathbf{u}_L] \quad (6.39)$$

$$0 = \begin{pmatrix} \frac{1}{n}\mathbb{E}[\mathbf{u}^\top \mathbf{u}_{LL} + \mathbf{u}_L^\top \mathbf{u}_L] & -\frac{1}{n}\mathbb{E}[\mathbf{u}_L^\top \mathbf{u}_{LL}] & 0 \end{pmatrix} \begin{pmatrix} \lambda \\ \lambda^2 \\ \sigma^2 \end{pmatrix} - \frac{1}{n}\mathbb{E}[\mathbf{u}^\top \mathbf{u}_L] \quad (6.40)$$

At this point it is important to realized that we have have three equations an three unknowns, λ , λ^2 and σ^2 . Consider the following three-equations system implied by Equations (6.38), (6.39) and (6.40):

$$\mathbf{\Gamma}_n \boldsymbol{\alpha} = \boldsymbol{\gamma}_n \quad (6.41)$$

where $\mathbf{\Gamma}_n$ is given in Equation (6.42), and $\boldsymbol{\alpha} = (\lambda, \lambda^2, \sigma^2)$.⁶ If $\mathbf{\Gamma}_n$ where known, Assumption 6.13 (Identification) implies that Equation (6.41) determines $\boldsymbol{\alpha}$ as:

$$\boldsymbol{\alpha} = \mathbf{\Gamma}_n^{-1} \boldsymbol{\gamma}_n$$

where:

$$\mathbf{\Gamma}_n = \begin{pmatrix} \frac{2}{n}\mathbb{E}[\mathbf{u}^\top \mathbf{u}_L] & -\frac{1}{n}\mathbb{E}[\mathbf{u}_L^\top \mathbf{u}_L] & 1 \\ \frac{2}{n}\mathbb{E}[\mathbf{u}_{LL}^\top \mathbf{u}_L] & -\frac{1}{n}\mathbb{E}[\mathbf{u}_{LL}^\top \mathbf{u}_{LL}] & \frac{1}{n}\text{tr}(\mathbf{M}^\top \mathbf{M}) \\ \frac{1}{n}\mathbb{E}[\mathbf{u}^\top \mathbf{u}_{LL} + \mathbf{u}_L^\top \mathbf{u}_L] & -\frac{1}{n}\mathbb{E}[\mathbf{u}_L^\top \mathbf{u}_{LL}] & 0 \end{pmatrix} \quad (6.42)$$

and

$$\boldsymbol{\gamma}_n = \begin{pmatrix} \frac{1}{n}\mathbb{E}[\mathbf{u}^\top \mathbf{u}] \\ \frac{1}{n}\mathbb{E}[\mathbf{u}_L^\top \mathbf{u}_L] \\ \frac{1}{n}\mathbb{E}[\mathbf{u}^\top \mathbf{u}_L] \end{pmatrix} \quad (6.43)$$

Now we express the moment conditions $\boldsymbol{\gamma}_n = \mathbf{\Gamma}_n \boldsymbol{\alpha}$ as sample averages in observables spatial lags of OLS residuals:

$$\mathbf{g}_n = \mathbf{G}_n \boldsymbol{\alpha} + \mathbf{v}_n(\lambda, \sigma^2) \quad (6.44)$$

Note also that

$$\mathbf{G}_n = \begin{pmatrix} \frac{2}{n}\hat{\mathbf{u}}^\top \hat{\mathbf{u}}_L & -\frac{1}{n}\hat{\mathbf{u}}_L^\top \hat{\mathbf{u}}_L & 1 \\ \frac{2}{n}\hat{\mathbf{u}}_{LL}^\top \hat{\mathbf{u}}_L & -\frac{1}{n}\hat{\mathbf{u}}_{LL}^\top \hat{\mathbf{u}}_{LL} & \frac{1}{n}\text{tr}(\mathbf{M}^\top \mathbf{M}) \\ \frac{1}{n}[\hat{\mathbf{u}}^\top \hat{\mathbf{u}}_{LL} + \hat{\mathbf{u}}_L^\top \hat{\mathbf{u}}_L] & -\frac{1}{n}\hat{\mathbf{u}}_L^\top \hat{\mathbf{u}}_{LL} & 0 \end{pmatrix}$$

and

$$\mathbf{g}_n = \begin{pmatrix} \frac{1}{n}\hat{\mathbf{u}}^\top \hat{\mathbf{u}} \\ \frac{1}{n}\hat{\mathbf{u}}_L^\top \hat{\mathbf{u}}_L \\ \frac{1}{n}\hat{\mathbf{u}}^\top \hat{\mathbf{u}}_L \end{pmatrix}$$

⁶Note that we are assuming that λ^2 is a new parameter.

where \mathbf{G}_n is a 3×3 matrix, and where $\mathbf{v}_n(\lambda, \sigma^2)$ can be viewed as a vector of residuals. This can be thought as a OLS regression where (Kelejian and Prucha, 1998):

$$\tilde{\boldsymbol{\alpha}}_n = \mathbf{G}_n^{-1} \mathbf{g}_n \quad (6.45)$$

However, the estimator in (6.45) is based on an overparameterization in the sense that it does not use the information that the second element of $\boldsymbol{\alpha}$, λ^2 , is the squared of the first. Given this, Kelejian and Prucha (1998) and Kelejian and Prucha (1999) define the GM estimator for λ and σ^2 as the nonlinear least square estimator corresponding to Equation (6.44):⁷

$$(\hat{\lambda}_{NLS,n}, \hat{\sigma}_{NLS,n}^2) = \operatorname{argmin} \left\{ \mathbf{v}_n(\lambda, \sigma^2)^\top \mathbf{v}_n(\lambda, \sigma^2) : \lambda \in [-a, a], \sigma^2 \in [0, b] \right\} \quad (6.46)$$

Note that $(\hat{\lambda}_{NLS,n}, \hat{\sigma}_{NLS,n}^2)$ are defined as the minimizers of

$$\left[\mathbf{g}_n - \mathbf{G}_n \begin{pmatrix} \lambda \\ \lambda^2 \\ \sigma^2 \end{pmatrix} \right]^\top \left[\mathbf{g}_n - \mathbf{G}_n \begin{pmatrix} \lambda \\ \lambda^2 \\ \sigma^2 \end{pmatrix} \right]$$

Assumption 6.11 — Bounded Matrices (Kelejian and Prucha, 1999). The row and column sums of the matrices \mathbf{M}_n and $(\mathbf{I} - \lambda \mathbf{M}_n)$ are bounded uniformly in absolute value.

Assumption 6.12 — Residuals (Kelejian and Prucha, 1999). Let $\tilde{u}_{i,n}$ denote the i -th element of $\tilde{\mathbf{u}}_n$. We then assume that

$$\tilde{u}_{i,n} - u_{i,n} = \mathbf{d}_{i,n} \boldsymbol{\Delta}_n$$

where $\mathbf{d}_{i,n}$ and $\boldsymbol{\Delta}_n$ are $1 \times p$ and $p \times 1$ dimensional random vectors. Let $d_{ij,n}$ be the j th element of $\mathbf{d}_{i,n}$. Then, we assume that for some $\delta > 0$, $\mathbb{E} |d_{ij,n}|^{2+\delta} \leq c_d < \infty$, where c_d does not depend on n , and that

$$\sqrt{n} \|\boldsymbol{\Delta}_n\| = O_p(1). \quad (6.47)$$

This assumption should be satisfied for most cases in which $\tilde{\mathbf{u}}$ is based on \sqrt{n} -consistent estimators of the regression coefficients (non-linear OLS, linear OLS, 2SLS). Assumption 6.12 comes from Kelejian and Prucha (2010) and is a bit stronger than the same assumption in Kelejian and Prucha (1999).

Assumption 6.13 — Identification (Kelejian and Prucha, 1999). Let $\boldsymbol{\Gamma}_n$ be the matrix in Equation (6.42). The smallest eigenvalues of $\boldsymbol{\Gamma}_n^\top \boldsymbol{\Gamma}_n$ is bounded away from zero, that is, $\omega_{\min}(\boldsymbol{\Gamma}_n^\top \boldsymbol{\Gamma}_n) \geq \omega_* > 0$, where ω_* may depend on λ and σ^2

⁷They state that is more efficient than the OLS estimator. However, both estimator are consistent. See Theorem 2 in (Kelejian and Prucha, 1998).

Theorem 6.14 — Consistency. Let $(\hat{\lambda}_{NLS,n}, \hat{\sigma}_{NLS,n}^2)$ given by:

$$(\hat{\lambda}_{NLS,n}, \hat{\sigma}_{NLS,n}^2) = \operatorname{argmin} \left\{ \mathbf{v}_n(\lambda, \sigma^2)^\top \mathbf{v}_n(\lambda, \sigma^2) : \lambda \in [-a, a], \sigma^2 \in [0, b] \right\}$$

Then, given Assumptions 6.1 (Heterokedastic errors), 6.10 (Weight Matrix \mathbf{M}_n), 6.11 (Bounded Matrices), 6.12 (Residuals), and 6.13 (Identification),

$$(\hat{\lambda}_{NLS,n}, \hat{\sigma}_{NLS,n}^2) \xrightarrow{p} (\lambda, \sigma^2) \quad \text{as } n \rightarrow \infty \quad (6.48)$$

An important remark is that Theorem 6.14 states only that the NLS estimates are consistent, but it does not tell us about the asymptotic distribution of $\hat{\lambda}_{NLS,n}$.

Sketch or proof for GM estimator of $\hat{\lambda}$. The proof is based on Kelejian and Piras (2017) and consist into two steps. First, we prove consistency of $\hat{\lambda}$ for the OLS estimate of α —which is more simple—and assuming that the vector \mathbf{u} is observed. We then show that \mathbf{u} can be replaced in the GM estimator for λ by $\hat{\mathbf{u}}$. For a more general proof see Kelejian and Prucha (1998, 1999).

- (a) *Assuming that \mathbf{u} is observed.* Recall that in Equation (6.44) the sample moments are based on the estimated $\hat{\mathbf{u}}$. But, if \mathbf{u} were observed, then we would use the following sample moments:

$$\mathbf{g}_n^* = \mathbf{G}_n^* \alpha$$

where

$$\mathbf{G}_n^* = \begin{pmatrix} \frac{2}{n} \mathbf{u}^\top \mathbf{u}_L & -\frac{1}{n} \mathbf{u}_L^\top \mathbf{u}_L & 1 \\ \frac{2}{n} \mathbf{u}_{LL}^\top \mathbf{u}_L & -\frac{1}{n} \mathbf{u}_{LL}^\top \mathbf{u}_{LL} & \frac{1}{n} \operatorname{tr}(\mathbf{M}^\top \mathbf{M}) \\ \frac{1}{n} [\mathbf{u}^\top \mathbf{u}_{LL} + \mathbf{u}_L^\top \mathbf{u}_L] & -\frac{1}{n} \mathbf{u}_L^\top \mathbf{u}_{LL} & 0 \end{pmatrix}$$

and

$$\mathbf{g}_n^* = \begin{pmatrix} \frac{1}{n} \mathbf{u}^\top \mathbf{u} \\ \frac{1}{n} \mathbf{u}_L^\top \mathbf{u}_L \\ \frac{1}{n} \mathbf{u}^\top \mathbf{u}_L \end{pmatrix}$$

Recall that:

$$\begin{aligned} \mathbf{u} &= (\mathbf{I}_n - \lambda \mathbf{M})^{-1} \boldsymbol{\varepsilon} \\ \mathbf{u}_L &= \mathbf{M} (\mathbf{I}_n - \lambda \mathbf{M})^{-1} \boldsymbol{\varepsilon} \\ \mathbf{u}_{LL} &= \mathbf{M} \mathbf{M} (\mathbf{I}_n - \lambda \mathbf{M})^{-1} \boldsymbol{\varepsilon} \end{aligned}$$

and first and second column of \mathbf{G}^* are quadratic forms of $\boldsymbol{\varepsilon}$. Since \mathbf{M} is uniformly bounded then, using Theorem of consistency of quadratic forms (3.27), we can state that:

$$\mathbf{G}^* \xrightarrow{p} \mathbf{\Gamma}_n$$

Also:

$$\begin{aligned}\text{plim } \mathbf{g}_n^* &= \text{plim } \mathbf{G}^* \boldsymbol{\alpha} \\ &= \boldsymbol{\Gamma} \boldsymbol{\alpha}\end{aligned}$$

If \mathbf{u} would be observed, a linear GMM estimator for λ , say $\tilde{\lambda}$, would be the first element of the least squared estimator $\boldsymbol{\alpha}$, namely:

$$\tilde{\boldsymbol{\alpha}} = \mathbf{G}_n^{-1*} \mathbf{g}_n^*$$

since \mathbf{G}_n^* is a 3×3 matrix which is nonsingular. Thus, using our previous results:

$$\text{plim } \tilde{\boldsymbol{\alpha}} = \text{plim } \mathbf{G}_n^{-1*} \text{plim } \mathbf{g}_n^* = \boldsymbol{\Gamma}_n^{-1} \boldsymbol{\gamma}_n = \boldsymbol{\alpha} \quad (6.49)$$

- (b) *Replacing \mathbf{u} by $\hat{\mathbf{u}}$.* Now consider the estimator $\boldsymbol{\alpha}$ based on $\hat{\mathbf{u}}$. The OLS estimator is consistent and can be expressed as:

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + \Delta_n, \quad \Delta_n \xrightarrow{p} \mathbf{0}.$$

Then, the OLS estimator $\hat{\mathbf{u}}$ is:

$$\begin{aligned}\hat{\mathbf{u}} &= \mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}} \\ &= \mathbf{y} - \mathbf{X} (\boldsymbol{\beta}_0 + \Delta_n) \\ &= \mathbf{y} - \mathbf{X} \boldsymbol{\beta}_0 - \mathbf{X} \Delta_n \\ &= \mathbf{u} - \mathbf{X} \Delta_n\end{aligned}$$

Note that, with the exception of the constants in the third column of \mathbf{G}_n^* , every element of \mathbf{G}_n^* and \mathbf{g}_n^* can be expressed as a quadratic of the form $\boldsymbol{\varepsilon}^\top \mathbf{S} \boldsymbol{\varepsilon} / n$, where \mathbf{S} is an $n \times n$ matrix whose row and columns are uniformly bounded in absolute value given our assumption 6.11. For example:

$$\frac{1}{n} \mathbf{u}^\top \mathbf{u}_L = \frac{1}{n} \boldsymbol{\varepsilon}^\top (\mathbf{I}_n - \lambda \mathbf{M})^{-1\top} \mathbf{M} (\mathbf{I}_n - \lambda \mathbf{M})^{-1} \boldsymbol{\varepsilon} = \frac{1}{n} \boldsymbol{\varepsilon}^\top \mathbf{S} \boldsymbol{\varepsilon}$$

Then:

$$\begin{aligned}\frac{\hat{\mathbf{u}}^\top \mathbf{S} \hat{\mathbf{u}}}{n} &= \frac{(\mathbf{u} - \mathbf{X} \Delta_n)^\top \mathbf{S} (\mathbf{u} - \mathbf{X} \Delta_n)}{n} \\ &= \frac{\mathbf{u}^\top \mathbf{S} \mathbf{u}}{n} - \frac{2 \Delta_n^\top \mathbf{X}^\top \mathbf{S} \mathbf{u}}{n} + \frac{\Delta_n^\top \mathbf{X}^\top \mathbf{S} \mathbf{X} \Delta_n}{n}\end{aligned}$$

