

Intermediate Spatial Econometrics for Cross-Sectional Data with Applications in R: Lecture Notes and Codes

Mauricio Sarrias
Universidad de Talca

June 13, 2025

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 Weights	12
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	28
1.5.1	Cloropeth Graphs	28
1.5.2	Moran's I Test	29
1.6	Exercises	32
2	Spatial Models	35
2.1	Taxonomy of Models	35
2.1.1	Spatial Lag Model	35
2.1.2	Spatial Durbin Model	36
2.1.3	Spatial Error Model	37
2.1.4	Spatial Autocorrelation Model	38
2.2	Reduced Form and Parameter Space	40
2.2.1	Eigenvalues in R	44
2.2.2	Generate spatial DGP in R	47

2.3	Motivation of Spatial Models	47
2.3.1	SLM as a Long-run Equilibrium	47
2.3.2	SEM and Omitted Variables Motivation	48
2.3.3	SDM and Omitted Variables Motivation	49
2.4	Interpreting Spatial Models	49
2.4.1	Measuring Spillovers	49
2.4.2	Marginal Effects	50
2.4.3	Partitioning Global Effects Estimates Over Space	56
2.5	Lesage's Book Example	57
2.5.1	Commuting Times and Congestion	57
2.5.2	Computing Effects in R	59
2.5.3	Cumulative Effects	63
2.6	Empirical Example: Growth and Spatial Externalities	66
2.7	Exercises	68
II	Estimation Methods	71
3	Review of Asymptotic Theory	73
3.1	Convergence of Deterministic Sequences	73
3.2	Convergence in Probability	77
3.2.1	Convergence in Quadratic Mean	81
3.3	Law of Large Numbers	83
3.4	Convergence in Distribution	90
3.5	Central Limit Theorems	93
3.6	Orders in Probability	97
3.7	Triangular Arrays	100
3.8	Bounded Matrices and Useful Lemmas for Spatial Econometrics	103
3.9	Linear and Quadratic Forms	105
3.9.1	Moments	105
3.9.2	Law of Large Numbers	109
3.10	CLT for Spatial Models	110
3.11	Exercises	113
	Appendix 3.A Matrix Norm	113
	Appendix 3.B Inequalities	114
4	Maximum Likelihood Estimation	115
4.1	What Are The Consequences of Applying OLS?	115
4.1.1	Finite and Asymptotic Properties	115
4.1.2	Illustration of Bias	117
4.2	Maximum Likelihood Estimation of SLM	119
4.2.1	Maximum Likelihood Function	120
4.2.2	Score Vector and Estimates	121
4.2.3	Hessian	123
4.2.4	Ord's Jacobian	124
4.3	Maximum Likelihood Estimation of SEM	125
4.3.1	What Are The Consequences of Applying OLS on a SEM Model?	125

4.3.2	Log-likelihood function	126
4.3.3	Score Function and ML Estimates	127
4.4	Asymptotic Properties of SLM	129
4.4.1	Consistency of QMLE	129
4.4.2	Asymptotic Normality	134
4.5	Computing the Standard Errors For The Marginal Effects	135
4.6	Spillover Effects on Crime: An Application in R	136
4.6.1	Estimation of Spatial Models in R	136
4.6.2	Estimation of Marginal Effects in R	140
4.7	Programing the SLM in R	148
4.7.1	First approach	148
4.7.2	Second approach	154
4.8	Exercises	157
Appendix 4.A	Consistency of SLM Model	160
Appendix 4.B	Expected Value of Hessian for SLM	164
Appendix 4.C	Variance of the Score Function	166
Appendix 4.D	Proof of Asymptotic Normality	168
5	Hypothesis Testing	173
5.1	Test for Residual Spatial Autocorrelation Based on the Moran I Statistic	173
5.1.1	Cliff and Ord Derivation	173
5.1.2	Kelijan and Prucha (2001) Derivation of Moran's I	175
5.1.3	Example	175
5.2	Common Factor Hypothesis	176
5.3	Hausman Test: OLS vs SEM	177
5.4	Tests Based on ML	178
5.4.1	Likelihood Ratio Test	178
5.4.2	Wald Test	180
5.4.3	Lagrange Multiplier Test	182
5.4.4	Anselin and Florax Recipe	185
5.4.5	Lagrange Multiplier Test Statistics in R	185
5.5	Exercises	186
Appendix 5.A	Asymptotic Properties of Moran's I	186
6	Instrumental Variables and GMM	191
6.1	A Review of GMM	191
6.1.1	Model Specification	192
6.1.2	Asymptotic Distribution of One-Step GMM Estimator	194
6.1.3	Asypmtotic Distribution of Two-Step GMM Estimator	195
6.2	Spatial Two Stage Estimation of SLM	197
6.2.1	Instruments in the Spatial Context	197
6.2.2	Defining the S2SLS Estimator	198
6.2.3	Additional Endogenous Variables	203
6.2.4	Consistency of S2SLS Estimator	204
6.2.5	Asymptotic Distribution of S2SLS Estimator	205
6.2.6	Coding S2SLS in R	207
6.2.7	Best S2SLS Estimator	213

6.2.8	Coding BS2SLS Estimator	217
6.3	GMM Estimator of SLM	221
6.3.1	GMM Estimator Under Homoskedasticity	221
6.3.2	OGMM Estimator	231
6.3.3	Coding GMME and OGMME for SLM	232
6.3.4	Best GMM estimator	242
6.3.5	S2SLS Estimator as GMM Estimator	244
6.4	Feasible Generalized Least Squares Estimator for SEM Model	245
6.4.1	Generalized Least Squares Estimator	246
6.4.2	Moment Conditions	247
6.4.3	Feasible Generalized Least Squares Estimator	252
6.4.4	Coding the FSGLS estimator in R	255
6.5	GMM Estimator for SEM	260
6.5.1	GMM Estimator Under Homoskedasticity	260
6.5.2	Coding GMM Estimator for SEM	262
6.6	Estimation of SAC Model: The Feasible Generalized Two Stage Least Squares estimator Procedure	272
6.6.1	Intuition Behind the Procedure	272
6.6.2	Moment Conditions Revised	274
6.6.3	Assumptions	276
6.6.4	Estimators and Estimation Procedure in a Nutshell	278
6.6.5	Asymptotic Distribution	283
6.6.6	Coding FG2SLS Estimator Procedure	284
6.7	GMM Estimator for the SAC model	293
6.8	Exercises	294
Appendix 6.A	Asymptotic Distribution of GMME for SEM Model	294
Appendix 6.B	Proof Theorem 3 in KP 1998	296

Index	305
--------------	------------

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	15
1.9	Commune with largest number of contiguities	17
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	29
1.14	Moran Plot for Poverty	32
2.1	The SLM for two tegions	36
2.2	The SDM for Two Regions	37
2.3	The SEM for two regions	38
2.4	Taxonomy of spatial models	39
2.5	Regions east and west of the CBD	57
3.1	Convergence of sequence $2 + 3/n$	74
3.2	Bounded sequence	77
3.3	Illustration of convergnce in probability to a constant	78
3.4	Convergence of mean from normal distribution	86
3.5	Convergence of mean from binomial distribution	87
3.6	Chebychev's Convergence	90
3.7	Convergence of the sample mean and speed of convergence	100
4.1	Distribution of $\hat{\rho}$	119
4.2	Distances from R3 to all Regions	130
4.3	Spatial Distribution of Crime in Columbus, Ohio Neighborhoods	137

4.4	Effects of a Change in Region 30: Categorization	142
4.5	Effects of a Change in Region 30: Magnitude	143
6.1	Estimation steps for SAC model	279

List of Tables

4.1	Spatial Models for Crime in Columbus, Ohio Neighborhoods.	140
4.2	Comparing coefficients for SLM.	158
6.1	Comparing coefficients for SLM.	241
6.2	Comparing SE for SLM.	242

Part I

Introduction to Spatial Dependence

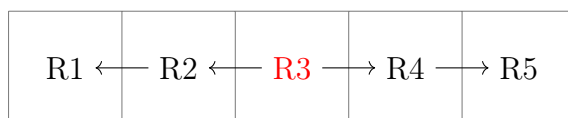
Introduction to Spatial Econometric

1.1 Why do We Need Spatial Econometric?

An essential consideration in any study involving spatial units, such as cities, regions, or countries, is the potential relationships and interactions between them. For instance, when modeling pollution at the regional level, it becomes impractical to treat each region as an independent entity. Regions cannot be analyzed in isolation since they are spatially interrelated through ecological and economic interactions.

Consider Figure 1.1, where Region 3 (R3) is highly industrialized, while Regions 1, 2, 4, and 5 are residential areas. If Region 3 increases its economic activity, pollution will not only rise within that region but will also affect neighboring regions. It is anticipated that contamination will increase in Regions 1 and 5, albeit to a lesser extent. These spatial externalities from R3 can be generated by both spatial-economic interactions (e.g., transportation of inputs and outputs from Region 3) and spatial-ecological interactions (e.g., carbon emissions).

Figure 1.1: Environmental Externalities



Similarly, when studying crime at the city level, it is crucial to incorporate the possibility that crime is localized. The identification of concentrations or clusters of higher criminal activity has become a central mechanism for targeting criminal justice and crime prevention responses. These crime clusters, commonly referred to as **hotspots**, represent geographic locations with high crime concentrations relative to the overall distribution of crime across the entire region of interest.

Both examples implicitly highlight the significance of geographic location and distance. They underscore the importance of the first law of geography, as articulated by Waldo Tobler: *“everything is related to everything else, but near things are more related than distant things”*. This foundational principle gives rise to fundamental concepts such as **spatial dependence** and **spatial autocorrelation**.

1.1.1 Spatial Dependence

Spatial dependence occurs when the values observed at one location or region, say observation i , depend on the values of neighboring observations at nearby locations. Formally, we can express this as:

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

In simpler terms, what happens in region i depends on what happens in region j for all $j \neq i$. For example, using our previous example displayed in Figure 1.1, 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} represents the effect of pollution in region j on region i . However, this model becomes impractical as it would result in a system with many more parameters than observations. With $n = 5$ observations, we would have to estimate 20 parameters, exceeding the available degrees of freedom. Intuitively, allowing for dependence relations between a set of n observations/locations introduces potentially $n^2 - n$ relations, accounting for the exclusion of dependence on oneself.

The crucial point is that, under standard econometric modeling, incorporating spatial dependency in such a manner is impractical. However, as we will explore in the next sections, we can efficiently integrate spatial relationships using the so-called **spatial weight matrix**.

1.1.2 Spatial Autocorrelation

Another crucial concept is **spatial autocorrelation**. In a spatial context, autocorrelation denotes the correlation between the values of a variable at two different locations. It can also be defined as the correlation between the same attribute at two (or more) different locations or the coincidence of values' similarity with location similarity. Essentially, spatial autocorrelation investigates whether the presence of a variable in one region of a spatial 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 implies the absence of any discernible spatial pattern in the data. In other words, the value observed in one 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, there is evidence of spatial structure.

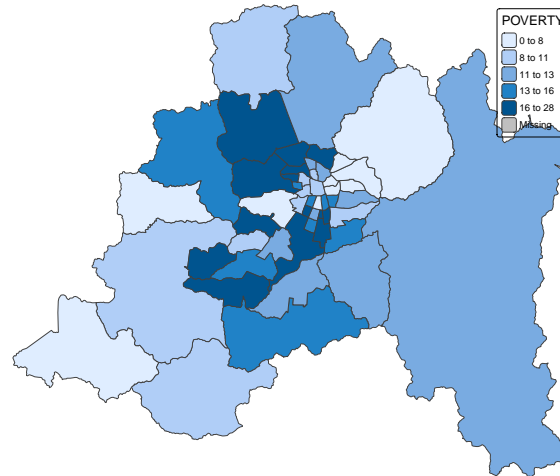
As an illustrative example, Figure 1.2 depicts the spatial distribution of poverty in the communes of the Metropolitan Region, Chile. It reveals a discernible spatial pattern where communes with similar poverty rates are clustered.

Formally, the presence of spatial autocorrelation can be expressed through the following moment conditions:

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

where y_i and y_j are observations on a random variable at locations i and j in space. Here, i and j can represent either points or areal units. Therefore, nonzero spatial autocorrelation

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



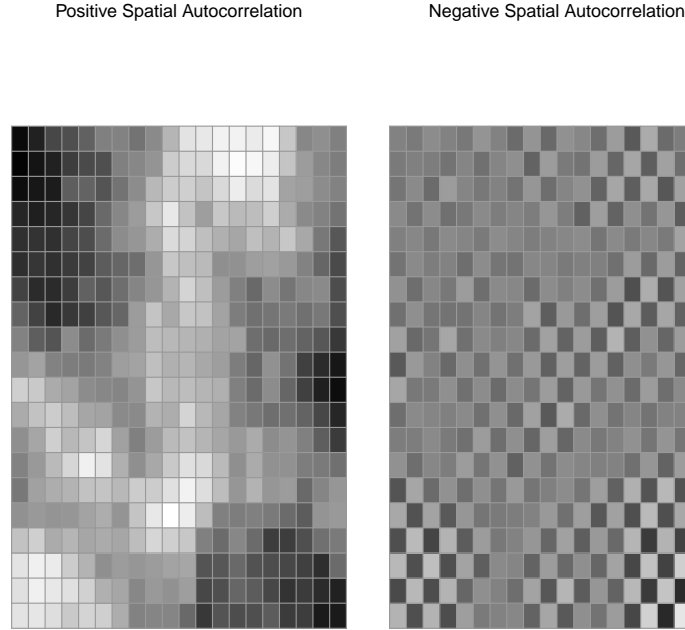
Notes: This graph shows the spatial distribution of poverty in the Metropolitan Region, Chile.