We need to show that (This would be part of your homework):

$$\begin{aligned}\frac{2 \Delta_n^\top \mathbf{X}^\top \mathbf{S} \mathbf{u}}{n} &\xrightarrow{p} 0 \\ \frac{\Delta_n^\top \mathbf{X}^\top \mathbf{S} \mathbf{X} \Delta_n}{n} &\xrightarrow{p} 0\end{aligned}$$

so that we can say that:

$$\frac{\hat{\mathbf{u}}^\top \mathbf{S} \hat{\mathbf{u}}}{n} \xrightarrow{p} \frac{1}{n} \mathbf{u}^\top \mathbf{S} \mathbf{u},$$

and finally say that:

$$\mathbf{g}_n \xrightarrow{p} \mathbf{g}_n^* \xrightarrow{p} \gamma_n, \quad \mathbf{G}_n \xrightarrow{p} \mathbf{G}_n^* \xrightarrow{p} \Gamma_n$$

Given Equation (6.49), consistency is proved. ■

6.3.3 Feasible Generalized Least Squares Model

In Section 6.3.1 we derived that the GLS estimator is given by:

$$\beta_{GLS}(\lambda) = [\mathbf{X}^\top \boldsymbol{\Omega}(\lambda)^{-1} \mathbf{X}]^{-1} \mathbf{X}^\top \boldsymbol{\Omega}(\lambda)^{-1} \mathbf{y}, \quad (6.50)$$

where $\boldsymbol{\Omega}(\lambda) = (\mathbf{I} - \lambda \mathbf{W})^{-1} (\mathbf{I} - \lambda \mathbf{W}^\top)^{-1}$. But now we have a consistent estimate for λ . Thus, we can get an estimate of $\hat{\beta}$ using the FGLS estimator defined as:

$$\beta_{FGLS}(\lambda) = [\mathbf{X}^\top \boldsymbol{\Omega}(\hat{\lambda})^{-1} \mathbf{X}]^{-1} \mathbf{X}^\top \boldsymbol{\Omega}(\hat{\lambda})^{-1} \mathbf{y}. \quad (6.51)$$

Assumption 6.15 — Limiting Behavior. The elements of \mathbf{X} are non-stochastic and bounded in absolute value by $c_X, 0 < c_X < \infty$. Also, \mathbf{X} has full rank, and the matrix $\mathbf{Q}_X = \lim_{n \rightarrow \infty} n^{-1} \mathbf{X}^\top \mathbf{X}$ is finite and nonsingular. Furthermore, the matrices $\mathbf{Q}_X(\lambda) = \lim_{n \rightarrow \infty} n^{-1} \mathbf{X}^\top \boldsymbol{\Omega}(\lambda)^{-1} \mathbf{X}$ is finite and nonsingular for all $|\lambda| < 1$

The following Theorem proposes the asymptotic distribution for the FGLS Estimator:

Theorem 6.16 — Asymptotic Properties of FGLS Estimator. If assumptions 6.1 (Homokedastic errors), 6.10 (Weight Matrix \mathbf{M}_n), 6.11 (Bounded Matrices), and 6.15 (Limiting Behavior) hold:

- (a) The true GLS estimator $\hat{\beta}_{GLS}$ is a consistent estimator for β , and

$$\sqrt{n} (\hat{\beta}_{GLS} - \beta) \xrightarrow{d} N(\mathbf{0}, \sigma^2 \mathbf{Q}_X(\lambda)^{-1}) \quad (6.52)$$

- (b) Let $\hat{\lambda}_n$ be a consistent estimator for λ . Then the true GLS estimator $\hat{\beta}_{GLS}$ and the feasible GLS estimator $\hat{\beta}_{FGLS}$ have the same asymptotic distribution.

- (c) Suppose further than $\hat{\sigma}_n^2$ is a consistent estimator for σ^2 . Then $\hat{\sigma}_n^2 [n^{-1} \mathbf{X}^\top \boldsymbol{\Omega}(\hat{\lambda}_n)^{-1} \mathbf{X}]$ is a consistent estimator for $\sigma^2 \mathbf{Q}_X(\lambda)^{-1}$.

Note that Theorem 6.16 assumes the existence of a consistent estimator of λ and σ^2 . It can be shown that the OLS estimator:

$$\hat{\beta}_n = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

is \sqrt{n} -consistent. Thus, the OLS residuals $\tilde{u}_i = y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_n$ satisfy Assumption 6.12 with $d_{i,n} = |\mathbf{x}_i|$ and $\Delta_n = \hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}$. Thus, OLS residuals can be used to obtain consistent estimators of λ and σ^2 .

Then, the feasible GLS is given by

$$\hat{\boldsymbol{\beta}}_{FGLS} = [\mathbf{X}^\top(\tilde{\lambda})\mathbf{X}(\tilde{\lambda})]^{-1} \mathbf{X}^\top(\tilde{\lambda})\mathbf{y}(\tilde{\lambda})$$

where:

$$\begin{aligned} \mathbf{X}(\tilde{\lambda}) &= (\mathbf{I} - \tilde{\lambda}\mathbf{M})\mathbf{X} \\ \mathbf{y}(\tilde{\lambda}) &= (\mathbf{I} - \tilde{\lambda}\mathbf{M})\mathbf{y} \end{aligned}$$

The variance covariance matrix of $\hat{\boldsymbol{\beta}}_{FGLS}$ is estimated as:

$$\hat{\mathbb{V}}(\hat{\boldsymbol{\beta}}_{FGLS}) = \hat{\sigma}^2 [\mathbf{X}^\top(\tilde{\lambda})\mathbf{X}(\tilde{\lambda})]^{-1},$$

where:

$$\begin{aligned} \hat{\sigma}^2 &= \hat{\boldsymbol{\varepsilon}}^\top(\tilde{\lambda})\hat{\boldsymbol{\varepsilon}}(\tilde{\lambda}) \\ \hat{\boldsymbol{\varepsilon}}(\tilde{\lambda}) &= \mathbf{y}(\tilde{\lambda}) - \mathbf{X}(\tilde{\lambda})\hat{\boldsymbol{\beta}}_{FGLS} = (\mathbf{I} - \tilde{\lambda}\mathbf{M})\hat{\mathbf{u}} \\ \hat{\mathbf{u}} &= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{FGLS} \end{aligned}$$

Sketch of Proof of Theorem 6.16. We first prove part (a). Recall that the GLS and FGSL estimator are given by:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{GLS} &= [\mathbf{X}^\top \boldsymbol{\Omega}(\lambda)^{-1} \mathbf{X}]^{-1} \mathbf{X}^\top \boldsymbol{\Omega}(\lambda)^{-1} \mathbf{y} \\ \hat{\boldsymbol{\beta}}_{FGLS} &= [\mathbf{X}^\top \hat{\boldsymbol{\Omega}}(\lambda)^{-1} \mathbf{X}]^{-1} \mathbf{X}^\top \hat{\boldsymbol{\Omega}}(\lambda)^{-1} \mathbf{y} \end{aligned}$$

Since $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} = \mathbf{X}\boldsymbol{\beta} + (\mathbf{I}_n - \lambda\mathbf{M})^{-1}\boldsymbol{\varepsilon}$, the sampling error of $\hat{\boldsymbol{\beta}}_{GLS}$ is,

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \boldsymbol{\beta} + [\mathbf{X}^\top \boldsymbol{\Omega}(\lambda)^{-1} \mathbf{X}]^{-1} \mathbf{X}^\top \boldsymbol{\Omega}(\lambda)^{-1} \mathbf{u} \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= [\mathbf{X}^\top \boldsymbol{\Omega}(\lambda)^{-1} \mathbf{X}]^{-1} \mathbf{X}^\top (\mathbf{I}_n - \lambda\mathbf{M})^\top (\mathbf{I}_n - \lambda\mathbf{M}) (\mathbf{I}_n - \lambda\mathbf{M})^{-1} \boldsymbol{\varepsilon} \\ \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= [\mathbf{X}^\top \boldsymbol{\Omega}(\lambda)^{-1} \mathbf{X}]^{-1} \mathbf{X}^\top (\mathbf{I}_n - \lambda\mathbf{M})^\top \boldsymbol{\varepsilon} \\ \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= \left[\frac{1}{n} \mathbf{X}^\top \boldsymbol{\Omega}(\lambda)^{-1} \mathbf{X} \right]^{-1} \frac{1}{\sqrt{n}} \mathbf{A}^\top \boldsymbol{\varepsilon} \end{aligned}$$

where $\mathbf{A} = (\mathbf{I}_n - \lambda\mathbf{M})\mathbf{X}$. By Assumption 6.15 (Limiting Behavior):

$$\frac{1}{n} \mathbf{X}^\top \boldsymbol{\Omega}(\lambda)^{-1} \mathbf{X} \rightarrow \mathbf{Q}_X(\lambda)$$

Since \mathbf{Q}_X is not singular:

$$\left[\frac{1}{n} \mathbf{X}^\top \boldsymbol{\Omega}(\lambda)^{-1} \mathbf{X} \right]^{-1} \rightarrow \mathbf{Q}_X^{-1}(\lambda)$$

Since \mathbf{A} is bounded in absolute value, by Theorem 3.28 it follows that:

$$\frac{1}{\sqrt{n}} \mathbf{A}^\top \boldsymbol{\varepsilon} \xrightarrow{d} N \left(\mathbf{0}, \lim_{n \rightarrow \infty} n^{-1} \sigma^2 \mathbf{A}^\top \mathbf{A} \right) \quad (6.53)$$

where $\lim_{n \rightarrow \infty} n^{-1} \sigma^2 \mathbf{A}^\top \mathbf{A} = \sigma^2 \lim_{n \rightarrow \infty} n^{-1} \mathbf{X}^\top (\mathbf{I}_n - \lambda \mathbf{M})^\top (\mathbf{I}_n - \lambda \mathbf{M}) \mathbf{X} = \sigma^2 \mathbf{Q}_X(\lambda)$. Consequently:

$$\begin{aligned} \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= \underbrace{\left[\frac{1}{n} \mathbf{X}^\top \boldsymbol{\Omega}(\lambda)^{-1} \mathbf{X} \right]^{-1}}_{\rightarrow \mathbf{Q}_X^{-1}(\lambda)} \underbrace{\frac{1}{\sqrt{n}} \mathbf{A}^\top \boldsymbol{\varepsilon}}_{\xrightarrow{d} N(\mathbf{0}, \sigma^2 \mathbf{Q}_X(\lambda))} \\ &\xrightarrow{d} N \left[\mathbf{0}, \mathbf{Q}_X^{-1}(\lambda) \sigma^2 \mathbf{Q}_X(\lambda) \mathbf{Q}_X^{-1}(\lambda)^\top \right] \\ &\xrightarrow{d} N \left[\mathbf{0}, \sigma^2 \mathbf{Q}_X^{-1}(\lambda) \right] \end{aligned}$$

This also implies that $\hat{\boldsymbol{\beta}}_{GLS}$ is consistent. To show part (b), we can show that:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{GLS} - \hat{\boldsymbol{\beta}}_{FGLS}) \xrightarrow{p} 0$$

Following [Kelejian and Prucha \(1999\)](#), it suffices to show that

$$\frac{1}{n} \mathbf{X}^\top \left[\boldsymbol{\Omega}(\hat{\lambda}_n)^{-1} - \boldsymbol{\Omega}(\lambda)^{-1} \right] \mathbf{X} \xrightarrow{p} \mathbf{0} \quad (6.54)$$

and

$$\frac{1}{n} \mathbf{X}^\top \left[\boldsymbol{\Omega}(\hat{\lambda}_n)^{-1} - \boldsymbol{\Omega}(\lambda)^{-1} \right] \mathbf{u} \xrightarrow{p} \mathbf{0}$$

Note that:

$$\boldsymbol{\Omega}(\hat{\lambda}_n)^{-1} - \boldsymbol{\Omega}(\lambda)^{-1} = (\lambda - \hat{\lambda}_n)(\mathbf{M} + \mathbf{M}^\top) + (\lambda^2 - \hat{\lambda}_n^2) \mathbf{M}^\top \mathbf{M}$$

Then using the fact that we have summable matrices,

$$\frac{1}{n} \mathbf{X}^\top \left[\boldsymbol{\Omega}(\hat{\lambda}_n)^{-1} - \boldsymbol{\Omega}(\lambda)^{-1} \right] \mathbf{X} = \underbrace{(\lambda - \hat{\lambda}_n) n^{-1} \mathbf{X}^\top (\mathbf{M} + \mathbf{M}^\top) \mathbf{X}}_{\xrightarrow{p} 0} + \underbrace{(\lambda^2 - \hat{\lambda}_n^2) n^{-1} \mathbf{X}^\top \mathbf{M}^\top \mathbf{M} \mathbf{X}}_{\xrightarrow{p} 0}$$

where $(\lambda - \hat{\lambda}_n) = o_p(1)$ since $\hat{\lambda}_n$ is a consistent estimate of λ , and :

$$\begin{aligned} \frac{1}{n} \mathbf{X}^\top \left[\boldsymbol{\Omega}(\hat{\lambda}_n)^{-1} - \boldsymbol{\Omega}(\lambda)^{-1} \right] \mathbf{u} &= \underbrace{(\lambda - \hat{\lambda}_n) n^{-1/2} \mathbf{X}^\top (\mathbf{M} + \mathbf{M}^\top) \mathbf{u}}_{\xrightarrow{p} 0} + \underbrace{(\lambda^2 - \hat{\lambda}_n^2) n^{-1/2} \mathbf{X}^\top \mathbf{M}^\top \mathbf{M} \mathbf{u}}_{\xrightarrow{p} 0} \\ &= o_p(1) * O_p(1) + o_p(1) * O_p(1) \\ &= o_p(1) + o_p(1) \\ &= o_p(1) \\ &\xrightarrow{p} 0 \end{aligned} \quad (6.55)$$

To see that $n^{-1/2} \mathbf{X}^\top (\mathbf{M} + \mathbf{M}^\top) \mathbf{u} = O_p(1)$ note

$$\begin{aligned}\mathbb{E}\left[n^{-1/2}\mathbf{X}^\top(\mathbf{M} + \mathbf{M}^\top)\mathbf{u}\right] &= 0 \\ \mathbb{V}\left[n^{-1/2}\mathbf{X}^\top(\mathbf{M} + \mathbf{M}^\top)\mathbf{u}\right] &= n^{-1}\mathbf{X}^\top \underbrace{(\mathbf{M} + \mathbf{M}^\top)\boldsymbol{\Omega}(\mathbf{M}^\top + \mathbf{M})\mathbf{X}}_{\substack{\text{absolutely summable} \\ O(n)}} = O(1)\end{aligned}$$

A similar result holds for $n^{-1/2}\mathbf{X}^\top\mathbf{M}^\top\mathbf{M}\mathbf{u}$.

Part 3 of the theorem follows from (6.54) and the fact that $\hat{\sigma}^2$ is a consistent estimator for σ^2 . ■

A Feasible GLS (FGLS) can be obtained along with the following steps:

Algorithm 6.17 — GLS (FGLS) Algorithm of SEM. The steps are the following:

- (a) First of all obtain a consistent estimate of β , say $\tilde{\beta}$ using either OLS or NLS.
- (b) Use this estimate to obtain an estimate of \mathbf{u} , say $\hat{\mathbf{u}}$,
- (c) Use $\hat{\mathbf{u}}$, to estimate λ , say $\hat{\lambda}$, using (6.46),
- (d) Estimate β using Equation (6.51)

6.3.4 FGLS in R

The estimation procedure by GM is carried out by the `GMerrorsar` function from **spatialreg** package. In order to show its functionalities we first load the required packages and dataset:

```
# Load data and packages
library("memisc")
library("spdep")
library("spatialreg")
data("columbus")
listw <- nb2listw(col.gal.nb)
source("getSummary.sarlm.R")
```

Now we estimate the SEM model by ML using Ord's eigen approximation of the determinant and the Kelejian and Prucha (1999)'s GM procedure:

```
# Estimate the SEM model by ML and GM
sem_ml <- errorsarlm(CRIME ~ INC + HOVAL,
  data = columbus,
  listw,
  method = "eigen")
sem_mm <- GMerrorsar(CRIME ~ HOVAL + INC,
  data = columbus,
  listw = listw,
  returnHcov = TRUE)
```

6.4. ESTIMATION OF SAC MODEL: THE FEASIBLE GENERALIZED TWO STAGE LEAST SQUARE

A Hausman test comparing an OLS and SEM model can be obtained using

```
# Hausman test
summary(sem_mm, Hausman = TRUE)

##
## Call:GMerrorsar(formula = CRIME ~ HOVAL + INC, data = columbus, listw = listw,
##      returnHcov = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -30.8212  -6.8764  -2.1781   9.5693  28.5779
##
## Type: GM SAR estimator
## Coefficients: (GM standard errors)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 63.487150   5.083612 12.4886 < 2.2e-16
## HOVAL       -0.300365   0.096799  -3.1030 0.0019160
## INC        -1.180414   0.341788  -3.4536 0.0005531
##
## Lambda: 0.3643 (standard error): 0.4318 (z-value): 0.84366
## Residual variance (sigma squared): 109.37, (sigma: 10.458)
## GM argmin sigma squared: 108.93
## Number of observations: 49
## Number of parameters estimated: 5
## Hausman test: 6.4506, df: 3, p-value: 0.091633
```

The default model specification shown above. The output follows the familiar R format. Note that even though the estimation procedure is the GM, the output presents inference for λ . In this case, the inference is based on the analytical method described in <http://econweb.umd.edu/~prucha/STATPROG/OLS/desols.pdf>. The output also shows the Hausman test. Recall that this test can be used whenever there are two estimators, one of which is inefficient but consistent (OLS in this case under the maintained hypothesis of the SEM), while the other is efficient (SEM in this case). The null hypothesis is that the SEM and OLS estimates are not significantly different (see LeSage and Pace, 2010, pag. 62). We reject the null hypothesis, thus the SEM model is more appropriate. Table 6.2 compares the estimates.

6.4 Estimation of SAC Model: The Feasible Generalized Two Stage Least Squares estimator Procedure

6.4.1 Intuition Behind the Procedure


Consider the following SAC model:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}\mathbf{y} + \mathbf{u} = \mathbf{Z}\boldsymbol{\delta} + \mathbf{u} \\ \mathbf{u} &= \lambda\mathbf{M}\mathbf{u} + \boldsymbol{\varepsilon} \end{aligned} \tag{6.56}$$

Table 6.2: Spatial Models for Crime in Columbus: ML vs GM

	ML	GM
<i>Constant</i>	61.054*** (5.315)	63.487*** (5.084)
INC	-0.995** (0.337)	-1.180*** (0.342)
HOVAL	-0.308*** (0.093)	-0.300** (0.097)
λ	0.521*** (0.141)	0.364 (0.432)
N	49	49
Significance: *** $\equiv p < 0.001$; ** $\equiv p < 0.01$; * $\equiv p < 0.05$		

where $\mathbf{Z} = [\mathbf{X}, \mathbf{W}\mathbf{y}]$, $\boldsymbol{\delta} = [\boldsymbol{\beta}^\top, \lambda]^\top$, \mathbf{y} is the $n \times 1$ vector of observations of the dependent variables, \mathbf{X} is the $n \times k$ matrix of observations on **nonstochastic (exogenous)** regressors, \mathbf{W} and \mathbf{M} are the $n \times n$ nonstochastic weights matrices, \mathbf{u} is the $n \times 1$ vector of regression disturbances, $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of innovations. Note that we allow different spatial weight matrices for each process. However, in practice, there is a seldom sound basis for assuming this.

 This model is generally referred to as the Spatial-ARAR(1, 1) model to emphasize its autoregressive structure both in the dependent variable and the error term.

The SAC model can be estimated by ML procedure (see [Anselin, 1988](#)). However, the estimation process requires the inversion of \mathbf{A} and \mathbf{B} , which can be very costly in terms of computation in large samples. Furthermore, the ML relies on the normality assumption of the error terms. One way of dealing with this issue is to incorporate the estimation ideas from the S2SLS and GM we previously presented. To see this, we can re-write the first equation in model (6.56) by applying the following spatial Cochrane-Orcutt transformation:

$$\begin{aligned}
 \mathbf{y} &= \mathbf{Z}\boldsymbol{\delta} + (\mathbf{I} - \lambda\mathbf{M})^{-1}\boldsymbol{\varepsilon} \\
 (\mathbf{I} - \lambda\mathbf{M})\mathbf{y} &= (\mathbf{I} - \lambda\mathbf{M})\mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\varepsilon} \\
 \mathbf{y}_s(\lambda) &= \mathbf{Z}_s(\lambda)\boldsymbol{\delta} + \boldsymbol{\varepsilon}
 \end{aligned} \tag{6.57}$$

where the spatially filtered variables are given by:

$$\begin{aligned}
 \mathbf{y}_s(\lambda) &= \mathbf{y} - \lambda\mathbf{M}\mathbf{y} \\
 &= \mathbf{y} - \lambda\mathbf{y}_L \\
 &= (\mathbf{I} - \lambda\mathbf{M})\mathbf{y} \\
 \mathbf{Z}_s(\lambda) &= \mathbf{Z} - \lambda\mathbf{M}\mathbf{Z} \\
 &= \mathbf{Z} - \lambda\mathbf{Z}_L \\
 &= (\mathbf{I} - \lambda\mathbf{M})\mathbf{Z}
 \end{aligned}$$

If we knew λ , we would be able to apply an **IV approach on the transformed model** (6.57). For the discussion below, assume that we know λ . Note that the ideal instruments in this case will be:

$$\begin{aligned}\mathbb{E}(\mathbf{Z}) &= \mathbb{E}[\mathbf{X}, \mathbf{W}\mathbb{E}(\mathbf{y})] \\ \mathbb{E}(\mathbf{MZ}) &= \mathbb{E}[\mathbf{MX}, \mathbf{MW}\mathbb{E}(\mathbf{y})]\end{aligned}$$

Given that all the columns of $\mathbb{E}(\mathbf{Z})$ and $\mathbb{E}(\mathbf{MZ})$ are linear in

$$\mathbf{X}, \mathbf{WX}, \mathbf{W}^2\mathbf{X}, \dots, \mathbf{MX}, \mathbf{MWX}, \mathbf{MW}^2\mathbf{X}, \dots \quad (6.58)$$

the matrix of instruments \mathbf{H} is a subset of the linearly independent columns in (6.58), for example

$$\mathbf{H} = [\mathbf{X}, \mathbf{WX}, \dots, \mathbf{W}^l\mathbf{X}, \mathbf{MX}, \mathbf{MWX}, \dots, \mathbf{MW}^l\mathbf{X}]_{LI},$$

where typically, $l \leq 2$.

Since we have the instruments \mathbf{H} , and we have assumed that we have $\hat{\lambda}$ such that $\hat{\lambda} \xrightarrow{p} \lambda_0$ we might apply a GMM-type procedure using the following moment conditions for the transformed model (6.57):

$$\mathbf{m}(\lambda_0, \boldsymbol{\delta}_0) = \mathbb{E} \left[\frac{1}{\sqrt{n}} \mathbf{H}^\top \boldsymbol{\varepsilon} \right] = 0$$

Now let $\tilde{\lambda}$ some consistent estimator for λ_0 which can be obtained in a previous step, then the sample moment vector is:

$$\mathbf{m}^\delta(\tilde{\lambda}, \boldsymbol{\delta}) = \frac{1}{\sqrt{n}} \mathbf{H}^\top \underbrace{[\mathbf{y}_s(\tilde{\lambda}) - \mathbf{Z}_s(\tilde{\lambda})\boldsymbol{\delta}]}_{\tilde{\boldsymbol{\varepsilon}}},$$

where we explicitly state that the moments depends on $\boldsymbol{\delta}$ —which will be estimated—and a consistent estimate of λ . Under **homoskedasticity** the variance-covariance matrix of the moment vector $\mathbf{g}(\lambda_0, \boldsymbol{\delta}_0)$ is given by:

$$\mathbb{V}(\mathbf{m}(\lambda_0, \boldsymbol{\delta}_0)) = \mathbb{E}(\mathbf{m}(\lambda_0, \boldsymbol{\delta}_0)\mathbf{m}(\lambda_0, \boldsymbol{\delta}_0)^\top) = \sigma^2 n^{-1} \mathbf{H}^\top \mathbf{H},$$

which motivates the following two-step GMM estimator for $\boldsymbol{\delta}_0$:

$$\hat{\boldsymbol{\delta}} = \underset{\boldsymbol{\delta}}{\operatorname{argmin}} \quad \mathbf{g}_n^\delta(\tilde{\lambda}, \boldsymbol{\delta})^\top \boldsymbol{\mathcal{R}}_n^{\delta\delta} \mathbf{g}_n^\delta(\tilde{\lambda}, \boldsymbol{\delta})$$

with

$$\boldsymbol{\mathcal{R}}_n^{\delta\delta} = \left[\frac{1}{n} \mathbf{H}^\top \mathbf{H} \right]^{-1}.$$

Note that:

$$\begin{aligned}
J_n &= \left[\frac{1}{\sqrt{n}} \mathbf{H}^\top [\mathbf{y}_s(\tilde{\lambda}) - \mathbf{Z}_s(\tilde{\lambda})\boldsymbol{\delta}] \right]^\top \left[\frac{1}{n} \mathbf{H}^\top \mathbf{H} \right]^{-1} \left[\frac{1}{\sqrt{n}} \mathbf{H}^\top [\mathbf{y}_s(\tilde{\lambda}) - \mathbf{Z}_s(\tilde{\lambda})\boldsymbol{\delta}] \right] \\
&= \frac{1}{n} [\mathbf{y}_s(\tilde{\lambda}) - \mathbf{Z}_s(\tilde{\lambda})\boldsymbol{\delta}]^\top \mathbf{H} \left[\frac{1}{n} \mathbf{H}^\top \mathbf{H} \right]^{-1} \mathbf{H}^\top [\mathbf{y}_s(\tilde{\lambda}) - \mathbf{Z}_s(\tilde{\lambda})\boldsymbol{\delta}] \\
&= [\mathbf{y}_s(\tilde{\lambda}) - \mathbf{Z}_s(\tilde{\lambda})\boldsymbol{\delta}]^\top \mathbf{H} [\mathbf{H}^\top \mathbf{H}]^{-1} \mathbf{H}^\top [\mathbf{y}_s(\tilde{\lambda}) - \mathbf{Z}_s(\tilde{\lambda})\boldsymbol{\delta}] \\
&= [\mathbf{y}_s(\tilde{\lambda}) - \mathbf{Z}_s(\tilde{\lambda})\boldsymbol{\delta}]^\top \mathbf{P}_H [\mathbf{y}_s(\tilde{\lambda}) - \mathbf{Z}_s(\tilde{\lambda})\boldsymbol{\delta}]
\end{aligned}$$

Then, the estimator of $\boldsymbol{\delta}$ will be:

$$\hat{\boldsymbol{\delta}} = \left[\widehat{\mathbf{Z}}_s^\top \mathbf{Z}_s \right]^{-1} \widehat{\mathbf{Z}}_s^\top \mathbf{y}_s$$

where $\widehat{\mathbf{Z}}_s = \mathbf{H} (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H} \mathbf{Z}_s$. This estimator has been called the feasible generalized spatial two-stage least squares (FGS2SLS) estimator (Kelejian and Prucha, 1998). However, this estimator is not fully efficient.

The question is: How to obtain a consistent estimator of $\hat{\lambda}$? As probably you can guess, this consistent estimator is obtained in a previous step by GM.

6.4.2 Moment Conditions Revised

Since we will require a consistent estimate of λ , in this section we will specialized in other ways of expressing the moment conditions under homokedasticity (Kelejian and Prucha, 1999) and heteroskedasticity (Kelejian and Prucha, 2010). Furthermore, recall that the Kelejian and Prucha (1999)'s GM approach presented in Section 6.3.2 does not yield a consistent estimate for λ in the presence of heteroskedasticity: Theorem 6.14 is derived under homokedasticity. Extensions that include the form of a generalized method of moments were made by Kelejian and Prucha (2010), Arraiz et al. (2010) and Drukker et al. (2013).

The GMM approach offers three main extensions relative to GM. First, the estimator is robust to the presence of heteroskedasticity. Second, an asymptotic variance matrix is obtained for the parameter λ . Finally, joint inference is implemented for the spatial lag coefficient ρ and the spatial error coefficient λ .

The expression for the moment conditions in the articles cited above change a bit. In particular, the moment conditions are reduced from three to two and their expressions are generalized. This is so since no condition can be now derived from the parameter σ^2 under heteroskedasticity. To see this, consider the **homokedastic** model and the following three moment conditions:

$$\begin{aligned}
\mathbb{E} [\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}] &= \sigma^2 \\
\mathbb{E} [\boldsymbol{\varepsilon}^\top \mathbf{M} \mathbf{M} \boldsymbol{\varepsilon}] &= \sigma^2 \text{tr} (\mathbf{M}^\top \mathbf{M}) \\
\mathbb{E} [\boldsymbol{\varepsilon}^\top \mathbf{M} \boldsymbol{\varepsilon}] &= 0
\end{aligned}$$

Substituting out σ^2 into the second moment equation yields:

$$\begin{aligned}
\mathbb{E} [\boldsymbol{\varepsilon}^\top \mathbf{M} \mathbf{M} \boldsymbol{\varepsilon}] - \mathbb{E} [\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}] \operatorname{tr} (\mathbf{M}^\top \mathbf{M}) &= 0 \\
\mathbb{E} [\boldsymbol{\varepsilon}^\top \mathbf{M} \mathbf{M} \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} \operatorname{tr} (\mathbf{M}^\top \mathbf{M})] &= 0 \\
\mathbb{E} [\boldsymbol{\varepsilon}^\top \mathbf{M} \mathbf{M} \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}^\top \operatorname{tr} (\mathbf{M}^\top \mathbf{M}) \boldsymbol{\varepsilon}] &= 0 \\
\mathbb{E} [\boldsymbol{\varepsilon}^\top (\mathbf{M} \mathbf{M} - \operatorname{tr} (\mathbf{M}^\top \mathbf{M}) \mathbf{I}) \boldsymbol{\varepsilon}] &= 0 \\
\mathbb{E} [\boldsymbol{\varepsilon}^\top \mathbf{A}_1 \boldsymbol{\varepsilon}] &= 0.
\end{aligned}$$

Generalizing this expression for the third moment we end up with two instead of three quadratic moment conditions:

$$\begin{aligned}
\frac{1}{n} \mathbb{E} [\boldsymbol{\varepsilon}^\top \mathbf{A}_1 \boldsymbol{\varepsilon}] &= \mathbf{0} \\
\frac{1}{n} \mathbb{E} [\boldsymbol{\varepsilon}^\top \mathbf{A}_2 \boldsymbol{\varepsilon}] &= \mathbf{0}
\end{aligned} \tag{6.59}$$

with

$$\begin{aligned}
\mathbf{A}_1 &= \mathbf{M} \mathbf{M} - n^{-1} \operatorname{tr} (\mathbf{M}^\top \mathbf{M}) \mathbf{I} \\
\mathbf{A}_2 &= \mathbf{M}.
\end{aligned}$$

Note that \mathbf{A}_1 is symmetric with $\operatorname{tr}(\mathbf{A}_1) = 0$ (you should be able to prove this), but its diagonal elements are non zero (In the heteroskedasticity case it is!). In [Drukker et al. \(2013\)](#), an additional scaling factor is included as:

$$\nu = 1 / \left[1 + \left[(1/n) \operatorname{tr} (\mathbf{M}^\top \mathbf{M}) \right]^2 \right].$$

Under this case the weighting matrices for quadratic moments are:

$$\begin{aligned}
\mathbf{A}_1 &= \nu [\mathbf{M} \mathbf{M} - n^{-1} \operatorname{tr} (\mathbf{M}^\top \mathbf{M}) \mathbf{I}] \\
\mathbf{A}_2 &= \mathbf{M}_n.
\end{aligned}$$

If the errors are **heterokedastic**, then:

$$\begin{aligned}
\mathbf{A}_1 &= \mathbf{M}^\top \mathbf{M} - n^{-1} \operatorname{diag} (\mathbf{M}^\top \mathbf{M}) = \mathbf{M}^\top \mathbf{M} - n^{-1} \operatorname{diag} (\mathbf{m}_i^\top \mathbf{m}_i) \\
\mathbf{A}_2 &= \mathbf{M},
\end{aligned}$$

where \mathbf{m}_i is the i th column of the weights matrix \mathbf{M} . Note that $\operatorname{diag} (\mathbf{m}_i^\top \mathbf{m}_i)$ consists of the sum of the squares of the weight in the i th column. Denote this matrix as \mathbf{D} .