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 refer to it as **positive spatial autocorrelation**; conversely, if the opposite is true, we term it **negative spatial autocorrelation**. Figure 1.3 illustrates examples of positive and negative spatial autocorrelation.

Positive autocorrelation is more commonly observed, but negative autocorrelation does exist. Instances of negative autocorrelation can be found in studies on welfare competition or federal grants competitions among local governments (Saavedra, 2000; Boarnet and Glazer, 2002), regional employment (Filiztekin, 2009; Pavlyuk, 2011), cross-border lottery shopping (Garrett and Marsh, 2002), foreign direct investment in OECD countries (Garretsen and Peeters, 2009), and the locations of the Turkish manufacturing industry (Basdas, 2009). In essence, our interest lies in studying non-random spatial patterns and explaining this non-randomness. Potential causes of non-randomness, as outlined by Gibbons et al. (2015), include:

- (a) Economic agents may be randomly allocated across space, but some characteristics of locations vary across space and influence 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) Individuals or firms may be randomly allocated across space, but they interact in a way that decisions by one agent affect the outcomes of other agents.
- (d) Individuals or firms may be non-randomly allocated across space, and the characteristics of others nearby directly influence individual outcomes.

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.

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. Assuming there are n spatial objects (regions, cities, countries), the \mathbf{W} matrix is a square matrix of dimension $n \times n$. This matrix imposes a structure in terms of identifying neighbors for each location, assigning weights that measure the intensity of the relationship among pairs of spatial units. Each element (i, j) of \mathbf{W} , denoted as w_{ij} , expresses the degree of spatial proximity between the pair. The matrix can be represented as follows:

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

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

A more formal definition of a spatial weight matrix is as follows:

Definition 1.2.1 — Spatial Weight Matrix. Let n be the number of spatial units. The spatial weight matrix, \mathbf{W} , is a $n \times n$ **positive** 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 means that $w_{ij} \geq 0$ for all $i \neq j$. 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

Polygon or lattice data allow for the construction of contiguity-based spatial weight matrices, which represent spatial relationships among regions. A typical specification of the contiguity relationship in such matrices is given by:

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

In regular grids, the definition of contiguity can vary. Drawing an analogy to chess, three common contiguity criteria are rook contiguity, bishop contiguity, and queen contiguity. Each criterion defines neighborhood relationships differently, as detailed below.

Rook Contiguity

Rook contiguity considers two regions as neighbors if they share a common border or side. For example, in the regular grid shown in Figure 1.4, which contains 9 regions (each represented as a square), the neighbors of region 5 under the rook criterion are regions 2, 4, 6, and 8 (highlighted in red).

Using this criterion, the 9×9 spatial weight matrix \mathbf{W} is:

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

Figure 1.4: Rook Contiguity

1	2	3
4	5	6
7	8	9

Bishop Contiguity

Under bishop contiguity, two regions are neighbors if they share a common corner. This criterion is less frequently used in practice. In Figure 1.5, the neighbors of region 5 are regions 1, 3, 7, and 9 (highlighted in red). Note that regions located at the grid's interior have more neighbors than those along the periphery.

Figure 1.5: Bishop Contiguity

1	2	3
4	5	6
7	8	9

The corresponding \mathbf{W} matrix is:

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

Queen Contiguity

Queen contiguity considers two regions as neighbors if they share either a common side or a common corner. In Figure 1.6, the neighbors of region 5 include all surrounding regions: 1, 2, 3, 4, 6, 7, 8, and 9 (highlighted in red).

Figure 1.6: Queen Contiguity

1	2	3
4	5	6
7	8	9

1.2.2 Weights Based on Distance

Weights can also be defined as a function of the distance between regions i and j , denoted as d_{ij} . Typically, this distance is computed as the separation between centroids, although other significant points, such as capital cities or major urban centers, could also be used. Notably, unlike contiguity-based weights, distance-based matrices only require the coordinates of the relevant points.

Several methods exist to compute the distance between two spatial units. Let x_i and x_j represent the longitudes, and y_i and y_j denote the latitudes of regions i and j . The **Minkowski** metric provides a general formula for distance:

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

where p is a parameter that allows flexibility in distance computation. A commonly used variant is the Euclidean distance, which corresponds to $p = 2$:

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

Another popular alternative is the Manhattan (or block) distance, which considers movement along east-west and north-south directions, i.e., along straight angles. This metric corresponds to $p = 1$:

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

While these metrics are suitable for treating the Earth as a plane, they may lack accuracy over larger distances due to the Earth's curvature. The Euclidean distance, for instance, represents the straight-line length on a map but might not align with the shortest path on the Earth's surface. For more accurate analyses, particularly for long-distance travel, the **great circle distance** is often preferred. This method considers the Earth's curvature and calculates the shortest path between two points on a sphere. The great circle distance is given by:

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

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 must translate information about distances among spatial points into a weight scheme. The objective is to ensure that $w_{ij} \rightarrow 0$ as $d_{ij} \rightarrow \infty$. In simpler terms, as point j gets farther from point i , the spatial weight w_{ij} should decrease, aligning with Tobler’s first law.

In the inverse distance weighting scheme, the weights are inversely proportional to the separation distance, as expressed by the following formula:

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

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 $\alpha = 2$. Consequently, the weights become the reciprocal of the distance: the greater the distance between spatial units, the smaller the spatial weight or connection. Conventionally, diagonal elements of spatial weights are set to zero to avoid division by zero in the case of inverse distance weights. Plugging in a value of $d_{ii} = 0$ would yield division by zero for inverse distance weights.

Negative Exponential Model

In the negative exponential model, weights decrease exponentially with separation distance:

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

where α is again 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 negative exponential distance models rely on the parameter value, functional form, and the chosen distance metric. As the weights are inversely related to distance, larger distances yield smaller weights, and vice versa. A potential challenge arises when distances are so vast that inverse distance weights approach zero, possibly resulting in a spatial weight matrix with zero values. Additionally, issues may occur if the distance metric yields values less than one, which is typically undesirable ([Anselin and Rey, 2014](#)).

k -nearest Neighbors

An alternative approach to spatial weights that mitigates the issue of isolates involves selecting the k -nearest neighbors. Unlike the distance band, this relation is not symmetric. However, a challenge arises when ties occur—when multiple locations j share the same distance from i . Several solutions exist to address ties, ranging from randomly selecting one of 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 prevent isolates resulting from an excessively stringent critical distance, the distance must be chosen so 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.

Importantly, a weights matrix derived from a distance band is always symmetric, as distance is inherently a symmetric relation.

1.2.3 Row-Standardized Weights Matrix

In practical applications, spatial weights are rarely used in their raw binary (or distance) form. Instead, they are often transformed or standardized to improve interpretability and comparability. One common approach is to compute weighted averages, placing greater emphasis on nearby observations than on distant ones. To achieve this, a **row-standardized** weight matrix \mathbf{W}^s is defined, with elements w_{ij}^s given by:

$$w_{ij}^s = \frac{w_{ij}}{\sum_{j=1}^n w_{ij}}.$$

This formulation ensures that all weights lie within the range of 0 to 1, facilitating the interpretation of matrix operations as an averaging of neighboring values. Moreover, row-standardization enhances the comparability of spatial parameters across models in various spatial stochastic processes (Anselin and Bera, 1998).

An additional feature of row-standardized matrices is that the sum of the weights in each row equals unity, and the total sum of all weights, $S_0 = \sum_i \sum_j w_{ij}$, equals the number of observations, n . This property simplifies the interpretation and application of the weights, as we will explore further in subsequent sections.

However, a key consequence of row-standardization is the loss of symmetry in the spatial weights matrix. While symmetric matrices have the desirable property of real eigenvalues, row-standardized matrices are generally asymmetric and do not preserve this characteristic. Nevertheless, row-standardization remains prevalent in applied work due to its intuitive interpretation and computational convenience.

Row-standardized matrices are also referred to as row-stochastic or Markov matrices. Formally, a row-stochastic matrix is defined as follows:

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 property of row-stochastic matrices concerns their eigenvalues:

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

Having explored the spatial weight matrix, we can now introduce the concept of **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 $\mathbf{y}_{Li} = \sum_{j=1}^n w_{ij}y_j$, i.e., a weighted average of the \mathbf{y} values in the neighbor of i .

For example, consider

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

Using a row-standardized weight matrix:

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


Consequently, for spatial unit i , the spatial lag of y_i , denoted a \mathbf{y}_{Li} (the variable $\mathbf{W}\mathbf{y}$ observed for location i) is:

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

or equivalently,

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

where the weights w_{ij} are the elements of the i th row of the matrix \mathbf{W}^s matched with the corresponding elements of the vector \mathbf{y} . In other words, this represents a weighted sum of values observed at neighboring locations, excluding non-neighbors.

 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 Weights

To expand our understanding of geographical space as defined by the matrix \mathbf{W} , we turn to the concept of higher-order neighbors. These are neighbors that are not directly adjacent but are reachable through other spatial units. For instance, we might consider the neighbors of a spatial unit's neighbors or even their neighbors' neighbors. To formalize this, we introduce the notion of **higher-order spatial weight matrices**.

The higher-order spatial weight matrix of order l , denoted as \mathbf{W}^l , is defined as:

$$\mathbf{W}^l = \underbrace{\mathbf{W}\mathbf{W} \cdots \mathbf{W}}_{l \text{ times}}.$$

For example:

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

- The second-order spatial weight matrix is $\mathbf{W}^2 = \mathbf{W}\mathbf{W}$.
- The third-order spatial weight matrix is $\mathbf{W}^3 = \mathbf{W}\mathbf{W}\mathbf{W}$, and so on.

The element w_{ij} in these higher-order matrices represents whether spatial unit j is a neighbor of order l to spatial unit i . Specifically:

- For $l = 2$, w_{ij} equals 1 if j is adjacent to a first-order neighbor of i , and 0 otherwise.
- For $l = n$, w_{ij} equals 1 if j is adjacent to the $(n - 1)$ -order neighbors of i , and 0 otherwise.

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

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

Then, $\mathbf{W}^2 = \mathbf{W}\mathbf{W}$ based on the 5×5 first-order contiguity matrix \mathbf{W} from (1.1) is:

$$\mathbf{W}^2 = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 2 & 0 & 1 & 0 \\ 1 & 0 & 2 & 0 & 1 \\ 0 & 1 & 0 & 2 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix} \quad (1.2)$$

For region $R1$, the second-order neighbors are $R1$ and $R3$. This indicates that $R1$ is its own second-order neighbor (via feedback) and also shares second-order adjacency with $R3$, which is a first-order neighbor of $R2$.

Now consider $R2$. The first panel of Figure 1.7 shows the first-order neighbors of $R2$ based on the spatial weight matrix in (1.1): $R1$ and $R3$. Panel B considers the second-order neighbors of $R2$: $R2$ 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 $w_{22}^2 = 2$. Additionally, there is an indirect effect coming from $R4$ through $R3$ that finally impacts $R2$, yielding a value of 1 for the element w_{24}^2 .

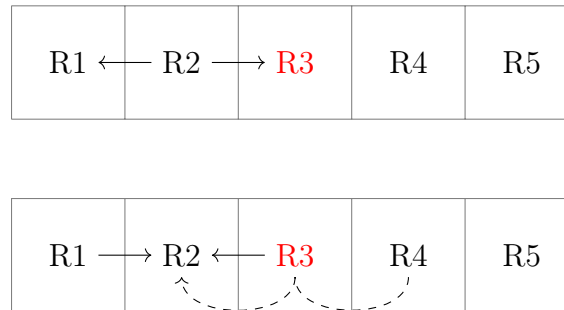
Similarly, for region $R3$, the second-order neighbors are regions $R1$ (which is a neighbor to the neighboring region $R2$), $R3$ (a second-order neighbor to itself), and $R5$ (which is a neighbor to the neighboring region $R4$).

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?

Figure 1.7: Higher-Order Neighbors



1.3 Examples of Weight Matrices in R

Creating spatial weight matrices manually is a tedious and error-prone process, especially for large datasets. Fortunately, modern statistical software provides robust tools to simplify this task. To start, we typically need a **shapefile**, a widely used format for storing geographical information.

The shapefile format is a digital vector storage format that supports geometric data types such as points, lines, and polygons, along with associated attributes. It enables diverse representations of geographic data by combining shapes with their corresponding attribute data. A complete shapefile consists of three mandatory files with the following extensions:

- **.shp**: The shape file containing the feature geometry.
- **.shx**: The shape index file, which facilitates fast positional indexing of the feature geometry.
- **.dbf**: The attribute file, formatted in dBase IV format, storing columnar attributes for each shape.