The sample moments are obtained by replacing $\boldsymbol{\varepsilon}$ by their counterpart expressed as a function of the regression residuals. Since $\mathbf{u} = \lambda \mathbf{u}_L + \boldsymbol{\varepsilon}$, it follows that $\boldsymbol{\varepsilon} = \mathbf{u} - \lambda \mathbf{u}_L = \mathbf{u}_s$, the spatially filtered residuals. Then:

$$\begin{aligned}
\frac{1}{n} \mathbb{E} [\mathbf{u}_s^\top \mathbf{A}_1 \mathbf{u}_s] &= \mathbf{0} \\
\frac{1}{n} \mathbb{E} [\mathbf{u}_s^\top \mathbf{A}_2 \mathbf{u}_s] &= \mathbf{0}
\end{aligned} \tag{6.60}$$

or more general

$$\frac{1}{n} \mathbb{E} \left[\mathbf{u}^\top (\mathbf{I} - \lambda \mathbf{M}^\top) \mathbf{A}_q (\mathbf{I} - \lambda \mathbf{M}^\top) \mathbf{u} \right] = 0 \quad (6.61)$$

where $q = 1, 2$. Note that:

$$\begin{aligned} \frac{1}{n} \boldsymbol{\varepsilon}^\top \mathbf{A}_q \boldsymbol{\varepsilon} &= \frac{1}{n} (\mathbf{u} - \lambda \mathbf{u}_L)^\top \mathbf{A}_q (\mathbf{u} - \lambda \mathbf{u}_L) \\ &= \frac{1}{n} \mathbf{u}^\top \mathbf{A}_q \mathbf{u} - \frac{1}{n} \lambda (\mathbf{u}^\top \mathbf{A}_q \mathbf{u}_L + \mathbf{u}_L^\top \mathbf{A}_q \mathbf{u}) + \frac{1}{n} \lambda^2 \mathbf{u}_L^\top \mathbf{A}_q \mathbf{u}_L \\ &= \frac{1}{n} \mathbf{u}^\top \mathbf{A}_q \mathbf{u} - 2 \frac{1}{n} \lambda \mathbf{u}_L^\top \mathbf{A}_q \mathbf{u} + \frac{1}{n} \lambda^2 \mathbf{u}_L^\top \mathbf{A}_q \mathbf{u}_L \\ &= \mathbf{0} \end{aligned} \quad (6.62)$$

In the third line of Equation 6.62, we assume that \mathbf{A}_q is symmetric such that:

$$\begin{aligned} \mathbf{u}^\top \mathbf{A}_q \mathbf{u}_L + \mathbf{u}_L^\top \mathbf{A}_q \mathbf{u} &= \mathbf{u}_L^\top \mathbf{A}_q^\top \mathbf{u} + \mathbf{u}_L \mathbf{A}_q \mathbf{u} \\ &= \mathbf{u}_L^\top (\mathbf{A}_q + \mathbf{A}_q^\top) \mathbf{u} \\ &= 2 \mathbf{u}_L^\top \mathbf{A}_q \mathbf{u} \end{aligned}$$

Here it is important to note that in some cases $\mathbf{A}_2 = \mathbf{M}$ might not be symmetric. However, we can use Definition 3.11.1 and set:

$$\mathbf{A}_2 = (1/2) (\mathbf{M} + \mathbf{M}^\top) \quad (6.63)$$

Taking expectation over (6.62):

$$\begin{aligned} \frac{1}{n} \mathbb{E} (\boldsymbol{\varepsilon}^\top \mathbf{A}_q \boldsymbol{\varepsilon}) &= n^{-1} \mathbb{E} (\mathbf{u}^\top \mathbf{A}_q \mathbf{u}) - 2n^{-1} \lambda \mathbb{E} (\mathbf{u}_L^\top \mathbf{A}_q \mathbf{u}) + \lambda^2 n^{-1} \mathbb{E} (\mathbf{u}_L^\top \mathbf{A}_1 \mathbf{u}_L) \\ \mathbf{0} &= n^{-1} \mathbb{E} (\mathbf{u}^\top \mathbf{A}_q \mathbf{u}) - \left(2n^{-1} \mathbb{E} (\mathbf{u}_L^\top \mathbf{A}_q \mathbf{u}) \quad -n^{-1} \mathbb{E} (\mathbf{u}_L^\top \mathbf{A}_q \mathbf{u}_L) \right) \begin{pmatrix} \lambda \\ \lambda^2 \end{pmatrix} \end{aligned}$$

Then, we have the following system of equations for $q = 1, 2$ (see (Kelejian and Prucha, 2010, pag 56)):

$$\begin{aligned} \begin{pmatrix} n^{-1} \mathbb{E} (\mathbf{u}^\top \mathbf{A}_1 \mathbf{u}) \\ n^{-1} \mathbb{E} (\mathbf{u}^\top \mathbf{A}_2 \mathbf{u}) \end{pmatrix} - \begin{pmatrix} 2n^{-1} \mathbb{E} (\mathbf{u}_L^\top \mathbf{A}_1 \mathbf{u}) & -n^{-1} \mathbb{E} (\mathbf{u}_L^\top \mathbf{A}_1 \mathbf{u}_L) \\ 2n^{-1} \mathbb{E} (\mathbf{u}_L^\top \mathbf{A}_2 \mathbf{u}) & -n^{-1} \mathbb{E} (\mathbf{u}_L^\top \mathbf{A}_2 \mathbf{u}_L) \end{pmatrix} \begin{pmatrix} \lambda \\ \lambda^2 \end{pmatrix} &= \mathbf{0} \\ \begin{pmatrix} n^{-1} \mathbb{E} (\mathbf{u}^\top \mathbf{A}_1 \mathbf{u}) \\ n^{-1} \mathbb{E} (\mathbf{u}^\top \mathbf{u}_L) \end{pmatrix} - \begin{pmatrix} 2n^{-1} \mathbb{E} (\mathbf{u}^\top \mathbf{M}^\top \mathbf{A}_1 \mathbf{u}) & -n^{-1} \mathbb{E} (\mathbf{u}^\top \mathbf{M}^\top \mathbf{A}_1 \mathbf{M} \mathbf{u}) \\ n^{-1} \mathbb{E} (\mathbf{u}_L^\top (\mathbf{M} + \mathbf{M}^\top) \mathbf{u}) & -n^{-1} \mathbb{E} (\mathbf{u}^\top \mathbf{M}^\top \mathbf{A}_2 \mathbf{M} \mathbf{u}) \end{pmatrix} &= \mathbf{0} \\ \boldsymbol{\gamma}_n - \boldsymbol{\Gamma}_n \boldsymbol{\alpha}_n &= \mathbf{0}. \end{aligned} \quad (6.64)$$

where we use Equation (6.63) for the second moment. Now, we can express the **sample moment conditions** as in Section 6.3.2:

$$\widetilde{\mathbf{m}}_{2 \times 1} = \widetilde{\mathbf{g}}_{2 \times 1} - \widetilde{\mathbf{G}}_{2 \times 2} \begin{pmatrix} \lambda \\ \lambda^2 \end{pmatrix} = \mathbf{0}$$

The elements of $\widetilde{\mathbf{g}}$ the following:

$$\begin{aligned}\tilde{\mathbf{g}}_1 &= \frac{1}{n} \tilde{\mathbf{u}}^\top \mathbf{A}_1 \tilde{\mathbf{u}} \\ \tilde{\mathbf{g}}_2 &= \frac{1}{n} \tilde{\mathbf{u}}^\top \mathbf{A}_2 \tilde{\mathbf{u}} = \frac{1}{n} \tilde{\mathbf{u}}^\top \tilde{\mathbf{u}}_L\end{aligned}$$

The $\widehat{\mathbf{G}}$ matrix is given by:

$$\widetilde{\mathbf{G}}_{11} = 2n^{-1} \tilde{\mathbf{u}}^\top \mathbf{M}^\top \mathbf{A}_1 \tilde{\mathbf{u}} \quad (6.65)$$

$$\widetilde{\mathbf{G}}_{12} = -n^{-1} \tilde{\mathbf{u}}^\top \mathbf{M}^\top \mathbf{A}_1 \mathbf{M} \tilde{\mathbf{u}} \quad (6.66)$$

$$\widetilde{\mathbf{G}}_{21} = -n^{-1} \tilde{\mathbf{u}}^\top \mathbf{M}^\top (\mathbf{A}_2 + \mathbf{A}_2^\top) \tilde{\mathbf{u}} \quad (6.67)$$

$$\widetilde{\mathbf{G}}_{22} = -n^{-1} \tilde{\mathbf{u}}^\top \mathbf{M} \mathbf{A}_2 \mathbf{M} \tilde{\mathbf{u}} \quad (6.68)$$

A more compact notation is:

$$\begin{aligned}\widetilde{\mathbf{G}} &= \frac{1}{n} \begin{pmatrix} \tilde{\mathbf{u}}^\top (\mathbf{A}_1 + \mathbf{A}_1^\top) \tilde{\mathbf{u}}_s & -\tilde{\mathbf{u}}_s^\top \mathbf{A}_1 \tilde{\mathbf{u}}_s^\top \\ \vdots & \vdots \\ \tilde{\mathbf{u}}^\top (\mathbf{A}_q + \mathbf{A}_q^\top) \tilde{\mathbf{u}}_s & -\tilde{\mathbf{u}}_s^\top \mathbf{A}_q \tilde{\mathbf{u}}_s^\top \end{pmatrix} \\ \tilde{\mathbf{g}} &= \frac{1}{n} \begin{pmatrix} \tilde{\mathbf{u}}^\top \mathbf{A}_1 \tilde{\mathbf{u}} \\ \vdots \\ \tilde{\mathbf{u}}^\top \mathbf{A}_q \tilde{\mathbf{u}} \end{pmatrix}\end{aligned}$$

for $q = 1, 2$.

Now, let Ψ be 2×2 matrix of variance-covariance matrix of the moment conditions $\frac{1}{n} \mathbb{E} [\varepsilon^\top \mathbf{A}_1 \varepsilon]$. Then, the using Equation (3.34) from Lemma 3.25:

$$\psi_{s,r} = \frac{1}{2n} \text{tr} \left[(\mathbf{A}_s + \mathbf{A}_s^\top) \Sigma (\mathbf{A}_r + \mathbf{A}_r^\top) \Sigma \right] + \frac{1}{n} \mu^\top (\mathbf{A}_1 + \mathbf{A}_1^\top) \Sigma (\mathbf{A}_2 + \mathbf{A}_2^\top) \mu$$

where $s, r = 1, 2$ correspond to the moment conditions; Σ is a diagonal matrix in the heteroskedasticity case with elements:

$$\hat{\varepsilon}_i^2 = (\tilde{u}_i - \lambda \tilde{u}_{L_i})^2 = \tilde{u}_{s_i}^2$$

R Kelejian and Prucha (1999) show consistency of the Method of Moment estimator of λ , but not asymptotic normality of the estimator.

6.4.3 Assumptions

Now we will state the assumption for the SAC model under heteroskedasticity following Arraiz et al. (2010). The assumptions regarding the spatial weight matrix are the following:

Assumption 6.18 — Spatial Weight Matrices (Arraiz et al., 2010). Assume the following:

- (a) All diagonal elements \mathbf{W}_n and \mathbf{M}_n are zero.
- (b) $\lambda \in (-1, 1)$, $\rho \in (-1, 1)$.
- (c) The matrices $\mathbf{I}_n - \rho\mathbf{W}_n$ and $\mathbf{I}_n - \lambda\mathbf{M}_n$ are nonsingular for all $\lambda \in (-1, 1)$ and $\rho \in (-1, 1)$.

Assumption 6.18(a) is a normalization rule: a region cannot be a neighbor of itself. Assumption 6.18(b) has to do with the parameter space. This assumption is discussed by Kelejian and Prucha (2010, section 2.2). Assumption 6.18(c) ensures that \mathbf{y} and \mathbf{u} are uniquely defined. Thus, under assumption 6.18 (Spatial Weight Matrices), we can write the model as:

$$\begin{aligned}\mathbf{y}_n &= (\mathbf{I}_n - \rho\mathbf{W}_n)^{-1} [\mathbf{X}_n\boldsymbol{\beta} + \mathbf{u}_n] \\ \mathbf{u}_n &= (\mathbf{I}_n - \rho\mathbf{M}_n)^{-1} \boldsymbol{\varepsilon}_n.\end{aligned}$$

The reduced form is:

$$\mathbf{y} = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I} - \rho\mathbf{W})^{-1}(\mathbf{I} - \lambda\mathbf{M})^{-1}\boldsymbol{\varepsilon}$$

The reduced form represents a system of n simultaneous equations. As in the standard spatial lag model, we can include endogenous explanatory variables on the right hand side of model specification. In this case:

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Y}\boldsymbol{\gamma} + (\mathbf{I} - \lambda\mathbf{W})^{-1}\boldsymbol{\varepsilon}.$$

Assumption 6.19 — Heteroskedastic Errors (Arraiz et al., 2010). The error term $\{\epsilon_{i,n} : 1 \leq i \leq n, n \geq 1\}$ satisfy $\mathbb{E}(\epsilon_{i,n}) = 0$, $\mathbb{E}(\epsilon_{i,n}^2) = \sigma_{i,n}^2$, with $0 < \underline{a}^\sigma \leq \sigma_{i,n}^2 \leq \bar{a}^\sigma < \infty$. Furthermore, for each $n \geq 1$ the random variables $\epsilon_{1,n}, \dots, \epsilon_{n,n}$ are totally independent.

Assumption 6.19 allows the innovations to be heteroskedastic with uniformly bounded variances. This assumption also allows for the innovations to depend on the sample size n , i.e., to form a triangular arrays.

Assumption 6.20 — Bounded Spatial Weight Matrices (Arraiz et al., 2010). The row and column sums of the matrices \mathbf{W}_n and \mathbf{M}_n are bounded uniformly in absolute value, by , respectively, one and some finite constant, and the row and column sums of the matrices $(\mathbf{I}_n - \rho\mathbf{W}_n)^{-1}$ and $(\mathbf{I} - \rho\mathbf{M}_n)^{-1}$ are bounded uniformly in absolute value by some finite constant.