These three components must be present for a shapefile to be functional. The **.shp** file stores the actual geometric data, while the **.shx** and **.dbf** files provide supporting information to enable efficient access and attribute storage.

In this example, we demonstrate how to create spatial weight matrices using R. Specifically, we focus on a map of the communes in the Metropolitan Region of Chile. To begin, we load the shapefile into R using the **sf** package (Pebesma, 2018), which is designed for handling spatial data.

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

If the shapefile **mr_chile.shp** is located in the working directory, we can load it using the **read_sf** function:

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

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


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

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

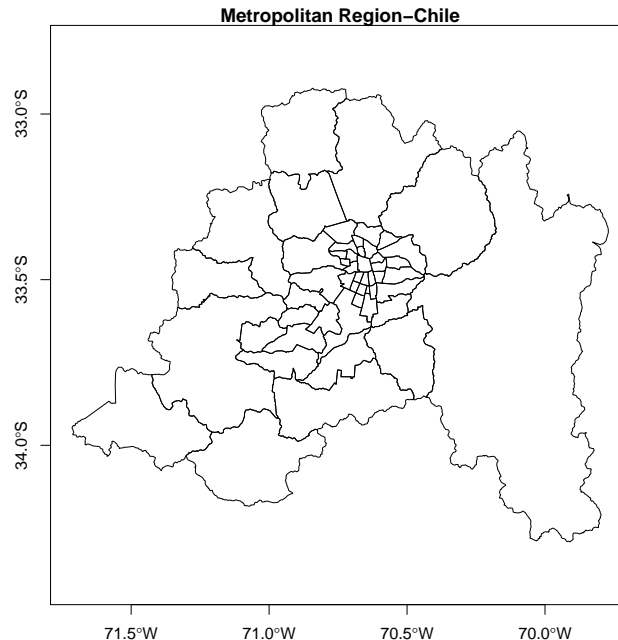
## [1] "ID"          "NAME"        "NAME2"       "URB_POP"     "RUR_POP"
## [6] "MALE_POP"    "TOT_POP"     "FEM_POP"     "N_PARKS"     "N_PLAZA"
## [11] "CONS_HOUSE"  "M2_CONS_HA"  "GREEN_AREA"  "AREA"        "POVERTY"
## [16] "PER_CONTR_"  "PER_HON_SA"  "PER_PLANT_"  "NURSES"      "DOCTORS"
## [21] "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"
```

Now, let’s visualize the shapefile using the `plot` function:

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

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

Figure 1.8: Plotting a Map in R



1.3.1 Creating Contiguity Neighbors

To construct spatial weight matrices, the `spdep` package (Bivand et al., 2013) provides an extensive suite of tools. After installation, the package can be loaded as follows:

Mauricio Sarrias

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

In **spdep**, neighbor relationships between n observations are represented by objects of class **nb**. These objects are lists of length n , where each element is an integer vector containing the indices of neighboring regions. If a region has no neighbors, the corresponding element is assigned an integer value of zero.

The **poly2nb** function constructs neighbor relationships based on **contiguity**, generating a neighbors list of class **nb**. This function supports both the Queen and Rook criteria for defining spatial relationships. Refer to **help(poly2nb)** for detailed documentation.

As explained in previous section, the Queen criterion defines neighbors as regions that share at least one vertex or edge. The following example demonstrates how to construct a neighbor list for the communes in the Metropolitan Region of Chile:

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

Once the **nb** object is created, it can be explored using standard methods such as **print**, **summary**, and **plot**. For example, the summary of the Queen-based neighbors list provides information about the spatial relationships:

```
# 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:
## Tiltil San Pedro Maria Pinto with 2 links
## 1 most connected region:
## San Bernardo with 12 links
```

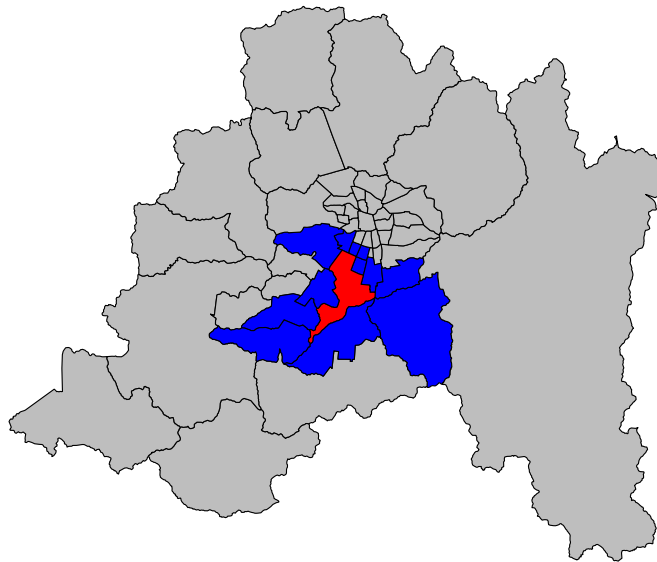
The output provides crucial information about the neighbors, including the number of regions (52 communes in this example), the number of nonzero links, the percentage of nonzero weights, the average number of links, and more.

For instance, the commune of San Bernardo stands out as the most connected region with 12 neighbors under the queen scheme. Conversely, the least connected regions are Tiltil, San Pedro, and Maria Pinto, each with only 2 neighbors. The output also shows the distribution of neighbors, revealing that 7 out of 52 regions have 4 neighbors, and only 2 communes have 8 neighbors.

To visualize the region with the largest number of neighbors (San Bernardo) and its immediate neighbors, use the following code:

```
# 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: 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.

Figure 1.9 provides a visual representation where the red-colored commune, San Bernardo, stands out as the spatial unit with the largest number of neighbors according to the queen criteria. Meanwhile, the communes in blue represent its neighboring units.

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

```
# 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:
## Tiltit San Pedro Maria Pinto with 2 links
## 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 for `nb2listw` 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 sums to 1. Other options include "B" for basic binary coding; "C" for global standarization, 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).

Additionally, the `summary` function provides several constants essential for global spatial autocorrelation statistics, which we will discuss later.

We can also inspect 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 Florida"
## [10] "La Pintana"        "La Reina"
## [13] "Lo Prado"          "Macul"
## [16] "Pedro Aguirre Cerda" "Providencia"
## [19] "Quinta Normal"     "Renca"
## [22] "San Joaquin"       "San Ramon"
## [25] "Independencia"     "Las Condes"
```

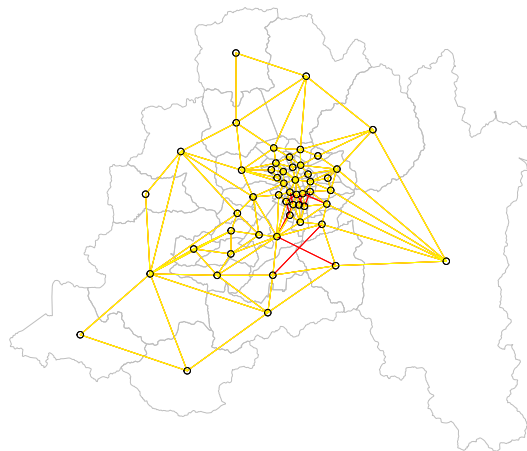


```
## Number of nonzero links: 272
## Percentage nonzero weights: 10.05917
## Average number of links: 5.230769
## Link number distribution:
##
##  2  3  4  5  6  7  8  9 10
##  3  3 12 16  7  6  2  1  2
## 3 least connected regions:
## Tiltit San Pedro Maria Pinto with 2 links
## 2 most connected regions: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

Finally, we can visualize the spatial connectivity implied by the Queen and Rook criterion 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 proceed to build 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
```

Here, the `coords` function extracts spatial coordinates from the shapefile, while `knearneigh` returns a matrix containing indices of points belonging to the set of k -nearest neighbors for each observation. The `k` argument specifies the number of nearest neighbors to return. If the point coordinates are given in longitude-latitude decimal degrees, distances are measured in kilometers when `longlat = TRUE`. For `longlat = FALSE`, great-circle distances are computed. The resulting objects `k1neigh` and `k2neigh` are of class `knn`.

Inverse distance weight matrices can be computed as follows (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 `dist` function from the **stats** package computes the distance matrix using the specified metric—Euclidean distance in this example. Other available methods include `maximum`, `manhattan`, `canberra`, `binary`, and `minkowski`. The `mat2listw` function converts a square spatial weight matrix into a list format suitable for spatial analysis. For additional details on spatial weight matrices, see [Stewart and Zhukov \(2010\)](#).

The following code demonstrates how to plot different weight matrices:

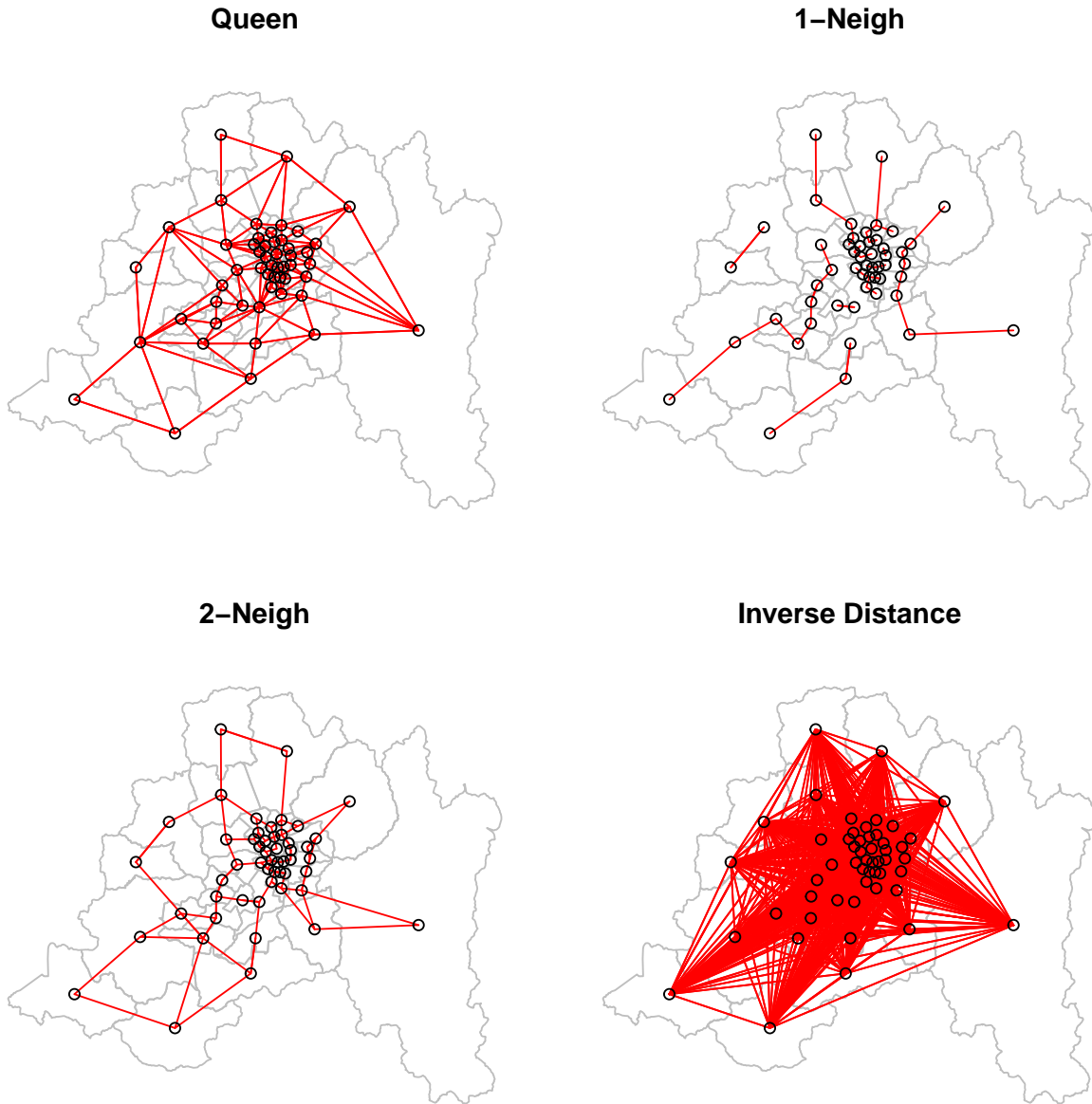
```
# Plot Weights
par(mfrow = c(2, 2))
plot(st_geometry(mr), border = "grey", main = "Queen")
plot(queen.w, coords, add = TRUE, col = "red")
plot(st_geometry(mr), border = "grey", main = "1-Neigh")
plot(knn2nb(k1neigh), coords, add = TRUE, col = "red")
plot(st_geometry(mr), border = "grey", main = "2-Neigh")
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 play a crucial role in various spatial tests and regression specifications. In the **spdep** package, these variables are created using the `lag.listw` function.

Let's begin by combining the variables `POVERTY` and `URB_POP` into a matrix and check the contents with `head`:

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 15919
## [2,]    9  65262
## [3,]   18 131850
## [4,]   12 104634
## [5,]   14 166514
```

Now, we can generate 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 introduced in Section 1.1.2, spatial autocorrelation refers to the relationship between the values of a variable across space. Positive spatial autocorrelation arises when similar values cluster together—high values near high values or low values near low values. In contrast, negative spatial autocorrelation reflects spatial dispersion, where high values are surrounded by low values, or vice versa.

A central empirical question is whether the observed spatial pattern reflects a true spatial process or is simply the result of random chance. Formal hypothesis tests are therefore necessary to evaluate whether the value of a variable at one location is statistically dependent on the values observed at neighboring locations.

1.4.1 Global Spatial Autocorrelation: Moran's I

Global spatial autocorrelation measures the overall degree of spatial clustering or dispersion in a dataset. These indices evaluate the similarity between values at different locations i and j , weighted by a spatial proximity matrix. High similarity in close proximity suggests clustering, while dissimilar values nearby imply spatial dispersion.

The most widely used measure of global spatial autocorrelation is **Moran's I**.² Moran's I tests the null hypothesis of spatial randomness and quantifies the extent of global spatial association.

Moran's I is given by:

$$I = \frac{\sum_{i=1}^n \sum_{j=1, j \neq i}^n w_{ij} (x_i - \bar{x}) (x_j - \bar{x})}{S_0 \sum_{i=1}^n (x_i - \bar{x})^2 / n} = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{ij} (x_i - \bar{x}) (x_j - \bar{x})}{S_0 \sum_{i=1}^n (x_i - \bar{x})^2}, \quad (1.3)$$

where w_{ij} denotes the spatial weight between regions i and j , and $S_0 = \sum_{i=1}^n \sum_{j=1}^n w_{ij}$ is the sum of all spatial weights.

²Other measures, such as Geary's C , are also available, but this section focuses on Moran's I due to its popularity and interpretability.

In matrix notation, this can be expressed as:

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

where:

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

If \mathbf{W} is row-standardized, then $S_0 = n$, and the expression simplifies to:

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

Moran's I ranges from -1 (perfect dispersion) to $+1$ (perfect clustering). A value close to zero suggests spatial randomness.

A helpful graphical tool for interpreting Moran's I is the **Moran Scatterplot**. It plots the standardized values of a variable (horizontal axis) against the standardized spatial lag (vertical axis), i.e., the average value of neighboring units (see Figure 1.12). As noted by Anselin (1996), standardizing both axes (mean zero, variance one) allows visualization of both global and local spatial autocorrelation. The slope of the best-fit line corresponds to Moran's I , while each quadrant reveals different types of local spatial association.

The quadrants of the Moran Scatterplot indicate:

- **Quadrant I (High-High)**: Regions with above-average values surrounded by neighbors also above average.
- **Quadrant II (Low-High)**: Regions with below-average values surrounded by neighbors with above-average values.
- **Quadrant III (Low-Low)**: Regions with below-average values surrounded by similarly low neighbors.
- **Quadrant IV (High-Low)**: Regions with above-average values surrounded by below-average neighbors.

Quadrants I and III reflect **positive spatial autocorrelation** (clustering of similar values), while Quadrants II and IV reflect **negative spatial autocorrelation** (clustering of dissimilar values).

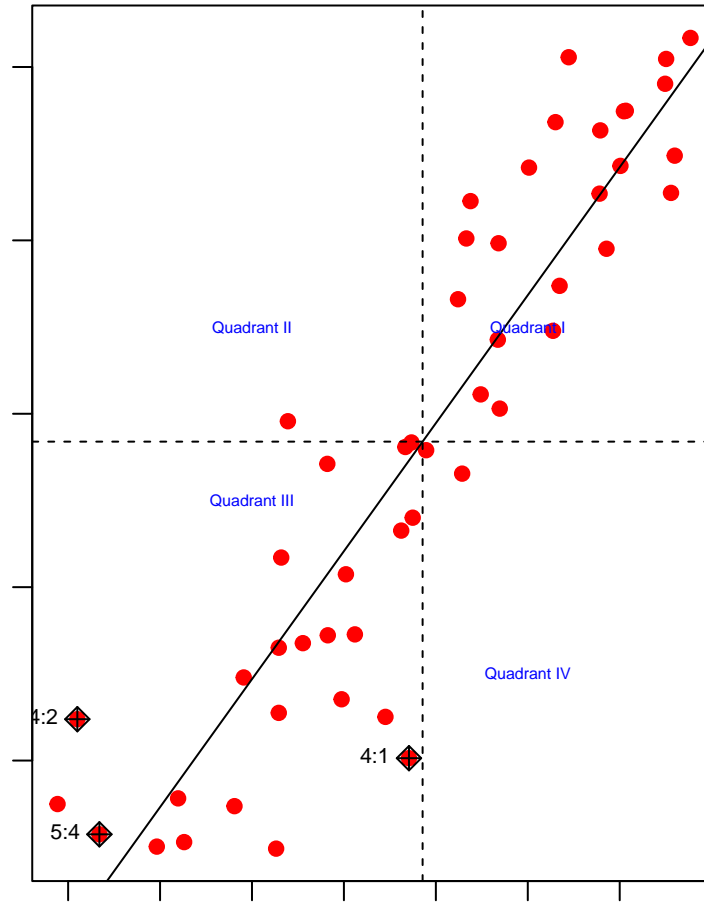
To grasp the essence of Moran's I , it is essential to draw parallels with the Ordinary Least Squares (OLS) coefficient. Recall the OLS coefficient formula:

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

Now, examining Equation (1.3), Moran's I equates to the slope coefficient of a linear regression, where the spatial lag $\mathbf{W}\mathbf{x}$ is regressed on the observation vector \mathbf{x} , both measured in deviation from their means. It's worth noting that Moran's I is not equivalent to the slope of \mathbf{x} on $\mathbf{W}\mathbf{x}$, which might seem more intuitive.

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.

Now, considering the distribution of Moran's I , our interest lies in the distribution of:

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

There are two methods to compute the mean and variance of Moran's I . The first assumes a normal distribution for x_i , while the second involves randomization of x_i . Under the normal assumption, it is assumed that the random variable x_i results from n independent draws from a normal population. Conversely, under the randomization assumption, irrespective of the underlying distribution of populations, observed values of x_i are repeatedly and 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 $\{x_i\} = \{x_1, x_2, \dots, x_n\}$ are independent and distributed as $N(\mu, \sigma^2)$, but μ and σ^2 are unknown. Then:

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

and

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

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:

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

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:

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

and

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

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:

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

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 essence of any Monte Carlo test, especially for Moran's I , involves the following steps:

- Specify a test statistic T for which large values indicate evidence against the null hypothesis H_0 (no spatial autocorrelation).
- Given an observed value t_{obs} of the test statistic, compute the p-value as $\Pr(T \geq t_{obs} | H_0)$. This involves understanding the distribution of T under the assumption of H_0 .

The algorithm for the Moran's I Monte Carlo test is as follows:

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:

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

- (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 conduct an Exploratory Spatial Data Analysis (ESDA) of poverty in the Metropolitan Region of Chile.

1.5.1 Cloropeth Graphs

To understand the spatial variation in poverty, a natural starting point is to visualize its geographic distribution. Such visualizations serve multiple purposes. National or aggregate indicators often mask significant disparities across smaller spatial units, such as communes. Poverty mapping allows us to uncover and highlight these localized differences. In addition, maps are powerful tools for communicating complex spatial information in an intuitive and accessible way.

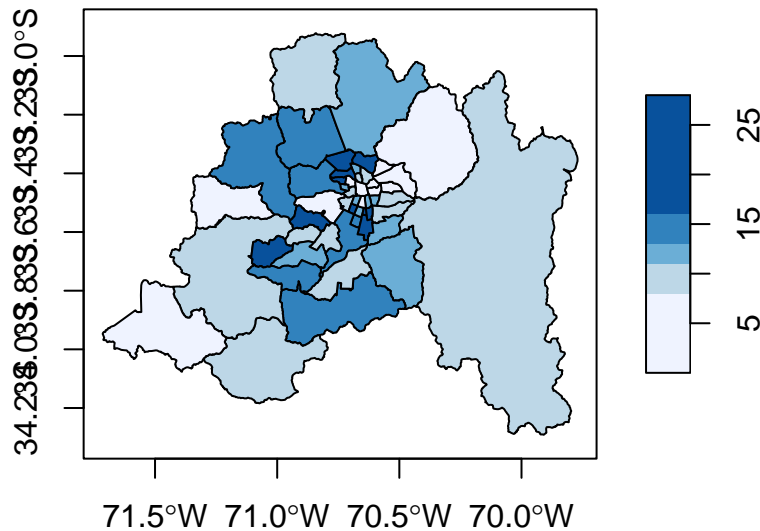
We begin by generating a choropleth map³ using the `plot` function. Specifically, we employ a quantile classification scheme, in which the range of the poverty variable is divided into intervals containing an equal number of spatial units. This facilitates comparisons across areas with varying levels of poverty by grouping them into ranked categories.

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

Figure 1.13 provides valuable insights into the spatial structure of poverty in the Metropolitan Region. First, it reveals that poverty is not uniformly distributed across the territory. Instead, there are marked differences in poverty levels among communes. Second, the map demonstrates the added value of disaggregated spatial indicators. For example, it becomes

³The term choropleth originates from the Greek words *choros* (space) and *pleth* (value).

Figure 1.13: Choropleth map: Poverty in the Metropolitan Region



evident that poverty is more pronounced in central communes, while peripheral areas tend to have lower levels of poverty—an insight that would be obscured in aggregate regional statistics.

In a quantile classification, data values are ranked and divided into intervals such that each class contains the same number of observations. This approach is particularly useful for comparing spatial units on a relative scale. Since each color represents an equal share of the data, quantile maps are easy to interpret and visually balanced, though they may sometimes exaggerate differences when the underlying data distribution is skewed.

While the visual inspection of poverty's spatial distribution in Figure 1.13 offers preliminary insights, it is important to recognize the sensitivity of such graphical analysis to arbitrary decisions—such as the number and boundaries of classification intervals. Therefore, a more formal statistical approach is necessary to evaluate whether the observed spatial pattern is the result of a genuinely spatial process or simply due to random variation.

To this end, we employ Moran's I test to formally assess the presence of spatial autocorrelation in the distribution of poverty.

1.5.2 Moran's I Test

We begin by constructing two types of spatial weight matrices—queen and rook contiguity—to evaluate the robustness of the test under alternative neighborhood definitions:

```
# 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 for spatial autocorrelation is implemented in the **spdep** package (Bivand and Piras, 2015). There are two main functions for conducting this test:

- `moran.test`: performs the test under either a normality or randomization assumption;

- `moran.mc`: computes a permutation-based test using Monte Carlo simulations.

The code below performs the test under the normality assumption (i.e., setting `randomisation = FALSE`) for both weight matrices:

```
# 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
```

By default, `moran.test` uses a one-sided test and the randomization assumption. To obtain a two-sided test under the normal approximation, we must explicitly set `randomisation = FALSE` and `alternative = 'two.sided'`. Note also that the option `zero.policy = FALSE` implies that spatial units with no neighbors (islands) will yield NA; setting this option to TRUE assigns them a spatial lag value of zero.

The test results reveal that Moran's I is approximately 0.30 (queen) and 0.34 (rook), both highly significant. This provides strong evidence of positive spatial autocorrelation in poverty rates, i.e., communes with similar levels of poverty tend to be located near each other.



When applying Moran's I to different variables using the same spatial weight matrix, the expected value and variance of the test statistic under the normal approximation remain unchanged. Why? Because under normality, the expectation and variance depend only on the spatial structure (i.e., the weights matrix), not the values of the variable.

We now replicate the test under the randomization assumption, which treats the observed poverty values as a random permutation of fixed values across space:

```
# 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 that the value of the Moran's I statistic and its expected value remain the same as in the normal case. However, the variance differs, leading to a different test statistic and p -value.

To implement a Monte Carlo version of the test, we use `moran.mc`, which requires specifying the number of permutations `nsim`. Best practices suggest using a number ending in 9 (e.g., 99 or 999), as this leads to pseudo p -values that round to familiar significance levels:

```
# 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
```

The result shows that none of the permuted datasets produced a Moran's I greater than the observed value of 0.3065, resulting in a pseudo p -value of $(0 + 1)/(99 + 1) = 0.01$.