This assumption is a technical assumption, which is used in large-sample derivation of the regression parameter estimator. This assumption limits the extent of spatial autocorrelation between \mathbf{u} and \mathbf{y} . It ensures that the disturbance process and the process of the dependent variable exhibit a “fading” memory. Note that:

$$\begin{aligned}\mathbb{E}[\mathbf{u}_n] &= \mathbb{E}[(\mathbf{I}_n - \lambda\mathbf{M}_n)^{-1} \boldsymbol{\varepsilon}_n] \\ &= (\mathbf{I}_n - \lambda\mathbf{M}_n)^{-1} \mathbb{E}[\boldsymbol{\varepsilon}_n] \\ &= \mathbf{0} \quad \text{by Assumption 6.19 (Heteroskedastic Errors)}\end{aligned}\tag{6.69}$$

$$\begin{aligned}
\mathbb{E} [\mathbf{u}_n \mathbf{u}_n^\top] &= \mathbb{E} \left[(\mathbf{I}_n - \lambda \mathbf{M}_n)^{-1} \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^\top (\mathbf{I}_n - \lambda \mathbf{M}_n^\top)^{-1} \right] \\
&= (\mathbf{I}_n - \lambda \mathbf{M}_n)^{-1} \mathbb{E} [\boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^\top] (\mathbf{I}_n - \lambda \mathbf{M}_n^\top)^{-1} \\
&= (\mathbf{I}_n - \lambda \mathbf{M}_n)^{-1} \boldsymbol{\Sigma} (\mathbf{I}_n - \lambda \mathbf{M}_n^\top)^{-1}
\end{aligned} \tag{6.70}$$

where $\boldsymbol{\Sigma} = \text{diag}(\sigma_{i,n}^2)$.

Assumption 6.21 — Regressors (Arraiz et al., 2010). The regressor matrices \mathbf{X}_n have full column rank (for n large enough). Furthermore, the elements of the matrices \mathbf{X}_n are uniformly bounded in absolute value.

This assumption rules out multicollinearity problems, as well as unbounded exogenous variables.

Assumption 6.22 — Instruments I (Arraiz et al., 2010). The instruments matrices \mathbf{H}_n have full column rank $L \geq K + 1$ (for all n large enough). Furthermore, the elements of the matrices \mathbf{H}_n are uniformly bounded in absolute value. Additionally, \mathbf{H}_n is assumed to, at least, contain the linearly independent columns of $(\mathbf{X}_n, \mathbf{M}_n \mathbf{X}_n)$

There are some papers that discuss the use of optimal instruments for the spatial (see for example Lee, 2003; Das et al., 2003; Kelejian et al., 2004; Lee, 2007).



The effect of the selection of instruments on the efficiency of the estimators remains to be further investigated.

Assumption 6.23 — Instruments II (Identification) (Arraiz et al., 2010). The instruments \mathbf{H}_n satisfy furthermore:

- (a) $\mathbf{Q}_{HH} = \lim_{n \rightarrow \infty} n^{-1} \mathbf{H}_n^\top \mathbf{H}_n$ is finite and nonsingular.
- (b) $\mathbf{Q}_{HZ} = \text{plim}_{n \rightarrow \infty} n^{-1} \mathbf{H}_n^\top \mathbf{Z}_n$ and $\mathbf{Q}_{HMZ} = \text{plim}_{n \rightarrow \infty} n^{-1} \mathbf{H}_n^\top \mathbf{M} \mathbf{Z}_n$ are finite and have full column rank. Furthermore $\mathbf{Q}_{HZ,s}(\lambda) = \mathbf{Q}_{HZ} - \lambda \mathbf{Q}_{HMZ}$ has full column rank.
- (c) $\mathbf{Q}_{H\Sigma H} = \lim_{n \rightarrow \infty} n^{-1} \mathbf{H}_n^\top \boldsymbol{\Sigma}_n \mathbf{H}_n$ is finite and nonsingular, where $\boldsymbol{\Sigma}_n = \text{diag}(\sigma_{i,n}^2)$

In treating \mathbf{X}_n and \mathbf{H}_n as non-stochastic our analysis should be viewed as conditional on \mathbf{X}_n and \mathbf{H}_n .

6.4.4 Estimators and Estimation Procedure in a Nutshell

Consider again the transformed model:

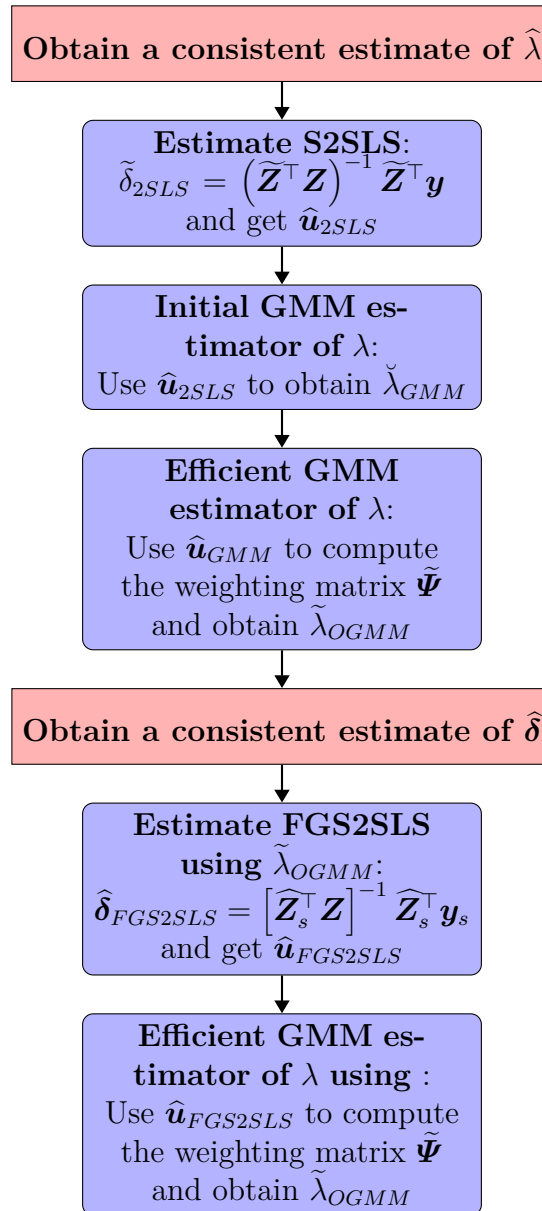
$$\mathbf{y}_s(\lambda_0) = \mathbf{Z}_s(\lambda_0) \boldsymbol{\delta}_0 + \boldsymbol{\epsilon}$$

where $\mathbf{y}_s(\lambda_0) = \mathbf{y} - \lambda_0 \mathbf{M} \mathbf{y}$ and $\mathbf{Z}_s(\lambda_0) = \mathbf{Z} - \lambda_0 \mathbf{M} \mathbf{Z}$. If we would know λ_0 , then we could apply the S2SLS to the transformed model. However, λ_0 is unknown and therefore we need to estimate it in a first place in order to estimate $\boldsymbol{\delta}$. The steps will be:

- (a) An initial IV estimator of δ leads to a set of consistent residuals.
- (b) With these residuals, derive the moment conditions that provide a consistent estimate of λ_0 using GMM Estimation procedure.
- (c) The estimate of λ_0 is then used to define a **weighting matrix** for the moment conditions in order to obtain a consistent and efficient estimator.
- (d) An estimate of δ_0 is obtained from the **transformed model**.
- (e) Finally, a **consistent and efficient** estimate of λ is based on GS2SLS residuals.

These steps are shown in Figure 6.1.

Figure 6.1: Estimation steps for SAC model



Now we will consider each step in detail:

Step 1a: 2SLS estimator

In the first step, δ is estimated by 2SLS applied to **untransformed model** $\mathbf{y} = \mathbf{Z}\delta + \mathbf{u}$ using the instruments matrix \mathbf{H} . Then:

$$\tilde{\delta}_{2SLS} = \left(\widetilde{\mathbf{Z}}^\top \mathbf{Z} \right)^{-1} \widetilde{\mathbf{Z}}^\top \mathbf{y} \quad (6.71)$$

where $\widetilde{\mathbf{Z}} = \mathbf{H} \left(\mathbf{H}^\top \mathbf{H} \right)^{-1} \mathbf{H}^\top \mathbf{Z} = \mathbf{P}_H \mathbf{Z} = (\mathbf{X}, \widetilde{\mathbf{W}}\mathbf{y})$. The estimates $\tilde{\delta}_{2SLS}$ yield an initial vector of residuals, \mathbf{u}_{2SLS} as:

$$\tilde{\mathbf{u}}_{2SLS} = \mathbf{y} - \mathbf{Z}\tilde{\delta}_{2SLS} \quad (6.72)$$

The following Theorem states that $\tilde{\delta}_{2SLS}$ is consistent:

Theorem 6.24 — Consistency of $\tilde{\delta}_{2SLS}$ (Kelejian and Prucha, 2010). Suppose the assumptions hold. Then $\tilde{\delta}_{2SLS} = \delta + O_p(n^{-1/2})$, and hence $\tilde{\delta}_{2SLS}$ is consistent for δ_0 :

$$\tilde{\delta}_{2SLS} \xrightarrow{p} \delta_0$$

Sketch of proof for Theorem 6.24. The model is:

$$\begin{aligned} \mathbf{y}_n &= \mathbf{Z}_n \delta + \mathbf{u}_n, \\ \mathbf{u}_n &= \lambda \mathbf{M}_n \mathbf{u}_n + \boldsymbol{\varepsilon}_n. \end{aligned}$$

The sampling error is given by:

$$\begin{aligned} \hat{\delta}_n &= \delta_0 + \left(\widehat{\mathbf{Z}}_n^\top \widehat{\mathbf{Z}}_n \right)^{-1} \widehat{\mathbf{Z}}_n^\top \mathbf{u}_n \\ &= \delta_0 + \left[\left(\mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{Z}_n \right)^\top \left(\mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{Z}_n \right) \right]^{-1} \left(\mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{Z}_n \right)^\top \mathbf{u}_n \\ &= \delta_0 + \left[\mathbf{Z}_n^\top \mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{Z}_n \right]^{-1} \mathbf{Z}_n^\top \mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top (\mathbf{I} - \lambda \mathbf{M}_n)^{-1} \boldsymbol{\varepsilon}_n \end{aligned}$$

Solving for $\hat{\delta}_n - \delta_0$ and multiplying by \sqrt{n} we obtain:

$$\begin{aligned} \sqrt{n}(\hat{\delta}_n - \delta_0) &= \left[\left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right)^\top \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right) \right]^{-1} \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right)^\top \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \frac{1}{\sqrt{n}} \mathbf{H}_n^\top (\mathbf{I} - \lambda \mathbf{M}_n)^{-1} \boldsymbol{\varepsilon}_n \\ &= \left[\left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right)^\top \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right) \right]^{-1} \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right)^\top \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \frac{1}{\sqrt{n}} \mathbf{F}_n^\top \boldsymbol{\varepsilon}_n, \end{aligned}$$

where:

$$\mathbf{F}_n^\top = \mathbf{H}_n^\top (\mathbf{I} - \lambda \mathbf{M}_n)^{-1} = \text{whose elements are bounded in absolute value}$$

Assumption 6.23 implies that:

$$\begin{aligned} \lim \frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n &= \mathbf{Q}_{HH}, \\ \text{plim} \frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n &= \mathbf{Q}_{HZ}, \end{aligned}$$

which are finite and nonsingular.

Furthermore, note that $\mathbb{E}(n^{-1/2} \mathbf{F}_n^\top \boldsymbol{\varepsilon}_n) = \mathbf{0}$ and

$$\begin{aligned} \mathbb{E} \left[(n^{-1/2} \mathbf{F}_n^\top \boldsymbol{\varepsilon}_n) (n^{-1/2} \mathbf{F}_n^\top \boldsymbol{\varepsilon}_n)^\top \right] &= \frac{1}{n} \mathbb{E} \left[\mathbf{H}_n^\top (\mathbf{I} - \lambda \mathbf{M}_n)^{-1} \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top (\mathbf{I} - \lambda \mathbf{M}_n^\top)^{-1} \mathbf{H}_n \right] \\ &= \sigma^2 \frac{1}{n} \mathbf{H}_n^\top (\mathbf{I} - \lambda \mathbf{M}_n)^{-1} (\mathbf{I} - \lambda \mathbf{M}_n^\top)^{-1} \mathbf{H}_n \end{aligned}$$

Assume that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{H}_n^\top (\mathbf{I} - \lambda \mathbf{M}_n)^{-1} (\mathbf{I} - \lambda \mathbf{M}_n^\top)^{-1} \mathbf{H}_n = \frac{1}{n} \mathbf{F}_n^\top \mathbf{F}_n = \boldsymbol{\Phi} \quad \text{exists}$$

Then assuming homocedasticity and using Theorem 6.16:

$$n^{-1/2} \mathbf{F}_n^\top \boldsymbol{\varepsilon}_n \xrightarrow{d} \mathbf{N}(\mathbf{0}, \sigma_\varepsilon^2 \boldsymbol{\Phi})$$

Therefore:

$$\sqrt{n}(\hat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0) \xrightarrow{d} \mathbf{N}(\mathbf{0}, \boldsymbol{\Delta})$$

and

$$\boldsymbol{\Delta} = \sigma_\varepsilon^2 \left[\mathbf{Q}_{HZ}^\top \mathbf{Q}_{HH}^{-1} \mathbf{Q}_{HZ} \right]^{-1} \mathbf{Q}_{HZ}^\top \mathbf{Q}_{HH}^{-1} \boldsymbol{\Phi} \mathbf{Q}_{HH}^{-1} \mathbf{Q}_{HZ} \left[\mathbf{Q}_{HZ}^\top \mathbf{Q}_{HH}^{-1} \mathbf{Q}_{HZ} \right]^{-1}$$

Then we can say that $\tilde{\boldsymbol{\delta}} = \boldsymbol{\delta} + O_p(n^{-1/2})$.

Consistency follows if $n^{-1} \mathbf{F}_n^\top \boldsymbol{\varepsilon}_n \xrightarrow{p} \mathbf{0}$. Note that $\mathbb{E}(n^{-1} \mathbf{F}_n^\top \boldsymbol{\varepsilon}_n) = \mathbf{0}$ and

$$\mathbb{V} \left(n^{-1} \mathbf{F}_n^\top \boldsymbol{\varepsilon}_n \right) = \sigma^2 \frac{1}{n^2} \mathbf{H}_n^\top (\mathbf{I} - \lambda \mathbf{M}_n)^{-1} (\mathbf{I} - \lambda \mathbf{M}_n^\top)^{-1} \mathbf{H}_n$$

which converges to $\mathbf{0}$, then using Chebyshev's Theorem 3.5:

$$n^{-1} \mathbf{F}_n^\top \boldsymbol{\varepsilon}_n \xrightarrow{p} \mathbf{0} \quad \text{and hence} \quad \tilde{\boldsymbol{\delta}}_n \xrightarrow{p} \boldsymbol{\delta}_0$$

■

Although $\tilde{\boldsymbol{\delta}}_{2SLS}$ is consistent, it does not utilize information relating to the spatial correlation error term. We therefore turn to the second step of the procedure. (Question: Why we cannot use the OLS residuals for the next step?)