We can also visualize the test using the Moran scatterplot, which plots each unit's value against its spatial lag:

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

Figure 1.14: Moran Plot for Poverty

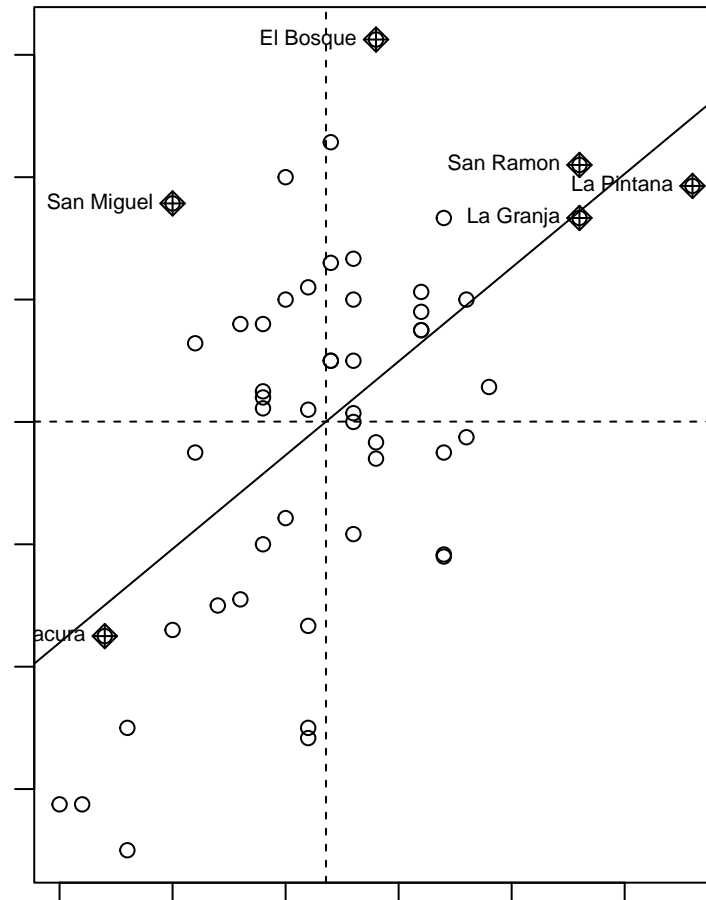


Figure 1.14 shows that most communes fall in Quadrants I and III, suggesting strong positive spatial autocorrelation: areas with high (low) poverty tend to be near others with high (low) poverty. Exceptions are visible in Quadrants II and IV. For example, San Miguel displays low poverty despite being surrounded by high-poverty neighbors.

A major limitation of Moran's I is that it only captures global spatial autocorrelation. While it indicates whether spatial clustering exists overall, it does not identify where such clusters or outliers occur. To investigate local patterns of spatial dependence, local indicators of spatial association (LISA) are needed.

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 denote the per capita income in cross-sectional unit $r = 1, \dots, n$. Consider the following definition for the spatial weight w_{ij} :

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

where $\alpha > 0$ is a fixed positive constant. Show that the constant α cancels out when the weight matrix $\mathbf{W} = [w_{ij}]$ is row-standardized.

Exercise 1.4 Let the Moran's I be defined as:

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

where $\mathbf{z} = \mathbf{P}\mathbf{x}$, $\mathbf{P} = \mathbf{I}_n - \frac{1}{n}\mathbf{J}_n$ is the centering matrix, \mathbf{I}_n is the $n \times n$ identity matrix, \mathbf{J}_n is the $n \times n$ matrix of ones, and $\mathbf{W}^s = (\mathbf{W} + \mathbf{W}^\top)/2$. The matrix \mathbf{P} is symmetric, positive semi-definite, and idempotent. Assume that $\mathbf{x} = (x_1, \dots, x_n)^\top$ is a vector of independent random variables such that $x_i \sim \mathcal{N}(\mu, \sigma^2)$, with unknown μ and σ^2 .

- Show that $\text{tr}(\mathbf{P}) = n - 1$.
- Using Lemma 3.25 on the moments of quadratic forms, show that: $\mathbb{E}(\mathbf{z}^\top \mathbf{z}) = \sigma^2(n - 1)$.
- Show that $\mathbb{E}(\mathbf{z}^\top \mathbf{W}^s \mathbf{z}) = \sigma^2 \text{tr}(\mathbf{W}^s \mathbf{P})$.
- Using the results above, show that under the null hypothesis of spatial independence: $\mathbb{E}(I) = -1/(n - 1)$.
- Derive $\mathbb{E}(I^2)$ using Lemma 3.25 for the second moment of quadratic forms.
- Check your results using R and compare them with the moments in Theorem 1.2

Exercise 1.5 In this applied example, we will work with the Columbus data set. Load the data and spatial weight matrix:

```
columbus      <- st_read(system.file("shapes/columbus.gpkg",
                                     package="spData"))[1], quiet=TRUE)
columbus_nb   <- poly2nb(columbus)
W             <- nb2listw(columbus_nb, style = "W")
```

- Plot a choropleth map of crime rates using the variable `CRIME`. What can you say about the spatial pattern? Are high-crime and low-crime areas clustered?

- (b) Create both `queen` and `rook` contiguity matrices using the `poly2nb()` function. Compare the two neighborhood structures: how many neighbors does each unit have under each structure?
- (c) Use `moran.test()` to compute Moran's I for the variable `CRIME` using both queen and rook matrices. Use the normality-based inference (`randomisation = FALSE`).
- (d) Repeat the test using the permutation-based method `moran.mc()` with 999 simulations. Comment on the statistical significance and interpretation of Moran's I in both approaches. Are your conclusions robust to the choice of spatial weights?
- (e) Generate a Moran scatterplot using `moran.plot()` for the variable `CRIME`. What does the Moran scatterplot tell you about the nature of spatial dependence?

Exercise 1.6 Create in R your own function to plot a Moran Scatterplot. Show that your function works well using a simulated example.

In the preceding sections, we reviewed fundamental concepts of spatial econometrics, such as spatial dependency and spatial autocorrelation. This chapter advances our understanding by examining the formulation of spatial models.

In Section 2.1, we establish a comprehensive taxonomy of spatial models, encompassing the Spatial Lag Model, Spatial Durbin Model, Spatial Error Model, and the Spatial Autocorrelation Model. Each model is motivated and illustrated with examples to provide a clear understanding.

Moving forward to Section 2.4, we explore the concept of “spillover” effects within the spatial model framework. Additionally, we delve into the interpretation of marginal effects, enhancing our ability to extract meaningful insights from spatial econometric analyses.

2.1 Taxonomy of Models

As demonstrated in the previous chapter, particularly in Section 1.1.1, traditional econometric methods are not equipped to handle spatial dependencies. The primary challenge arises from the fact that we often have more parameters than observations. However, by using the spatial weight matrix, we can address this issue by reducing the number of parameters to just one. This is achieved through the weighted average of the dependent variable values, y , in the neighborhood of unit i .

2.1.1 Spatial Lag Model

Given the problem of insufficient degrees of freedom, a natural question arises: how can we model a situation where the dependent variable is influenced by spatially lagged values? Instead of relying on a full system of equations, we can capture the spatial dependence with a model that includes a spatially lagged dependent variable. Specifically, we model the spatial dependence as follows:

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

where w_{ij} is the (i, j) th element of the spatial weight matrix \mathbf{W} matrix (refer to Definition 1.2.1). The variable y_i denotes the dependent variable for spatial unit i , and $\sum_{j=1}^n w_{ij} y_j$ is the

weighted average of the dependent variable for the neighbors of i (referred to as the spatial lag). The error term ϵ_i satisfies $\mathbb{E}(\epsilon_i) = 0$, and ρ is the spatial autoregressive parameter that quantifies the strength of spatial interdependence. Specifically, a positive value of ρ ($\rho > 0$) indicates positive spatial dependence, while a negative value ($\rho < 0$) suggests negative spatial dependence. If $\rho = 0$, the model reduces to the traditional linear regression model.

By introducing a spatially lagged variable, we explicitly account for spatial spillover effects, such as those arising from geographical proximity. This data-generating process is known as a *Spatial Autoregressive Process* (SAR) or, equivalently, the *Spatial Lag Model* (SLM). In the case where the model only includes the spatial lag and no explanatory variables, it is referred to as the pure Spatial Lag Model or SAR model.

Figure 2.1: The SLM for two tegions

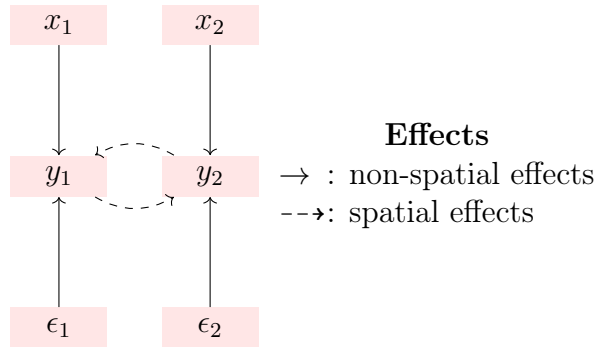


Figure 2.1 visually illustrates the spatial autoregressive model from Equation (2.1), focusing on two regions. In this representation, the variables (x_1, x_2) and the unobserved terms (ϵ_1, ϵ_2) directly influence the dependent variables y_1 and y_2 in their respective regions. The spatial spillover effects are captured through the influence of y_1 on y_2 and vice versa. This structure reflects *simultaneity* in the model, highlighting the presence of spatial autocorrelation and spillover effects.

It is important to have different notation forms for the SLM. In vector form, the model can be expressed as:

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

where \mathbf{w}_i is the i th row of \mathbf{W} . Additionally, a comprehensive Spatial Lag Model (SLM) specification, considering covariates in matrix form, takes the shape:

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

where \mathbf{y} is a $n \times 1$ vector containing 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.

2.1.2 Spatial Durbin Model

The Spatial Durbin Model (SDM) is defined as:

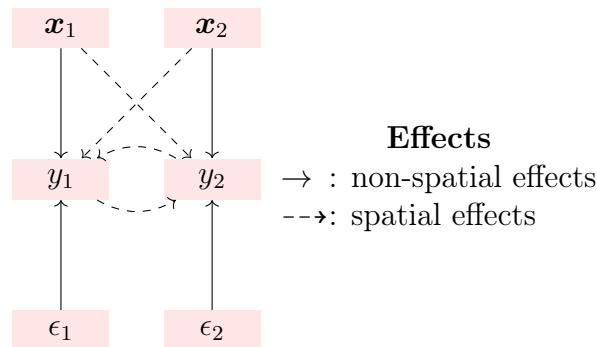
$$\mathbf{y} = \rho \mathbf{W} \mathbf{y} + \alpha \mathbf{1}_n + \mathbf{X} \boldsymbol{\beta} + \mathbf{W} \mathbf{X} \boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \quad (2.3)$$

where \mathbf{y} is the dependent variable, ρ is the spatial autoregressive parameter, \mathbf{W} is the spatial weights matrix, $\mathbf{1}_n$ is an n -dimensional vector of ones to account for the intercept α , \mathbf{X} is the matrix of explanatory variables, $\mathbf{W}\mathbf{X}$ represents the spatially lagged explanatory variables, and $\boldsymbol{\varepsilon}$ is the error term.

The SDM extends traditional spatial autoregressive models by including not only the spatially lagged dependent variable ($\mathbf{W}\mathbf{y}$) and explanatory variables (\mathbf{X}) but also the spatially lagged explanatory variables ($\mathbf{W}\mathbf{X}$). This means that \mathbf{y} depends on the local (own-region) factors from \mathbf{X} as well as the same factors averaged over neighboring regions, introducing an additional layer of spatial dependence.

This concept is illustrated in Figure 2.2. The diagram shows that Region 1 affects Region 2 not only through the spatially lagged dependent variable (\mathbf{y}) but also through the spatially lagged independent and exogenous variables (\mathbf{x}).

Figure 2.2: The SDM for Two Regions



Consider an example where \mathbf{y} represents air pollution levels in different regions. The term $\mathbf{W}\mathbf{y}$ implies that air pollution in Region 1 affects pollution in Region 2, and vice versa. If \mathbf{X} includes population density, the spatially lagged variable $\mathbf{W}\mathbf{X}$ captures the effect of population density in Region 1 (or Region 2) on air pollution in the neighboring region.

The SDM is particularly advantageous for calculating marginal effects, as it allows for the decomposition of direct, indirect, and total effects, which will be explored in later sections.

2.1.3 Spatial Error Model

Another form of spatial dependence occurs when it operates through the error process, meaning that the errors from different regions exhibit spatial autocorrelation. The Spatial Error Model (SEM) is formulated as:

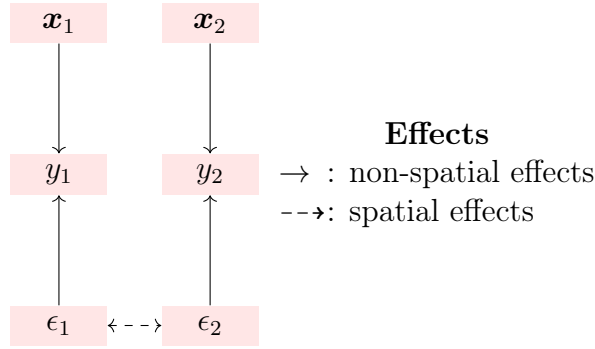
$$\begin{aligned} \mathbf{y} &= \alpha \mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \\ \mathbf{u} &= \lambda \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}. \end{aligned} \tag{2.4}$$

where λ is the autoregressive parameter capturing the spatial dependence in the error lag $\mathbf{W}\mathbf{u}$, $\boldsymbol{\varepsilon}$ is i.i.d. noise, and \mathbf{W} is the spatial weights matrix. The parameter λ distinguishes the spatial error dependence from the spatial autoregressive coefficient ρ in spatial lag models.

Figure 2.3 illustrates the SEM for two regions. The diagram highlights that the error terms (ϵ_1 and ϵ_2) of both regions are interconnected, and the spatial effect operates solely through this relationship.

As noted by [Anselin and Bera \(1998\)](#), spatial error dependence can be viewed as a nuisance, with λ acting as a nuisance parameter. This reflects spatial autocorrelation in mea-

Figure 2.3: The SEM for two regions



surement errors or in unaccounted variables that may influence the dependent variable. For instance, it may capture spillovers of omitted variables across spatial units.

Unlike models focusing on explicit spatial or social interaction processes, the SEM accommodates situations where the determinants of the dependent variable are spatially autocorrelated but not explicitly modeled. It can also represent scenarios where unobserved shocks exhibit a spatial pattern. This flexibility makes the SEM useful in dealing with spatially structured errors that arise from omitted variables or other latent spatial processes.

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

The Spatial Autocorrelation Model (SAC) is a more general framework that incorporates key features of both the spatial lag and spatial error models discussed earlier. Its structural representation is given by:

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

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

or more compactly in matrix form,

$$\mathbf{y} = \alpha \mathbf{1}_n + \rho \mathbf{W} \mathbf{y} + \mathbf{X} \boldsymbol{\beta} + \mathbf{u},$$

$$\mathbf{u} = \lambda \mathbf{M} \mathbf{u} + \boldsymbol{\varepsilon},$$

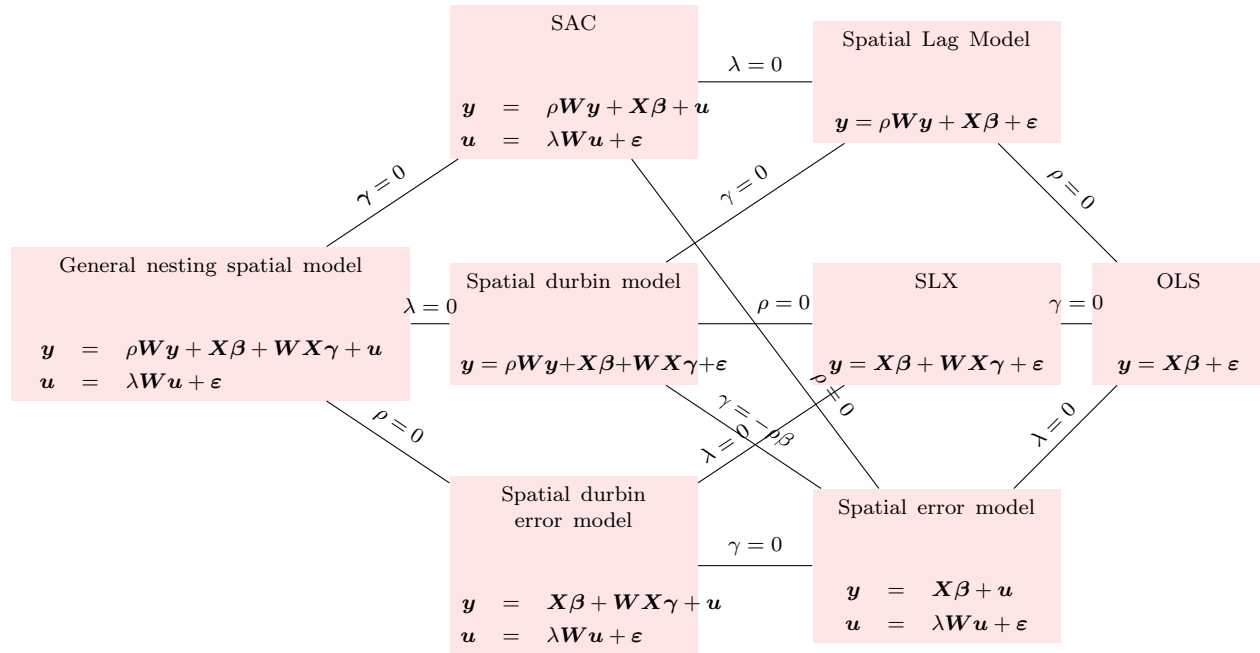
where the matrix \mathbf{W} and \mathbf{M} are $n \times n$ spatial-weighting matrices.¹ In this model, spatial interactions in the dependent variable and the disturbances are considered. As standard, the

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.

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.

A broader taxonomy of spatial models is shown in Figure 2.4. The most comprehensive model in this framework is the General Nesting Spatial Model (GNS or Manski's Model), which includes spatial dependence in the dependent variable, the explanatory variables, and the error term. Different restrictions on the GNS model yield various spatial model specifications, as detailed below.

Figure 2.4: Taxonomy of spatial models



Starting with the GNS model:

- Imposing the restriction $\gamma = \mathbf{0}$ 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 ($\gamma = -\rho\beta$) 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.

- Imposing the restriction $\gamma = 0$ leads to the spatial lag model (SLM), whereas the restriction $\rho = 0$ results in a least-squares spatially lagged \mathbf{X} regression model (labeled SLX) that assumes independence between regions in the dependent variable, but includes characteristics from neighboring regions in the form of spatially lagged explanatory variables.

Finally, if $\rho = \lambda = 0$ and $\boldsymbol{\theta} = 0$, then we obtain the traditional linear regression model.

2.2 Reduced Form and Parameter Space

An important distinction in the context of spatial models is the difference between structural and reduced form model. The reduced form expresses the endogenous variables as functions of the exogenous variables. In contrast, the structural form is the ‘behavioral model’ that defines the relationship between the variables in the system.

For example, Equation (2.2) represents the structural model for the SLM, where both exogenous and endogenous variables are related to the dependent variable \mathbf{y} . This form reflects the theoretical framework that captures the underlying relationship between the variables.

However, it is often more insightful to consider how the dependent variable is generated, which is referred to as the **data generating process** (DGP). The DGP represents the empirical process that gives rise to the observed data. By solving the structural model in Equation (2.2) for the endogenous variables, \mathbf{y} , yields the reduced-form model.

The implied DGP or “reduced form equation” for the SLM given in Equation (2.2) is:

$$\mathbf{y} = (\mathbf{I}_n - \rho \mathbf{W})^{-1} (\alpha \mathbf{z}_n + \mathbf{X} \boldsymbol{\beta}) + (\mathbf{I}_n - \rho \mathbf{W})^{-1} \boldsymbol{\varepsilon}, \quad (2.5)$$

which shows that any spatially lagged dependent variable is no longer present on the right-hand side. This equation illustrates 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, to identify the true coefficients from data, we need $(\mathbf{I}_n - \rho \mathbf{W})$ to be invertible. From linear algebra, a matrix \mathbf{A} is invertible if $\det(\mathbf{A}) \neq 0$. Thus, for the reduced form to be valid we require that $\det(\mathbf{I}_n - \rho \mathbf{W}) \neq 0$. The question then becomes: for which values of ρ lead to a non-singular $(\mathbf{I}_n - \rho \mathbf{W})$?

The following Lemma from [Kelejian and Prucha \(2010\)](#) establishes a bound for ρ :

Lemma 2.1 — Bounds for ρ ([Kelejian and Prucha, 2010](#)). Let τ denote the spectral radius of \mathbf{W} , defined as:

$$\tau = \max \{|\omega_1|, \dots, |\omega_n|\},$$

where $\omega_1, \dots, \omega_n$ denote the eigenvalues of \mathbf{W} . Then $\mathbf{I}_n - \rho \mathbf{W}$ is nonsingular for all ρ in

the interval $(-1/\tau, 1/\tau)$.

When ρ is restricted to the interval $(-1/\tau, 1/\tau)$, all eigenvalues of $\rho\mathbf{W}$ are guaranteed to have absolute values less than one. A related result, frequently used in the literature, is presented below:

Lemma 2.2 — 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:

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

As [Kelejian and Prucha \(2010\)](#) explain, Lemma 2.2 holds only when all eigenvalues of \mathbf{W} are real. However, non-symmetric matrices typically have complex eigenvalues, making Lemma 2.1 more general.

To facilitate interpretation, \mathbf{W} is often normalized so that each row sums to unity. This row-normalization ensures nonnegative weights between 0 and 1, which allows \mathbf{W} to be interpreted as a row-stochastic matrix that averages neighboring values. According to Theorem 1.1 (Eigenvalues of Row-Stochastic Matrix), the eigenvalues of such a row-stochastic matrix \mathbf{W} have absolute values less than or equal to one. This result, combined with Lemma 2.1, implies $\rho \in (1/\omega_{min}, 1)$.

However, interpreting ρ as a conventional correlation coefficient between \mathbf{y} and its spatial lag $\mathbf{W}\mathbf{y}$ can be misleading. The parameter space of ρ depends on the normalization of \mathbf{W} . For instance, standardization methods other than row-normalization may yield different bounds for ρ .

Lemma 2.3 — Invertibility of Row-Normalized \mathbf{W} matrix. If \mathbf{W} is row-normalized, then $(\mathbf{I}_n - \rho\mathbf{W})^{-1}$ exists for all $|\rho| < 1$

Despite its popularity, row-normalization has drawbacks. As discussed in Section 1.2.4, it alters the internal structure of \mathbf{W} , complicating inter-row comparisons. To address this, scalar normalization (multiplying \mathbf{W} by a constant a) can be applied. This approach preserves the relationships between rows and removes unit-of-measure effects. 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 $\mathbf{I}_n - \rho\mathbf{W}$ is nonsingular, it can be expressed as an infinite series, commonly referred to as the Leontief expansion:

Lemma 2.4 — Leontief Expansion. If $\mathbf{I} - \rho\mathbf{W}$ is nonsingular, then

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

Using Lemma 2.4 (Leontief Expansion), the reduced form for the SLM in Equation (2.5) can be written as:

$$\begin{aligned} \mathbf{y} &= (\mathbf{I}_n + \rho\mathbf{W} + \rho^2\mathbf{W}^2 + \dots) (\alpha\mathbf{z}_n + \mathbf{X}\beta) + (\mathbf{I}_n + \rho\mathbf{W} + \rho^2\mathbf{W}^2 + \dots) \boldsymbol{\varepsilon}, \\ &= \alpha\mathbf{z}_n + \rho\mathbf{W}\mathbf{z}_n\alpha + \rho^2\mathbf{W}^2\mathbf{z}_n\alpha + \dots + \mathbf{X}\beta + \rho\mathbf{W}\mathbf{X}\beta + \rho^2\mathbf{W}^2\mathbf{X}\beta + \dots \\ &\quad + \boldsymbol{\varepsilon} + \rho\mathbf{W}\boldsymbol{\varepsilon} + \rho^2\mathbf{W}^2\boldsymbol{\varepsilon}. \end{aligned} \quad (2.6)$$

Expression (2.6) 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 expansion reveals 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 2.4.

Considering the reduced form Equation (2.5), 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:

$$\begin{aligned}\mathbb{E}(\mathbf{y}|\mathbf{X}, \mathbf{W}) &= \mathbb{E}[(\mathbf{I}_n - \rho\mathbf{W})^{-1}(\alpha\mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}) + (\mathbf{I}_n - \rho\mathbf{W})^{-1}\boldsymbol{\varepsilon}|\mathbf{X}, \mathbf{W}] \\ &= (\mathbf{I}_n - \rho\mathbf{W})^{-1}(\alpha\mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}).\end{aligned}\tag{2.7}$$

From Equation (2.5), we derive the variance-covariance matrix of \mathbf{y} :

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

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:

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

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. ■

2.2.1 Eigenvalues in R

Computing eigenvalues is a crucial step in the estimation of spatial models, particularly when using Maximum Likelihood procedures. In this section, we demonstrate how to compute the eigenvalues of a spatial weight matrix \mathbf{W} in R, along with some of their properties.

First, we create an artificial spatial weight matrix using functions from the **spdep** package:

```
# Create a queen W matrix
library("spdep")
W1.queen <- cell2nb(3, 3, type = "queen")
summary(W1.queen)

## Neighbour list object:
## Number of regions: 9
## Number of nonzero links: 40
## Percentage nonzero weights: 49.38272
## Average number of links: 4.444444
## Link number distribution:
##
## 3 5 8
## 4 4 1
## 4 least connected regions:
## 1:1 3:1 1:3 3:3 with 3 links
## 1 most connected region:
## 2:2 with 8 links
```

The function `cell2nb` from **spdep** package generates a list of neighbors for a grid of cells. Here, we define a grid with 3 rows and 3 columns, resulting in a total of 9 regions. Since `type = "queen"` is specified, the neighbors include those sharing either an edge or a vertex, consistent with a regular grid spatial structure. The resulting object is of class `nb`, representing a list of neighbors.