Step 1b: Initial GMM estimator of λ based on 2SLS residuals

Using the consistent estimate \mathbf{u} in the previous step, now we create the sample moments corresponding to (6.61) for $q = 1, 2$ based on the estimated residuals, and $\tilde{\mathbf{u}}_s = \mathbf{M} \tilde{\mathbf{u}}$:

$$\begin{aligned} \mathbf{m}(\lambda, \tilde{\boldsymbol{\delta}}_{2SLS}) &= \frac{1}{n} \begin{pmatrix} \tilde{\mathbf{u}}_{2SLS}^\top (\mathbf{I} - \lambda \mathbf{M}^\top) \mathbf{A}_1 (\mathbf{I} - \lambda \mathbf{M}) \tilde{\mathbf{u}}_{2SLS} \\ \tilde{\mathbf{u}}_{2SLS}^\top (\mathbf{I} - \lambda \mathbf{M}^\top) \mathbf{A}_2 (\mathbf{I} - \lambda \mathbf{M}) \tilde{\mathbf{u}}_{2SLS} \end{pmatrix} \\ &= \tilde{\mathbf{G}} \begin{pmatrix} \lambda \\ \lambda^2 \end{pmatrix} - \tilde{\mathbf{g}} \end{aligned} \tag{6.73}$$

where,

$$\begin{aligned}\tilde{\mathbf{G}} &= \frac{1}{n} \begin{pmatrix} \tilde{\mathbf{u}}^\top (\mathbf{A}_1 + \mathbf{A}_1^\top) \tilde{\mathbf{u}}_s & -\tilde{\mathbf{u}}_s^\top \mathbf{A}_1 \tilde{\mathbf{u}}_s^\top \\ \vdots & \vdots \\ \tilde{\mathbf{u}}^\top (\mathbf{A}_q + \mathbf{A}_q^\top) \tilde{\mathbf{u}}_s & -\tilde{\mathbf{u}}_s^\top \mathbf{A}_q \tilde{\mathbf{u}}_s^\top \end{pmatrix} \\ \tilde{\mathbf{g}} &= \frac{1}{n} \begin{pmatrix} \tilde{\mathbf{u}}^\top \mathbf{A}_1 \tilde{\mathbf{u}} \\ \vdots \\ \tilde{\mathbf{u}}^\top \mathbf{A}_q \tilde{\mathbf{u}} \end{pmatrix}\end{aligned}$$

The initial GMM estimator for λ is then defined as

$$\check{\lambda}_{gmm} = \underset{\lambda}{\operatorname{argmin}} \left\{ \left[\tilde{\mathbf{G}} \begin{pmatrix} \lambda \\ \lambda^2 \end{pmatrix} - \tilde{\mathbf{g}} \right]^\top \left[\tilde{\mathbf{G}} \begin{pmatrix} \lambda \\ \lambda^2 \end{pmatrix} - \tilde{\mathbf{g}} \right] \right\} \quad (6.74)$$

where $\mathbf{\Upsilon}^{\lambda\lambda} = \mathbf{I}$. This estimator is consistent but not efficient. For efficiency we need to replace $\mathbf{\Upsilon}^{\lambda\lambda}$ by the variance-covariance matrix of the sample moments. Furthermore, the expression above can be interpreted as a nonlinear least squares system of equations. The initial estimate is thus obtained as a solution of the above system.

Now, we need to define the expression for the matrices \mathbf{A}_s . [Drukker et al. \(2013\)](#) suggest, for the homokedastic case, the following expressions:

$$\begin{aligned}\mathbf{A}_1 &= v \left[\mathbf{M}^\top \mathbf{M} - \frac{1}{n} \operatorname{tr}(\mathbf{M}^\top \mathbf{M}) \mathbf{I} \right] \\ \mathbf{A}_2 &= \mathbf{M}\end{aligned}$$

where v is the scaling factor needed to obtain the same estimator of [Kelejian and Prucha \(1998, 1999\)](#).

On the other hand, when heteroskedasticity is assumed, [Kelejian and Prucha \(2010\)](#) recommend the following expressions:

$$\begin{aligned}\mathbf{A}_1 &= \mathbf{M}^\top \mathbf{M} - \operatorname{diag}(\mathbf{M}^\top \mathbf{M}) \\ \mathbf{A}_2 &= \mathbf{M}\end{aligned}$$

Step 1c: Efficient GMM estimator of λ based on 2SLS residuals

The efficient GMM estimator of λ is a weighted nonlinear least squares estimator. Specifically, this estimator is $\tilde{\lambda}$ where:

$$\tilde{\lambda}_{ogmm} = \underset{\lambda}{\operatorname{argmin}} \left[\mathbf{m}(\lambda, \tilde{\boldsymbol{\delta}})^\top \tilde{\boldsymbol{\Psi}}^{-1} \mathbf{m}(\lambda, \tilde{\boldsymbol{\delta}}) \right] \quad (6.75)$$

and where the weighting matrix is $\tilde{\boldsymbol{\Psi}}_n^{-1}$, where $\boldsymbol{\Psi}$ is the variance of the moment conditions $\mathbf{m}(\lambda, \tilde{\boldsymbol{\delta}})$.

The matrix $\tilde{\boldsymbol{\Psi}}_n^{-1} = \tilde{\boldsymbol{\Psi}}_n^{-1}(\check{\lambda}_{gmm})$ is defined as follows. Let $\tilde{\boldsymbol{\Psi}} = [\hat{\Psi}_{rs}]_{r,s=1,2}$ with

$$\tilde{\Psi}_{rs} = (2n)^{-1} \operatorname{tr} \left[(\mathbf{A}_r + \mathbf{A}_r^\top) \tilde{\boldsymbol{\Sigma}} (\mathbf{A}_s + \mathbf{A}_s^\top) \tilde{\boldsymbol{\Sigma}} \right] + n^{-1} \tilde{\mathbf{a}}_r^\top \tilde{\boldsymbol{\Sigma}} \tilde{\mathbf{a}}_s, \quad (6.76)$$

where:

$$\begin{aligned}
\widetilde{\Sigma} &= \text{diag}_{i=1,\dots,n}(\tilde{\epsilon}_i^2) \\
\tilde{\epsilon} &= (\mathbf{I} - \check{\lambda}_{gmm}\mathbf{M})\tilde{\mathbf{u}} \\
\tilde{\mathbf{a}}_r &= (\mathbf{I} - \check{\lambda}_{gmm}\mathbf{M})\mathbf{H}\tilde{\mathbf{P}}\tilde{\alpha}_r \\
\tilde{\alpha}_r &= -n^{-1} \left[\mathbf{Z}^\top (\mathbf{I} - \check{\lambda}_{gmm}\mathbf{M}) (\mathbf{A}_r + \mathbf{A}_r^\top) (\mathbf{I} - \check{\lambda}_{gmm}\mathbf{M}) \tilde{\mathbf{u}} \right] \\
\tilde{\mathbf{P}} &= \left(\frac{1}{n} \mathbf{H}^\top \mathbf{H} \right)^{-1} \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right) \left[\left(\frac{1}{n} \mathbf{H}^\top \mathbf{Z} \right)^\top \left(\frac{1}{n} \mathbf{H}^\top \mathbf{H} \right)^{-1} \left(\frac{1}{n} \mathbf{H}^\top \mathbf{Z} \right) \right]^{-1}
\end{aligned} \tag{6.77}$$

It is important to note that this step is not necessary since the previous estimator of λ is already consistent.

Step 2a: FGS2SLS Estimator

Using $\check{\lambda}_{ogmm}$ from step 1c (or the consistent estimator from step 1b) in the transformed model we have:

$$\hat{\delta}_n(\check{\lambda}_{ogmm}) = \left[\widehat{\mathbf{Z}}_s^\top (\check{\lambda}_{ogmm}) \mathbf{Z} (\check{\lambda}_{ogmm}) \right]^{-1} \widehat{\mathbf{Z}}_s^\top (\check{\lambda}_{ogmm}) \mathbf{y}_s(\check{\lambda}_{ogmm}) \tag{6.78}$$

where

$$\begin{aligned}
\mathbf{y}_s &= \mathbf{y} - \check{\lambda}_{ogmm} \mathbf{M} \mathbf{y} \\
\mathbf{Z}_s &= \mathbf{Z} - \check{\lambda}_{ogmm} \mathbf{M} \mathbf{Z} \\
\widehat{\mathbf{Z}}_s &= \mathbf{P}_H \mathbf{Z}_s \\
\mathbf{P}_H &= \mathbf{H} (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top
\end{aligned} \tag{6.79}$$

Step 2b: Efficient GMM estimator of λ using FGS2SLS residual

In this last step, and **efficient** GMM estimator of λ based on the GS2SLS residuals is obtained by minimizing the following expression:

$$\hat{\lambda} = \underset{\lambda}{\operatorname{argmin}} \left\{ \left[\widehat{\mathbf{G}} \begin{pmatrix} \lambda \\ \lambda^2 \end{pmatrix} - \hat{\mathbf{g}} \right]^\top (\widehat{\Psi}^{\lambda\lambda})^{-1} \left[\widehat{\mathbf{G}} \begin{pmatrix} \lambda \\ \lambda^2 \end{pmatrix} - \hat{\mathbf{g}} \right] \right\} \tag{6.80}$$

where $\widehat{\Psi}^{\lambda\lambda}$ is an estimator for the variance-covariance matrix of the (normalized) sample moment vector based on the GS2SLS residuals. This estimator differs for the cases of homoskedastic and heteroskedastic errors.

For the **homoskedastic** case the r, s (with $r, s = 1, 2$) element of $\widehat{\Psi}^{\lambda\lambda}$ is given by:

$$\begin{aligned}
\widehat{\Psi}_{rs}^{\lambda\lambda} &= [\tilde{\sigma}^2]^2 (2n)^{-1} \operatorname{tr} \left[(\mathbf{A}_r + \mathbf{A}_r^\top) (\mathbf{A}_s + \mathbf{A}_s^\top) \right] \\
&\quad + \tilde{\sigma}^2 n^{-1} \tilde{\mathbf{a}}_r^\top \tilde{\mathbf{a}}_s^\top \\
&\quad + n^{-1} \left(\tilde{\mu}^{(4)} - 3 [\tilde{\sigma}^2]^2 \right) \operatorname{vec}_D(\mathbf{A}_r)^\top \operatorname{vec}_D(\mathbf{A}_s) \\
&\quad + n^{-1} \tilde{\mu}^{(3)} \left[\tilde{\mathbf{a}}_r^\top \operatorname{vec}_D(\mathbf{A}_s) + \tilde{\mathbf{a}}_s^\top \operatorname{vec}_D(\mathbf{A}_r) \right],
\end{aligned} \tag{6.81}$$

where

$$\begin{aligned}
\tilde{\mathbf{a}}_r &= \hat{\mathbf{T}} \tilde{\alpha}_r \\
\hat{\mathbf{T}} &= \mathbf{H} \hat{\mathbf{P}}, \\
\hat{\mathbf{P}} &= \hat{\mathbf{Q}}_{HH}^{-1} \hat{\mathbf{Q}}_{HZ} \left[\hat{\mathbf{Q}}_{HZ}^\top \hat{\mathbf{Q}}_{HH}^{-1} \hat{\mathbf{Q}}_{HZ} \right]^{-1} \\
\hat{\mathbf{Q}}_{HH}^{-1} &= \left(n^{-1} \mathbf{H}^\top \mathbf{H} \right), \\
\hat{\mathbf{Q}}_{HZ} &= \left(n^{-1} \mathbf{H}^\top \mathbf{Z} \right), \\
\mathbf{Z} &= \left(\mathbf{I} - \tilde{\lambda} \mathbf{M} \right) \mathbf{Z}, \\
\tilde{\alpha}_r &= -n^{-1} \left[\mathbf{Z}^\top \left(\mathbf{A}_r + \mathbf{A}_r^\top \right) \hat{\boldsymbol{\varepsilon}} \right] \\
\hat{\sigma}^2 &= n^{-1} \hat{\boldsymbol{\varepsilon}} \hat{\boldsymbol{\varepsilon}}^\top, \\
\hat{\mu}^{(3)} &= n^{-1} \sum_{i=1}^n \hat{\varepsilon}_i^3, \\
\hat{\mu}^{(4)} &= n^{-1} \sum_{i=1}^n \hat{\varepsilon}_i^4.
\end{aligned} \tag{6.82}$$

For the **heteroskedastic** case the r, s (with $r, s = 1, 2$) element of $\hat{\boldsymbol{\Psi}}^{\hat{\lambda}\hat{\lambda}}$ is given by:

$$\hat{\boldsymbol{\Psi}}_{rs}^{\hat{\lambda}\hat{\lambda}} = (2n)^{-1} \text{tr} \left[\left(\mathbf{A}_r + \mathbf{A}_r^\top \right) \hat{\boldsymbol{\Sigma}} \left(\mathbf{A}_s + \mathbf{A}_s^\top \right) \hat{\boldsymbol{\Sigma}} \right] + n^{-1} \tilde{\mathbf{a}}_r^\top \hat{\boldsymbol{\Sigma}} \tilde{\mathbf{a}}_s, \tag{6.83}$$

where, $\hat{\boldsymbol{\Sigma}}$ is a diagonal matrix whose i th diagonal element is $\hat{\varepsilon}_i^2$.

6.5 Application in R

In this example we will use the **simulated** US Driving Under the Influence (DUI) county data set used in [Drukker et al. \(2011\)](#). The dependent variable `dui` is defined as the alcohol-related arrest rate per 100,000 daily vehicle miles traveled (DVMT). The explanatory variables include

- `police`: number of sworn officers per 100,000 DVMT,
- `nondui`: non-alcohol-related arrests per 100,000 DVMT,
- `vehicles`: number of registered vehicles per 1,000 residents, and
- `dry`: a dummy for counties that prohibit alcohol sale within their borders

We load the required packages and dataset:

```

library("maptools")
library("spdep")
library("sphet")
# Load Data
us_shape <- readShapeSpatial("ccountyR") # Load shape file

```

```
## Warning: shapelib support is provided by GDAL through the sf and terra packages
among others
## Warning: shapelib support is provided by GDAL through the sf and terra packages
among others
## Warning: shapelib support is provided by GDAL through the sf and terra packages
among others

names(us_shape)                                # Names of variables in dbf

## [1] "dry"          "nondui"      "vehicles" "elect"      "dui"          "police"

# Load weight matrix
queen.w <- read.gal("ccountyR_w.gal")
lw <- nb2listw(queen.w, style = "W")
```

6.5.1 SAC Model with Homokedasticity (GS2SLS)

First, we estimate the SAC model assuming homoskedasticity ([Kelejian and Prucha, 1998](#)) using the `gstsls` function from **spdep** package. We will also assume that $\mathbf{W} = \mathbf{M}$. The code is the following:

```
GS2SLS <- gstsls(dui ~ police + nondui + vehicles + dry,
                 data = us_shape,
                 listw = lw)
summary(GS2SLS)

##
## Call:gstsls(formula = dui ~ police + nondui + vehicles + dry, data = us_shape,
##           listw = lw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.655535 -0.362165 -0.070363  0.277261  2.418849
##
## Type: GM SARAR estimator
## Coefficients: (GM standard errors)
##              Estimate Std. Error z value Pr(>|z|)
## Rho_Wy          0.04692763  0.01698220   2.7633  0.005721
## (Intercept) -6.40991922  0.41836312 -15.3214 < 2.2e-16
## police         0.59810726  0.01491778  40.0936 < 2.2e-16
## nondui         0.00024688  0.00108699   0.2271  0.820328
## vehicles       0.01571247  0.00066881  23.4933 < 2.2e-16
## dry            0.10608849  0.03496242   3.0344  0.002410
##
## Lambda: 0.00095701
## Residual variance (sigma squared): 0.31811, (sigma: 0.56402)
```

```
## GM argmin sigma squared: 0.31789
## Number of observations: 3109
## Number of parameters estimated: 8
```

The results show that all the variables are significant, except for `nondui`. Importantly, higher number of sworn officers is positively correlated with the DUI arrest rate, after controlling for `nondui`, `vehicles` and `dry!` The spatial autoregressive coefficient ρ is positive and significant indicating autocorrelation in the dependent variable. [Drukker et al. \(2011\)](#) give some theoretical explanation of this results. On the one hand, the positive coefficient may be explained in terms of coordination effort among police departments in different countries. On the other hand, it might well be that an enforcement effort in one of the counties leads people living close to the border to drink in neighboring counties. The estimate is λ negative, however the output does not produce inference for it. Lastly, it is important to stress that the standard errors has a degrees of freedom correction in the variance-covariance matrix.