To proceed with matrix computations, we convert this `nb` object into a spatial weight matrix of class `matrix`:

```
# From nb to matrix
W.queen <- nb2mat(W1.queen)
round(W.queen, 3)

##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
## 1:1 0.000 0.333 0.000 0.333 0.333 0.000 0.000 0.000 0.000
## 2:1 0.200 0.000 0.200 0.200 0.200 0.200 0.000 0.000 0.000
## 3:1 0.000 0.333 0.000 0.000 0.333 0.333 0.000 0.000 0.000
## 1:2 0.200 0.200 0.000 0.000 0.200 0.000 0.200 0.200 0.000
## 2:2 0.125 0.125 0.125 0.125 0.000 0.125 0.125 0.125 0.125
## 3:2 0.000 0.200 0.200 0.000 0.200 0.000 0.000 0.200 0.200
## 1:3 0.000 0.000 0.000 0.333 0.333 0.000 0.000 0.333 0.000
## 2:3 0.000 0.000 0.000 0.200 0.200 0.200 0.200 0.000 0.200
## 3:3 0.000 0.000 0.000 0.000 0.333 0.333 0.000 0.333 0.000
```

```
## attr("call")
## nb2mat(neighbours = W1.queen)
```

The resulting matrix is row-normalized, meaning that the sum of each row equals one. We can verify this as follows:

```
# Compute row-sums
rowSums(W.queen)

## 1:1 2:1 3:1 1:2 2:2 3:2 1:3 2:3 3:3
##  1  1  1  1  1  1  1  1  1
```

Additionally, note that this matrix is not symmetric. We can confirm this property using:

```
# Check whether the matrix is symmetric
is.sym <- all(W.queen == t(W.queen))
is.sym

## [1] FALSE
```

The `eigen` function in R computes the eigenvalues and eigenvectors of a matrix. It returns a named list with two components: `values` (the eigenvalues) and `vectors` (the eigenvectors). To illustrate, we compute the eigenvalues of the previously defined spatial weight matrix:

```
# Obtain eigenvalues
values <- eigen(W.queen, symmetric = is.sym)$values
values

## [1] 1.000000e+00 -4.527525e-01 -4.000000e-01 3.651484e-01 -3.651484e-01
## [6] 3.651484e-01 -3.651484e-01 -1.472475e-01 -2.259287e-16

range(values)

## [1] -0.4527525 1.0000000
```

The eigenvalues are returned in decreasing order. For the spatial weight matrix `W.queen`, all eigenvalues are real. The minimum eigenvalue is approximately -0.45, while the maximum eigenvalue is 1, corroborating the fact that if \mathbf{W} is row-normalized, then $|\omega| \leq 1$.

Since \mathbf{W} is row-normalized, and according to Lemma 2.3, $\mathbf{I}_n - \rho\mathbf{W}$ is invertible if $|\rho| < 1$. Note that if $\rho = 1$, then the matrix is nonsingular as shown in the following lines

```
solve(diag(9) - 1 * W.queen)

## Error in solve.default(diag(9) - 1 * W.queen): system is computationally singular:
## reciprocal condition number = 3.6909e-18
```

Next, we compute the eigenvalues for a rook spatial weight matrix:

```
W.rook <- nb2mat(cell2nb(3, 3, type = "rook"))
is.sym <- all(W.rook == t(W.rook))
values <- eigen(W.rook, symmetric = is.sym)$values
range(values)

## Error in min(x, na.rm = na.rm): invalid 'type' (complex) of argument
```

For the rook spatial weight matrix, the minimum eigenvalue is -1, and the maximum eigenvalue is 1. This highlights a difference in the spectral properties of rook versus queen spatial weight matrices.

Another approach to constructing spatial weight matrices, particularly for Monte Carlo studies, is the k -ahead and k -behind criterion in a circular world, as proposed by [Kelejian and Prucha \(1999\)](#). In this framework, each spatial unit is assumed to have k neighbors ahead of it and k neighbors behind it in the sample order. For example, we can create such a matrix with $k = 2$ using the `circular` function from the `sphet` package:

```
library("sphet")
W.cir <- nb2mat(circular(3, 3, 2))
W.cir

##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
## 1:1 0.00 0.25 0.25 0.00 0.00 0.00 0.00 0.25 0.25
## 2:1 0.25 0.00 0.25 0.25 0.00 0.00 0.00 0.00 0.25
## 3:1 0.25 0.25 0.00 0.25 0.25 0.00 0.00 0.00 0.00
## 1:2 0.00 0.25 0.25 0.00 0.25 0.25 0.00 0.00 0.00
## 2:2 0.00 0.00 0.25 0.25 0.00 0.25 0.25 0.00 0.00
## 3:2 0.00 0.00 0.00 0.25 0.25 0.00 0.25 0.25 0.00
## 1:3 0.00 0.00 0.00 0.00 0.25 0.25 0.00 0.25 0.25
## 2:3 0.25 0.00 0.00 0.00 0.00 0.25 0.25 0.00 0.25
## 3:3 0.25 0.25 0.00 0.00 0.00 0.00 0.25 0.25 0.00
## attr(,"call")
## nb2mat(neighbours = circular(3, 3, 2))

all(W.cir == t(W.cir))

## [1] TRUE

values <- eigen(W.cir, symmetric = all(W.cir == t(W.cir)))$values
values

## [1] 1.00000000 0.46984631 0.46984631 -0.08682409 -0.08682409 -0.38302222
## [7] -0.38302222 -0.50000000 -0.50000000

range(values)

## [1] -0.5 1.0
```


In this example, each spatial unit has $2k$ neighbors (two ahead and two behind). The eigenvalues for this matrix can provide insights into the connectivity and spatial structure defined by this criterion.

2.2.2 Generate spatial DGP in R

To construct a DGP of a spatial model, we first define the reduced form of the model. Consider the SLM with

$$\mathbf{y} = (\mathbf{I}_n - \rho \mathbf{W})^{-1} (\beta_0 \mathbf{1}_n + \beta_1 \mathbf{x} + \boldsymbol{\varepsilon}),$$

where $\mathbf{1}_n$ is an $n \times 1$ vector of ones and the elements of the vector \mathbf{x} are normally distributed with mean 0 and standard deviation 1. The parameter β_0 is set to 1, while $\beta_1 = 2$. The error term is normally distributed with mean zero and standard deviation one. The spatial autoregressive parameter is set to $\rho = 0.6$. The following lines create this DGP for a sample size of $n = 49$ using a circular spatial weight matrix:

```
# Set the random seed
set.seed(1)

# True parameters
b0 <- 1
b1 <- 2
rho <- 0.6

# Sample size and W matrix
n <- 49
W.cir <- nb2mat(circular(sqrt(n), sqrt(n), 3))

# Generate random variables
x <- rnorm(n, mean = 0, sd = 1)
epsilon <- rnorm(n, mean = 0, sd = 1)

# Generate dependent variable
y <- solve(diag(n) - rho * W.cir) %*% (b0 + b1 * x + epsilon)
```

2.3 Motivation of Spatial Models

2.3.1 SLM as a Long-run Equilibrium

A Spatial Lag Model (SLM) can be interpreted as a system of simultaneous dependencies over time that converges to a new steady-state equilibrium, even when using a cross-sectional dataset (LeSage and Pace, 2010). To demonstrate this, consider the vector of the dependent variable at time t , denoted by \mathbf{y}_t . Assume that this variable is determined by a spatial autoregressive process, where the current value depends on the spatially lagged values of the dependent variable from neighboring observations. This introduces a time lag in the average

values of the neighboring dependent variables observed during the previous period, $\mathbf{W}\mathbf{y}_{t-1}$. We can also include the current period's own-region characteristics \mathbf{X}_t in the model. If the characteristics of regions remain relatively fixed over time, we can express $\mathbf{X}_t = \mathbf{X}$.

For illustration, consider a model where pollution is the dependent variable \mathbf{y}_t , which depends on the past period's pollution values from neighboring regions, $\mathbf{W}\mathbf{y}_{t-1}$. The appropriate model in this case would be:

$$\mathbf{y}_t = \rho \mathbf{W} \mathbf{y}_{t-1} + \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}_t. \quad (2.8)$$

We can substitute \mathbf{y}_{t-1} from the right-hand side of equation (2.8) as:

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

resulting in:

$$\begin{aligned} \mathbf{y}_t &= \mathbf{X} \boldsymbol{\beta} + \rho \mathbf{W} (\mathbf{X} \boldsymbol{\beta} + \rho \mathbf{W} \mathbf{y}_{t-2} + \boldsymbol{\varepsilon}_{t-1}) + \boldsymbol{\varepsilon}_t \\ &= \mathbf{X} \boldsymbol{\beta} + \rho \mathbf{W} \mathbf{X} \boldsymbol{\beta} + \rho^2 \mathbf{W}^2 \mathbf{y}_{t-2} + \boldsymbol{\varepsilon}_t + \rho \mathbf{W} \boldsymbol{\varepsilon}_{t-1}. \end{aligned} \quad (2.9)$$

By recursively substituting past values of \mathbf{y}_{t-r} on the right-hand side of Equation (2.9) over q periods, we obtain:

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

The expected value of this spatial process is:

$$\mathbb{E}(\mathbf{y}_t) = (\mathbf{I}_n + \rho \mathbf{W} + \rho^2 \mathbf{W}^2 + \dots + \rho^{q-1} \mathbf{W}^{q-1}) \mathbf{X} \boldsymbol{\beta} + \rho^q \mathbf{W}^q \mathbf{y}_{t-q}, \quad (2.10)$$

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 equation (2.10) as $q \rightarrow \infty$, we obtain:

$$\lim_{q \rightarrow \infty} \mathbb{E}(\mathbf{y}_t) = (\mathbf{I}_n - \rho \mathbf{W})^{-1} \mathbf{X} \boldsymbol{\beta}. \quad (2.11)$$

Note that the magnitude of $\rho^q \mathbf{W}^q \mathbf{y}_{t-q}$ tends to zero as q increases, under the assumption that $|\rho| < 1$ and that \mathbf{W} is row-stochastic (i.e., it has a principal eigenvalue of 1).

Equation (2.11) indicates that we can interpret the observed cross-sectional relationship as the outcome or expectation of a long-run equilibrium or steady state. This provides a dynamic motivation for the data generating process (DGP) of the cross-sectional SLM, which is commonly used in spatial regression modeling. In other words, a cross-sectional SLM relationship can emerge from the time-dependent decisions of economic agents located at different points in space, where decisions are influenced by neighboring regions.

2.3.2 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:

$$\mathbf{z} = \rho \mathbf{W} \mathbf{z} + \mathbf{r}$$

$$\mathbf{z} = (\mathbf{I}_n - \rho \mathbf{W})^{-1} \mathbf{r}$$

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:

$$\begin{aligned} \mathbf{y} &= \mathbf{x}\beta + \mathbf{u} \\ \mathbf{u} &= (\mathbf{I}_n - \rho \mathbf{W})^{-1} \boldsymbol{\varepsilon} \end{aligned} \tag{2.12}$$

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

2.3.3 SDM and Omitted Variables Motivation

Now suppose that \mathbf{X} and $\boldsymbol{\varepsilon}$ from (2.12) are correlated, given by the following process:

$$\begin{aligned} \boldsymbol{\varepsilon} &= \mathbf{x}\gamma + \mathbf{v} \\ \mathbf{v} &\sim N(0, \sigma^2 \mathbf{I}_n) \end{aligned} \tag{2.13}$$

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 (2.13) into (2.12), we obtain:

$$\begin{aligned} \mathbf{y} &= \mathbf{x}\beta + (\mathbf{I}_n - \rho \mathbf{W})^{-1} \boldsymbol{\varepsilon} \\ &= \mathbf{x}\beta + (\mathbf{I}_n - \rho \mathbf{W})^{-1} (\mathbf{x}\gamma + \mathbf{v}) \\ &= \mathbf{x}\beta + (\mathbf{I}_n - \rho \mathbf{W})^{-1} \mathbf{x}\gamma + (\mathbf{I}_n - \rho \mathbf{W})^{-1} \mathbf{v} \\ (\mathbf{I}_n - \rho \mathbf{W}) \mathbf{y} &= (\mathbf{I}_n - \rho \mathbf{W}) \mathbf{x}\beta + \mathbf{v} \\ \mathbf{y} &= \rho \mathbf{W} \mathbf{y} + \mathbf{x}(\beta + \gamma) + \mathbf{W} \mathbf{x}(-\rho\beta) + \mathbf{v} \end{aligned} \tag{2.14}$$

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} .

2.4 Interpreting Spatial Models

2.4.1 Measuring Spillovers

A central concern in regional science is the measurement of spatial spillovers. In a spatial context, spillovers can be defined as the impacts that changes in one region have on neighboring regions (LeSage and Pace, 2014). Examples of spatial spillovers include:

- Changes in the tax rate in one region may influence tax rate decisions in neighboring regions, a phenomenon known as tax mimicking or yardstick competition among local governments.
- Home improvements made by one homeowner may increase the selling prices of nearby homes.
- Innovations by university researchers may diffuse to nearby firms.
- Air or water pollution generated in one region may spill over to neighboring regions.