6.5.2 SAC Model with Homokedasticity and Additional Endogeneity (GS2SLS)

The size of the `police` force may be related with the arrest rates `dui`. As a consequence, `police` produces endogeneity. We will use the dummy variable `elect`, where `elect` is 1 if a country government faces an election, 0 otherwise. To do so, we use the `spreg` function from `sphet`. Note that λ is ρ . The estimate of ρ is positive and significant thus indicating spatial autocorrelation in the dependent variable (coordination effort among police departments in different counties).

```
G2SLS_en_in <- spreg(dui ~ nondui + vehicles + dry,
  data = us_shape,
  listw = lw,
  endog = ~ police,
  instruments = ~ elect,
  model = "sarar",
  het = FALSE,
  lag.instr = TRUE)
summary(G2SLS_en_in)

##
## Generalized stsls
##
## Call:
## spreg(formula = dui ~ nondui + vehicles + dry, data = us_shape,
##       listw = lw, endog = ~police, instruments = ~elect, lag.instr = TRUE,
##       model = "sarar", het = FALSE)
##
## Residuals:
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -6.1862 -0.8838  0.0147 -0.0161  0.9213  8.3616
```

```
##
## Coefficients:
##              Estimate Std. Error t-value Pr(>|t|)
## (Intercept) 11.60596811  1.66674437  6.9633 3.325e-12 ***
## nondui      -0.00019624  0.00275912 -0.0711  0.943299
## vehicles     0.09299562  0.00564911 16.4620 < 2.2e-16 ***
## dry          0.39825983  0.09090201  4.3812 1.180e-05 ***
## police      -1.35130834  0.14101772 -9.5825 < 2.2e-16 ***
## lambda       0.19319018  0.04431011  4.3600 1.301e-05 ***
## rho         -0.08597523  0.03018333 -2.8484  0.004393 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Wald test that rho and lambda are both zero:
## Statistics: 7.6185 p-val: 0.0057773
```

An important issue here is that **the optimal instrument are unknown**. It is not recommended the inclusion of the spatial lag of these additional exogenous variables in the matrix of instruments. However, results reported in ? do consider the spatial lags of `elect`.

Now we assume that the error are heteroskedastic of unknown form.

6.6 Exercises

Exercise 6.1 Consider the following model:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \\ \mathbf{u} &= \lambda \mathbf{W}\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon} \end{aligned}$$

where $|\lambda| < 1$, $\boldsymbol{\varepsilon}$ has zero mean and variance $\sigma^2 \mathbf{I}_n$, respectively. Determine moment equations for a GMM approach you would use to estimate λ and σ^2 . (Hint: This model is known as the spatial moving average model for the error term).

Exercise 6.2 Consider the following model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \rho_1 \mathbf{W}_1 \mathbf{y} + \rho_2 \mathbf{W}_2 \mathbf{y} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\varepsilon}$ has zero mean and variance $\sigma^2 \mathbf{I}_n$, respectively, and \mathbf{W}_1 and \mathbf{W}_2 are observed exogenous weighting matrices. Suggest an instrumental variable estimation procedure for this model which accounts for the endogeneity of $\mathbf{W}_1 \mathbf{y}$ and $\mathbf{W}_2 \mathbf{y}$.

Exercise 6.3 Consider the following model:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \rho_1 \mathbf{W}_1 \mathbf{y} + \rho_2 \mathbf{W}_2 \mathbf{y} + \mathbf{u} \\ \mathbf{u} &= \lambda \mathbf{M}\mathbf{u} + \boldsymbol{\varepsilon} \end{aligned}$$

where $\boldsymbol{\varepsilon}$ has zero mean and variance $\sigma^2 \mathbf{I}_n$, respectively, and \mathbf{W}_1 , \mathbf{W}_2 and \mathbf{M} are observed exogenous weighting matrices. Suggest an instrumental variable estimation procedure for this model which accounts for the endogeneity of $\mathbf{W}_1 \mathbf{y}$ and $\mathbf{W}_2 \mathbf{y}$, as well as for the spatially correlated term.

Appendix

6.A Proof Theorem 3 in KP 1998

Recall that the GS2SLS is given by:

$$\hat{\delta}_n = \left[\widehat{\mathbf{Z}}_s(\lambda)^\top \widehat{\mathbf{Z}}_s(\lambda) \right]^{-1} \widehat{\mathbf{Z}}_s(\lambda)^\top \mathbf{y}_s(\lambda) \quad (6.84)$$

Whereas, the FGS2SLS is given by:

$$\hat{\delta}_{F,n} = \left[\widehat{\mathbf{Z}}_s(\hat{\lambda})^\top \widehat{\mathbf{Z}}_s(\hat{\lambda}) \right]^{-1} \widehat{\mathbf{Z}}_s(\hat{\lambda})^\top \mathbf{y}_s(\hat{\lambda}) \quad (6.85)$$

where

$$\begin{aligned} \widehat{\mathbf{Z}}_s(\hat{\lambda}_n) &= \mathbf{P}_{H_n} \mathbf{Z}_s(\hat{\lambda}_n) \\ \mathbf{Z}_s(\hat{\lambda}_n) &= \mathbf{Z}_n - \hat{\lambda}_n \mathbf{M}_n \mathbf{Z}_n \\ \mathbf{y}_s(\hat{\lambda}_n) &= \mathbf{y}_n - \hat{\lambda}_n \mathbf{M}_n \mathbf{y}_n \\ \widehat{\mathbf{Z}}_s(\hat{\lambda}_n) &= \left(\mathbf{X}_n - \hat{\lambda}_n \mathbf{M}_n \mathbf{X}_n, \mathbf{W}_n \mathbf{y}_n - \widehat{\widehat{\lambda}_n \mathbf{M}_n \mathbf{W}_n \mathbf{y}_n} \right) \\ \widehat{\widehat{\lambda}_n \mathbf{M}_n \mathbf{W}_n \mathbf{y}_n} &= \mathbf{P}_{H_n} \left(\mathbf{W}_n \mathbf{y}_n - \hat{\lambda}_n \mathbf{M}_n \mathbf{W}_n \mathbf{y}_n \right). \end{aligned} \quad (6.86)$$

The sampling error is:

$$\hat{\delta}_{F,n} - \delta = \left[\widehat{\mathbf{Z}}_s(\hat{\lambda})^\top \widehat{\mathbf{Z}}_s(\hat{\lambda}) \right]^{-1} \widehat{\mathbf{Z}}_s(\hat{\lambda})^\top \mathbf{u}_s(\hat{\lambda}_n) \quad (6.87)$$

where:

$$\begin{aligned} \mathbf{u}_s(\hat{\lambda}_n) &= (\mathbf{I} - \hat{\lambda}_n) \mathbf{u} \\ &= (\mathbf{I} - \hat{\lambda}_n) \mathbf{u} + \boldsymbol{\varepsilon}_n - \boldsymbol{\varepsilon}_n \\ &= \boldsymbol{\varepsilon}_n + (\mathbf{I} - \hat{\lambda}_n \mathbf{M}_n) \mathbf{u} - (\mathbf{I} - \lambda \mathbf{M}_n) \mathbf{u} \\ &= \boldsymbol{\varepsilon}_n + \mathbf{u} - \hat{\lambda}_n \mathbf{M}_n \mathbf{u} - \mathbf{u} + \lambda \mathbf{M}_n \mathbf{u} \\ &= \boldsymbol{\varepsilon}_n - (\hat{\lambda}_n - \lambda) \mathbf{M}_n \mathbf{u}_n \end{aligned} \quad (6.88)$$

Then:

$$\begin{aligned} \hat{\delta}_{F,n} - \delta &= \left[\widehat{\mathbf{Z}}_s(\hat{\lambda})^\top \widehat{\mathbf{Z}}_s(\hat{\lambda}) \right]^{-1} \widehat{\mathbf{Z}}_s(\hat{\lambda})^\top \left[\boldsymbol{\varepsilon}_n - (\hat{\lambda}_n - \lambda) \mathbf{M}_n \mathbf{u}_n \right] \\ &= \left[\widehat{\mathbf{Z}}_s(\hat{\lambda})^\top \widehat{\mathbf{Z}}_s(\hat{\lambda}) \right]^{-1} \widehat{\mathbf{Z}}_s(\hat{\lambda})^\top \boldsymbol{\varepsilon}_n - \left[\widehat{\mathbf{Z}}_s(\hat{\lambda})^\top \widehat{\mathbf{Z}}_s(\hat{\lambda}) \right]^{-1} \widehat{\mathbf{Z}}_s(\hat{\lambda})^\top (\hat{\lambda}_n - \lambda) \mathbf{M}_n \mathbf{u}_n \\ &= \left[\frac{1}{n} \widehat{\mathbf{Z}}_s(\hat{\lambda})^\top \widehat{\mathbf{Z}}_s(\hat{\lambda}) \right]^{-1} \frac{1}{n} \widehat{\mathbf{Z}}_s(\hat{\lambda})^\top \boldsymbol{\varepsilon}_n - \left[\frac{1}{n} \widehat{\mathbf{Z}}_s(\hat{\lambda})^\top \widehat{\mathbf{Z}}_s(\hat{\lambda}) \right]^{-1} (\hat{\lambda}_n - \lambda) \frac{1}{n} \widehat{\mathbf{Z}}_s(\hat{\lambda})^\top \mathbf{M}_n \mathbf{u}_n \\ \sqrt{n}(\hat{\delta}_{F,n} - \delta) &= \left[\frac{1}{n} \widehat{\mathbf{Z}}_s(\hat{\lambda})^\top \widehat{\mathbf{Z}}_s(\hat{\lambda}) \right]^{-1} \frac{1}{\sqrt{n}} \widehat{\mathbf{Z}}_s(\hat{\lambda})^\top \boldsymbol{\varepsilon}_n - \left[\frac{1}{n} \widehat{\mathbf{Z}}_s(\hat{\lambda})^\top \widehat{\mathbf{Z}}_s(\hat{\lambda}) \right]^{-1} (\hat{\lambda}_n - \lambda) \frac{1}{\sqrt{n}} \widehat{\mathbf{Z}}_s(\hat{\lambda})^\top \mathbf{M}_n \mathbf{u}_n \end{aligned} \quad (6.89)$$

By consistency $\hat{\lambda}_n - \lambda = o_p(1)$. Now, we need to show that:

$$\frac{1}{n} \widehat{\mathbf{Z}}_s(\hat{\lambda})^\top \widehat{\mathbf{Z}}_s(\hat{\lambda}) \xrightarrow{p} \frac{1}{n} \widehat{\mathbf{Z}}_s(\lambda)^\top \widehat{\mathbf{Z}}_s(\lambda) = \bar{\mathbf{Q}} \quad (6.90)$$

$$\frac{1}{\sqrt{n}} \widehat{\mathbf{Z}}_s(\widehat{\lambda})^\top \boldsymbol{\varepsilon}_n \xrightarrow{d} \mathbf{N}(\mathbf{0}, \sigma_\epsilon^2 \bar{\mathbf{Q}}), \quad (6.91)$$

$$(\widehat{\lambda}_n - \lambda) \frac{1}{\sqrt{n}} \widehat{\mathbf{Z}}_s(\widehat{\lambda})^\top \mathbf{M}_n \mathbf{u}_n \xrightarrow{p} 0 \quad (6.92)$$

where:

$$\bar{\mathbf{Q}} = [\mathbf{Q}_{HZ} - \lambda \mathbf{Q}_{mHZ}]^\top \mathbf{Q}_{HH}^{-1} [\mathbf{Q}_{HZ} - \lambda \mathbf{Q}_{mHZ}] \quad (6.93)$$

is finite and nonsingular. For 6.90, note that:

$$\begin{aligned} \frac{1}{n} \widehat{\mathbf{Z}}_s(\widehat{\lambda})^\top \widehat{\mathbf{Z}}_s(\widehat{\lambda}) &= \frac{1}{n} (\mathbf{Z}_n - \widehat{\lambda}_n \mathbf{M}_n \mathbf{Z}_n)^\top \mathbf{P}_{H_n} (\mathbf{Z}_n - \widehat{\lambda}_n \mathbf{M}_n \mathbf{Z}_n) \\ &= \frac{1}{n} (\mathbf{Z}_n - \widehat{\lambda}_n \mathbf{M}_n \mathbf{Z}_n)^\top \mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H})^{-1} \mathbf{H}_n^\top (\mathbf{Z}_n - \widehat{\lambda}_n \mathbf{M}_n \mathbf{Z}_n) \\ &= \left(\underbrace{\frac{1}{n} \mathbf{Z}_n^\top \mathbf{H}_n}_{\xrightarrow{p} \mathbf{Q}_{HZ}^\top} - \underbrace{\widehat{\lambda}_n}_{\xrightarrow{p} \lambda} \underbrace{\frac{1}{n} \mathbf{Z}_n^\top \mathbf{M}_n^\top \mathbf{H}_n}_{\xrightarrow{p} \mathbf{Q}_{mHZ}^\top} \right) \underbrace{\left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H} \right)^{-1}}_{\rightarrow \mathbf{Q}_{HH}^{-1}} \underbrace{\left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n - \widehat{\lambda}_n \frac{1}{n} \mathbf{H}_n^\top \mathbf{M}_n \mathbf{Z}_n \right)}_{\xrightarrow{p} \mathbf{Q}_{HZ} - \lambda \mathbf{Q}_{mHZ}} \end{aligned} \quad (6.94)$$

For 6.91, note that:

$$\begin{aligned} \frac{1}{\sqrt{n}} \widehat{\mathbf{Z}}_s(\widehat{\lambda})^\top \boldsymbol{\varepsilon}_n &= \frac{1}{\sqrt{n}} (\mathbf{Z}_n - \widehat{\lambda}_n \mathbf{M}_n \mathbf{Z}_n)^\top \mathbf{P}_{H_n} \boldsymbol{\varepsilon} \\ &= \left(\underbrace{\frac{1}{n} \mathbf{Z}_n^\top \mathbf{H}_n}_{\xrightarrow{p} \mathbf{Q}_{HZ}^\top} - \underbrace{\widehat{\lambda}_n}_{\xrightarrow{p} \lambda} \underbrace{\frac{1}{n} \mathbf{Z}_n^\top \mathbf{M}_n^\top \mathbf{H}_n}_{\xrightarrow{p} \mathbf{Q}_{mHZ}^\top} \right) \underbrace{\left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H} \right)^{-1}}_{\rightarrow \mathbf{Q}_{HH}^{-1}} \underbrace{\frac{1}{\sqrt{n}} \mathbf{H}_n^\top \boldsymbol{\varepsilon}_n}_{\xrightarrow{d} \mathbf{N}(\mathbf{0}, \sigma_\epsilon^2 \mathbf{Q}_{HH})} \end{aligned} \quad (6.95)$$

For 6.92 note that:

$$(\widehat{\lambda}_n - \lambda) \frac{1}{\sqrt{n}} \widehat{\mathbf{Z}}_s(\widehat{\lambda})^\top \mathbf{M}_n \mathbf{u}_n = \underbrace{(\widehat{\lambda}_n - \lambda)}_{o_p(1)} \left(\underbrace{\frac{1}{n} \mathbf{Z}_n^\top \mathbf{H}_n}_{\xrightarrow{p} \mathbf{Q}_{HZ}^\top} - \underbrace{\widehat{\lambda}_n}_{\xrightarrow{p} \lambda} \underbrace{\frac{1}{n} \mathbf{Z}_n^\top \mathbf{M}_n^\top \mathbf{H}_n}_{\xrightarrow{p} \mathbf{Q}_{mHZ}^\top} \right) \underbrace{\left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H} \right)^{-1}}_{\rightarrow \mathbf{Q}_{HH}^{-1}} \frac{1}{\sqrt{n}} \mathbf{H}_n^\top \mathbf{M}_n \mathbf{u}_n \quad (6.96)$$

Note that $\mathbb{E} \left(n^{-1/2} \mathbf{H}_n^\top \mathbf{M}_n \mathbf{u}_n \right) = 0$ and $\mathbb{E} \left(n^{-1} \mathbf{H}_n^\top \mathbf{M}_n \mathbf{u}_n \mathbf{u}_n^\top \mathbf{M}_n^\top \mathbf{H}_n^\top \right) = n^{-1} \mathbf{H}_n^\top \mathbf{M}_n \boldsymbol{\Sigma}_{u_n} \mathbf{M}_n^\top \mathbf{H}_n^\top$, whose elements are bounded, where

$$\boldsymbol{\Sigma}_{u_n} = \sigma_\epsilon^2 (\mathbf{I} - \lambda \mathbf{M}_n)^{-1} (\mathbf{I} - \lambda \mathbf{M}_n^\top)^{-1}$$

Then $\frac{1}{\sqrt{n}} \mathbf{H}_n^\top \mathbf{M}_n \mathbf{u}_n = O_p(1)$ and finally

$$\sqrt{n}(\widehat{\boldsymbol{\delta}}_{F,n} - \boldsymbol{\delta}) \xrightarrow{d} \mathbf{N}(\mathbf{0}, \sigma_\epsilon^2 \bar{\mathbf{Q}}^{-1}) \quad (6.97)$$

The small sample approximation is

$$\hat{\boldsymbol{\delta}}_{F,n} \sim \text{N} \left(\boldsymbol{\delta}, \hat{\sigma}^2 \left[\widehat{\mathbf{Z}}_s(\hat{\lambda})^\top \widehat{\mathbf{Z}}_s(\hat{\lambda}) \right]^{-1} \right) \quad (6.98)$$

where:

$$\hat{\sigma}^2 = \hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}} / n \quad (6.99)$$

and $\hat{\boldsymbol{\varepsilon}} = \mathbf{y}_s(\hat{\lambda}) - \mathbf{Z}_s(\hat{\lambda})\hat{\boldsymbol{\delta}}_F$.

Bibliography

- Allers, M. A. and Elhorst, J. P. (2005). Tax Mimicking and Yardstick Competition Among Local Governments in The Netherlands. *International tax and public finance*, 12(4):493–513.
- Anselin, L. (1988). *Spatial Econometrics: Methods and Models*, volume 4. Springer.
- Anselin, L. (1996). Chapter Eight: The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association. *Spatial Analytical*, 4:121.
- Anselin, L. (2003). Spatial Externalities, Spatial Multipliers, and Spatial Econometrics. *International regional science review*, 26(2):153–166.
- Anselin, L. (2007). *Spatial Econometrics*, pages 310–330. Blackwell Publishing Ltd.
- Anselin, L. and Bera, A. (1998). Spatial Dependence in Linear Regression Models with an Introduction to Spatial Econometrics. In Ullah, A. and Giles, D., editors, *Handbook of Applied Economic Statistics*, pages 237–289. Marcel Dekker, New York.
- Anselin, L. and Lozano-Gracia, N. (2008). Errors in Variables and Spatial Effects in Hedonic House Price Models of Ambient Air Quality. *Empirical economics*, 34(1):5–34.
- Anselin, L. and Rey, S. (1991). Properties of Tests for Spatial Dependence in Linear Regression Models. *Geographical analysis*, 23(2):112–131.
- Anselin, L. and Rey, S. (2014). *Modern Spatial Econometrics in Practice: A Guide to Geoda, Geodaspace and Pysal*. GeoDa Press LLC.
- Arraiz, I., Drukker, D. M., Kelejian, H. H., and Prucha, I. R. (2010). A Spatial Cliff-Ord-Type Model with Heteroskedastic Innovations: Small and Large Sample Results. *Journal of Regional Science*, 50(2):592–614.
- Baller, R. D., Anselin, L., Messner, S. F., Deane, G., and Hawkins, D. F. (2001). Structural Covariates of US County Homicide Rates: Incorporating Spatial Effects. *Criminology*, 39(3):561–588.

- Basdas, U. (2009). Spatial Econometric Analysis of the Determinants of Location in Turkish Manufacturing Industry. *Available at SSRN 1506888*.
- Bivand, R., Hauke, J., and Kossowski, T. (2013). Computing the Jacobian in Gaussian Spatial Autoregressive Models: An Illustrated Comparison of Available Methods. *Geographical Analysis*, 45(2):150–179.
- Bivand, R. and Lewin-Koh, N. (2015). *maptools: Tools for Reading and Handling Spatial Objects*. R package version 0.8-36.
- Bivand, R. and Piras, G. (2015). Comparing Implementations of Estimation Methods for Spatial Econometrics. *Journal of Statistical Software*, 63(1):1–36.
- Boarnet, M. G. and Glazer, A. (2002). Federal Grants and Yardstick Competition. *Journal of urban Economics*, 52(1):53–64.
- Cliff, A. and Ord, K. (1972). Testing for Spatial Autocorrelation Among Regression Residuals. *Geographical analysis*, 4(3):267–284.
- Cliff, A. D. and Ord, J. K. (1973). *Spatial Autocorrelation*. London:Pion.
- Cohen, J. and Tita, G. (1999). Diffusion in Homicide: Exploring a General Method for Detecting Spatial Diffusion Processes. *Journal of Quantitative Criminology*, 15(4):451–493.
- Cordy, C. B. and Griffith, D. A. (1993). Efficiency of least squares estimators in the presence of spatial autocorrelation. *Communications in Statistics-Simulation and Computation*, 22(4):1161–1179.
- Das, D., Kelejian, H. H., and Prucha, I. R. (2003). Finite Sample Properties of Estimators of Spatial Autoregressive Models with Autoregressive Disturbances. *Papers in Regional Science*, 82(1):1–26.
- Doreian, P. (1981). Estimating Linear Models with Spatially Distributed Data. *Sociological methodology*, pages 359–388.
- Drukker, D. M., Egger, P., and Prucha, I. R. (2013). On Two-step Estimation of a Spatial Autoregressive Model with Autoregressive Disturbances and Endogenous Regressors. *Econometric Reviews*, 32(5-6):686–733.
- Drukker, D. M., Prucha, I. R., and Raciborski, R. (2011). A Command for Estimating Spatial-autoregressive Models with Spatial-autoregressive Disturbances and Additional Endogenous Variables. *Econometric Reviews*, 32:686–733.
- Elhorst, J. P. (2010). Applied Spatial Econometrics: Raising the Bar. *Spatial Economic Analysis*, 5(1):9–28.
- Elhorst, J. P. (2014). *Spatial Econometrics: From Cross-Sectional Data to Spatial Panels*. Springer.
- Filiztekin, A. (2009). Regional Unemployment in Turkey. *Papers in Regional Science*, 88(4):863–878.

- Fischer, M. M., Bartkowska, M., Riedl, A., Sardadvar, S., and Kunnert, A. (2009). The Impact of Human Capital on Regional Labor Productivity in Europe. *Letters in Spatial and Resource Sciences*, 2(2-3):97–108.
- Garretsen, H. and Peeters, J. (2009). FDI and the Relevance of Spatial Linkages: Do Third-Country Effects Matter for Dutch FDI? *Review of World Economics*, 145(2):319–338.
- Garrett, T. A. and Marsh, T. L. (2002). The revenue impacts of cross-border lottery shopping in the presence of spatial autocorrelation. *Regional Science and Urban Economics*, 32(4):501–519.
- Gibbons, S., Overman, H. G., and Patacchini, E. (2015). Spatial Methods. *Handbook of Regional and Urban Economics SET*, page 115.
- Kelejian, H. and Piras, G. (2017). *Spatial econometrics*. Academic Press.
- Kelejian, H. H. and Prucha, I. R. (1998). A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances. *The Journal of Real Estate Finance and Economics*, 17(1):99–121.
- Kelejian, H. H. and Prucha, I. R. (1999). A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model. *International economic review*, 40(2):509–533.
- Kelejian, H. H. and Prucha, I. R. (2001). On the Asymptotic Distribution of the Moran I Test Statistic with Applications. *Journal of Econometrics*, 104(2):219–257.
- Kelejian, H. H. and Prucha, I. R. (2007). The Relative Efficiencies of Various Predictors in Spatial Econometric Models Containing Spatial Lags. *Regional Science and Urban Economics*, 37(3):363–374.
- Kelejian, H. H. and Prucha, I. R. (2010). Specification and Estimation of Spatial Autoregressive Models with Autoregressive and Heteroskedastic Disturbances. *Journal of Econometrics*, 157(1):53–67.
- Kelejian, H. H., Prucha, I. R., and Yuzefovich, Y. (2004). Instrumental Variable Estimation of a Spatial Autoregressive Model with Autoregressive Disturbances: Large and Small Sample Results. In Lesage, J. and Pace, R., editors, *Spatial and Spatiotemporal Econometrics*, pages 163–198. Emerald Group Publishing Limited.
- Kim, C. W., Phipps, T. T., and Anselin, L. (2003). Measuring the Benefits of Air Quality Improvement: A Spatial Hedonic Approach. *Journal of environmental economics and management*, 45(1):24–39.
- Kirby, D. K. and LeSage, J. P. (2009). Changes in Commuting to Work Times Over the 1990 to 2000 Period. *Regional Science and Urban Economics*, 39(4):460–471.
- Lee, L.-F. (2002). Consistency and Efficiency of Least Squares Estimation for Mixed Regressive, Spatial Autoregressive Models. *Econometric theory*, 18(02):252–277.
- Lee, L.-f. (2003). Best Spatial Two-Stage Least Squares Estimators for a Spatial Autoregressive Model with Autoregressive Disturbances. *Econometric Reviews*, 22(4):307–335.

- Lee, L.-F. (2004). Asymptotic Distributions of Quasi-Maximum Likelihood Estimators for Spatial Autoregressive Models. *Econometrica*, 72(6):1899–1925.
- Lee, L.-f. (2007). GMM and 2SLS Estimation of Mixed Regressive, Spatial Autoregressive Models. *Journal of Econometrics*, 137(2):489–514.
- LeSage, J. and Pace, R. K. (2010). *Introduction to Spatial Econometrics*. CRC press.
- LeSage, J. P. (1997). Bayesian Estimation of Spatial Autoregressive Models. *International Regional Science Review*, 20(1-2):113–129.
- LeSage, J. P. (2014). What Regional Scientists Need to Know about Spatial Econometrics. *The Review of Regional Studies*, 44(1):13–32.
- LeSage, J. P. and Pace, R. K. (2014). *Interpreting Spatial Econometric Models*, pages 1535–1552. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Liu, T., Xu, X., and Lee, L.-f. (2022). Consistency without compactness of the parameter space in spatial econometrics. *Economics Letters*, 210:110224.
- Mead, R. (1967). A Mathematical Model for the Estimation of Inter-Plant Competition. *Biometrics*, pages 189–205.
- Newey, W. K. and McFadden, D. (1994). Large Sample Estimation and Hypothesis Testing. *Handbook of econometrics*, 4:2111–2245.
- Ord, K. (1975). Estimation Methods for Models of Spatial Interaction. *Journal of the American Statistical Association*, 70(349):120–126.
- Pace, R. K. and LeSage, J. P. (2008). A spatial hausman test. *Economics Letters*, 101(3):282–284.
- Pavlyuk, D. (2011). Spatial Analysis of Regional Employment Rates in Latvia.
- Piras, G. (2010). sphet: Spatial Models with Heteroskedastic Innovations in R. *Journal of Statistical Software*, 35(1):1–21.
- Prucha, I. (2014). Instrumental Variables/Method of Moments Estimation. In Fischer, M. M. and Nijkamp, P., editors, *Handbook of Regional Science*, pages 1597–1617. Springer Berlin Heidelberg.
- Saavedra, L. A. (2000). A Model of Welfare Competition with Evidence from AFDC. *Journal of Urban Economics*, 47(2):248–279.
- Smirnov, O. and Anselin, L. (2001). Fast Maximum Likelihood Estimation of Very Large Spatial Autoregressive Models: A Characteristic Polynomial Approach. *Computational Statistics & Data Analysis*, 35(3):301–319.
- Stewart, B. M. and Zhukov, Y. (2010). Choosing Your Neighbors: The Sensitivity of Geographical Diffusion in International Relations. In *APSA 2010 Annual Meeting Paper*.
- Tiefelsdorf, M., Griffith, D., and Boots, B. (1999). A Variance-Stabilizing Coding Scheme for Spatial Link Matrices. *Environment and Planning A*, 31(1):165–180.

- big O, 68
- consistent estimator, 72
- Convergence
 - bounded sequences, 67
 - deterministic sequences, 65
- convergence in probability, 69
- eigen values, 37
- Endogeneity
 - additional endogenous variables, 187
 - error in variables, 187
- Generalized method of moments, 177
 - Moment conditions, 195
- GS2SLS
 - gstsls function, 220
 - spreg function, 221
- Heteroskedasticity
 - error term, 184
- Instrumental Variables
 - definition in the spatial context, 183
 - optimal instruments, 184
 - S2SLS, 182
- Leontief expansion, 38
- Maximum likelihood, 108
 - concentrated log-likelihood, 111
 - Jacobian, 109
- Moment conditions, 195
- Moran's I test, 24
 - Monte carlo, 28
 - Moran scatterplot, 25
 - moran.mc function, 32
 - moran.plot function, 32
 - moran.test function, 30
 - Normality, 27
 - Randomization, 28
- Multiplier effect, 39
- Parameter space, 38
- quadratic moment conditions, 195
- Reduced form
 - Spatial lag model, 36
- S2SLS
 - Asymptotic distribution, 189
 - consistency, 188
 - example, 190
 - stsls function, 190
- SAC model
 - FGS2SLS, 205
- Spatial autocorrelation, 4
- Spatial autoregressive process, 36
- Spatial dependence, 4
- Spatial durbin model, 40
 - reduced form, 40
- Spatial error model, 41
 - reduced form, 41
- Spatial lag model, 35
- Spillover effects, 36
 - Direct effects, 47
 - example, 53
 - for S2SLS, 192

- Global spillovers, 45
- Indirect effects, 47
- Local spillovers, 46
- Marginal effects, 46

Tobler's law, 3

Weight matrix, 6

- Based on distance, 9
- Bishop contiguity, 8
- Definition, 7
- Higher order, 13
- Invertibility, 37
- knearneigh function, 20
- lag.listw function, 24
- nb2listw function, 17
- poly2nb function, 16
- Queen contiguity, 9
- Rook contiguity, 7
- Row-standardization, 11
- Spatial lag, 12