The models discussed in the previous section can be used to formally define spatial spillovers and, more importantly, to estimate their quantitative magnitude and test their statistical significance. However, it is important to distinguish between global and local spillovers, a distinction explored in [Anselin \(2003\)](#) and [LeSage and Pace \(2014\)](#).

We begin by formally defining global spillovers:

Definition 2.4.1 — Global Spillovers. Global spillovers occur when changes in a characteristic of one region affect the outcomes of all other regions. This includes impacts on the region itself, as changes can propagate through neighboring regions and back to the original region (feedback). Specifically, global spillovers influence not only direct neighbors but also neighbors of neighbors, neighbors of neighbors of neighbors, and so on.

The endogenous interactions produced by global spillovers lead to a scenario where changes in one region trigger a sequence of adjustments across potentially all regions in the sample, ultimately resulting in a new long-run steady-state equilibrium ([LeSage, 2014](#)). As [LeSage \(2014\)](#) explains, global spillovers may arise from interactions between local policies. For example: *“It seems plausible that changes in the levels of public assistance (cigarette taxes) in state A would lead neighboring states (e.g., state B) to adjust their levels of assistance (taxes), which in turn triggers a feedback response from state A, and also responses from states C that are neighbors to state B, and so on.”*

The following definition describes local spillovers:

Definition 2.4.2 — Local Spillovers. Local spillovers occur when the impact is confined to nearby or immediate neighbors, with the effect diminishing before it reaches regions further removed, such as neighbors of neighbors.

As indicated by these definitions, the key difference between global and local spillovers is that feedback or endogenous interactions are only possible in the case of global spillovers.

2.4.2 Marginal Effects

Mathematically, the notion of spillover can be thought as the derivative $\partial y_i / \partial x_j$. This represents how changes in an explanatory variable in region i influence the dependent variable in another region $j \neq i$.

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

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

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 :

$$\begin{pmatrix} \mathbb{E}(y_1) \\ \mathbb{E}(y_2) \\ \vdots \\ \mathbb{E}(y_n) \end{pmatrix} = \sum_{r=1}^K \begin{pmatrix} \mathbf{S}_r(\mathbf{W})_{11} & \mathbf{S}_r(\mathbf{W})_{12} & \dots & \mathbf{S}_r(\mathbf{W})_{1n} \\ \mathbf{S}_r(\mathbf{W})_{21} & \mathbf{S}_r(\mathbf{W})_{22} & \dots & \mathbf{S}_r(\mathbf{W})_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_r(\mathbf{W})_{n1} & \mathbf{S}_r(\mathbf{W})_{n2} & \dots & \mathbf{S}_r(\mathbf{W})_{nn} \end{pmatrix} \begin{pmatrix} x_{1r} \\ x_{2r} \\ \vdots \\ x_{nr} \end{pmatrix}. \quad (2.15)$$

For the dependent variable for spatial unit i , Equation (2.15) would be:

$$\mathbb{E}(y_i) = \sum_{r=1}^k [\mathbf{S}_r(\mathbf{W})_{i1} x_{1r} + \mathbf{S}_r(\mathbf{W})_{i2} x_{2r} + \dots + \mathbf{S}_r(\mathbf{W})_{in} x_{nr}]. \quad (2.16)$$

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

$$\frac{\partial \mathbb{E}(y_i)}{\partial x_{jr}} = \mathbf{S}_r(\mathbf{W})_{ij} \quad (2.17)$$

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

$$\frac{\partial \mathbb{E}(y_i)}{\partial x_{ir}} = \mathbf{S}_r(\mathbf{W})_{ii}. \quad (2.18)$$

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:

$$\begin{aligned} \begin{pmatrix} \frac{\partial \mathbb{E}(\mathbf{y})}{\partial x_{1r}} & \frac{\partial \mathbb{E}(\mathbf{y})}{\partial x_{2r}} & \dots & \frac{\partial \mathbb{E}(\mathbf{y})}{\partial x_{nr}} \end{pmatrix}_{(n \times n)} &= \begin{pmatrix} \frac{\partial \mathbb{E}(y_1)}{\partial x_{1r}} & \frac{\partial \mathbb{E}(y_1)}{\partial x_{2r}} & \dots & \frac{\partial \mathbb{E}(y_1)}{\partial x_{nr}} \\ \frac{\partial \mathbb{E}(y_2)}{\partial x_{1r}} & \frac{\partial \mathbb{E}(y_2)}{\partial x_{2r}} & \dots & \frac{\partial \mathbb{E}(y_2)}{\partial x_{nr}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mathbb{E}(y_n)}{\partial x_{1r}} & \frac{\partial \mathbb{E}(y_n)}{\partial x_{2r}} & \dots & \frac{\partial \mathbb{E}(y_n)}{\partial x_{nr}} \end{pmatrix} \\ &= \mathbf{A}(\mathbf{W})^{-1} (\mathbf{I}_n \beta_r + \mathbf{W} \theta_r) = \mathbf{S}_r(\mathbf{W}) \\ &= (\mathbf{I}_n - \rho \mathbf{W})^{-1} \begin{pmatrix} \beta_r & w_{12} \theta_r & \dots & w_{1n} \theta_r \\ w_{21} \theta_r & \beta_r & \dots & w_{2n} \theta_r \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} \theta_r & w_{n2} \theta_r & \dots & \beta_r \end{pmatrix} \end{aligned} \quad (2.19)$$

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:

$$\mathbf{W} = \begin{pmatrix} 0 & 1 & 0 \\ w_{21} & 0 & w_{23} \\ 0 & 1 & 0 \end{pmatrix} \quad (2.20)$$

and

$$\mathbf{A}(\mathbf{W})^{-1} = \frac{1}{1 - \rho^2} \begin{pmatrix} 1 - w_{23}\rho^2 & \rho & \rho^2 w_{23} \\ \rho w_{21} & 1 & \rho w_{23} \\ \rho^2 w_{21} & \rho & 1 - w_{21}\rho^2 \end{pmatrix} \quad (2.21)$$

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 (2.20) and (2.21) into Equation (2.19) we get:

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

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 2.4.3 — 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

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

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.

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.

Definition 2.4.4 — 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:

$$\text{ATIT} = \frac{1}{n} \mathbf{1}_n' \mathbf{c}_r \quad (2.22)$$

Definition 2.4.5 — 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 $\mathbf{r}_r = \mathbf{1}_n' \mathbf{S}_r(\mathbf{W})$, so an average of these total impacts is:

$$\text{ATIF} = \frac{1}{n} \mathbf{r}_r \mathbf{1}_n \quad (2.23)$$

The definition 2.4.5 relates how changes in a single observation j influences all observations. In contrast, definition 2.4.4 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:

$$\bar{M}(r)_{\text{direct}} = n^{-1} \text{tr}(\mathbf{S}_r(\mathbf{W})) \quad (2.24)$$

$$\bar{M}(r)_{\text{total}} = n^{-1} \mathbf{1}_n' \mathbf{S}_r(\mathbf{W}) \mathbf{1}_n \quad (2.25)$$

$$\bar{M}(r)_{\text{indirect}} = \bar{M}(r)_{\text{total}} - \bar{M}(r)_{\text{direct}} \quad (2.26)$$

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

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

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 a Spatial Durbin Model (SDM) specification to examine changes in the (logged) number of workers in U.S. census tracts with commuting times exceeding 45 minutes one way, between 1990 and 2000. (See also the related discussion in Section 2.5.) This investigation is motivated by the observation that the percentage of U.S. workers with these long commute times increased from 12.5% in 1990 to 15.4% in 2000, representing a rise of over 10%.

When selecting the appropriate model, the authors identify two key dynamics:

- **Global spillover impacts:** Congestion effects from increased long-distance commuters in one part of a metropolitan area likely affect travel times across the entire roadway network.

- **Feedback effects:** Congestion from commuting decisions in one tract spills over into neighboring tracts, creating reciprocal feedback effects that further influence congestion in the originating tract.

These observations led the authors to specify the following SDM:

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

where:

- \mathbf{y} : Logged number of workers with long commute times.
- \mathbf{X} : Variables capturing household location decisions, including age, gender, income distribution, and geographical characteristics of the tract.
- $\mathbf{W} \mathbf{X}$: The corresponding variables for neighboring tracts, incorporating spatial interactions.
- ρ : Spatial autoregressive parameter capturing dependence in the dependent variable.
- \mathbf{W} : Spatial weight matrix reflecting neighborhood relationships.

Using this model, the authors compare **direct**, **indirect**, and **total effects** estimates for the years 1990 and 2000. Their analysis highlights that demographic factors, particularly age and gender distributions, explain much of the variation in long commute times during this period.

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 2.4.4).

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

$$\begin{aligned}
 n^{-1} \mathbf{z}^\top \mathbf{S}_r(\mathbf{W}) \mathbf{z} &= n^{-1} \mathbf{z}^\top [\mathbf{A}(\mathbf{W})^{-1} (\mathbf{I} \beta_r)] \mathbf{z} \\
 &= n^{-1} \mathbf{z}^\top [(\mathbf{I}_n - \rho \mathbf{W})^{-1}] (\mathbf{I} \beta_r) \mathbf{z} \\
 &= n^{-1} \mathbf{z}^\top [\mathbf{I}_n + \rho \mathbf{W} + \rho^2 \mathbf{W}^2 + \dots] (\mathbf{I} \beta_r) \mathbf{z} \quad \text{using Lemma 2.4} \\
 &= n^{-1} \mathbf{z}^\top [\mathbf{I}_n \beta_r + \rho \mathbf{W} \beta_r + \rho^2 \mathbf{W}^2 \beta_r + \dots] \mathbf{z} \\
 &= n^{-1} \mathbf{z}^\top [\mathbf{I}_n \mathbf{z} \beta_r + \rho \mathbf{W} \mathbf{z} \beta_r + \rho^2 \mathbf{W} (\mathbf{W} \mathbf{z}) \beta_r + \rho^3 \mathbf{W} \mathbf{W} (\mathbf{W} \mathbf{z})] \\
 &= n^{-1} \mathbf{z}^\top [\beta_r \mathbf{z} + \rho \beta_r \mathbf{z} + \rho^2 \beta_r \mathbf{z} + \rho^3 \beta_r \mathbf{z} + \dots] \quad \because \mathbf{W}^l \mathbf{z} = \mathbf{z} \\
 &= n^{-1} \mathbf{z}^\top [\beta_r + \rho \beta_r + \rho^2 \beta_r + \rho^3 \beta_r + \dots] \mathbf{z} \\
 &= n^{-1} [\beta_r + \rho \beta_r + \rho^2 \beta_r + \rho^3 + \dots] \mathbf{z}^\top \mathbf{z} \\
 &= n^{-1} [\beta_r + \rho \beta_r + \rho^2 \beta_r + \rho^3 \beta_r + \dots] n \\
 &= \frac{\beta_r}{(1 - \rho)}
 \end{aligned} \tag{2.27}$$

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:

$$\begin{aligned}
 \epsilon_{\text{SO}_2} &= \left(\frac{\text{SO}_2}{p} \right) \left(\frac{dp}{d\text{SO}_2} \right) \\
 &= \left(\frac{\text{SO}_2}{p} \right) \left(\frac{\beta_r}{(1 - \rho)} \cdot p \right) \quad \text{since the model is log-lin} \\
 &= \frac{\beta_r}{(1 - \rho)} \cdot \text{SO}_2
 \end{aligned} \tag{2.28}$$

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.

2.4.3 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 2.3.1). 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 2.4, 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 2.4:

$$\begin{pmatrix} \frac{\partial \mathbb{E}(\mathbf{y})}{\partial x_{1r}} & \frac{\partial \mathbb{E}(\mathbf{y})}{\partial x_{2r}} & \cdots & \frac{\partial \mathbb{E}(\mathbf{y})}{\partial x_{nr}} \end{pmatrix} \approx (\mathbf{I}_n + \rho \mathbf{W} + \rho^2 \mathbf{W}^2 + \rho^3 \mathbf{W}^3 + \dots + \rho^l \mathbf{W}^l) \mathbf{I}_n \beta_r \quad (2.29)$$

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

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:

$$\hat{\mathbf{y}}^{(2)} = (\mathbf{I}_n - \hat{\rho}\mathbf{W})^{-1} \widetilde{\mathbf{X}}\hat{\boldsymbol{\beta}} \quad (2.30)$$

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 $\widetilde{\mathbf{X}}$ 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 (2.18), 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 2.17. 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 2.4.4 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 2.4.5 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 (2.27) 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 2.4.3.

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 Empirical Example: Growth and Spatial Externalities

Ertur and Koch (2007) developed a theoretical growth model that explicitly incorporates technological interdependence among economies and examines the role of spatial spillover effects.

The model is based on an aggregate Cobb-Douglas production function for country i at time t , exhibiting constant returns to scale in labor and physical capital:

$$Y_i(t) = A_i(t)K_i^\alpha(t)L_i^{1-\alpha}(t), \quad (2.31)$$

where $Y_i(t)$ denotes total output, $K_i(t)$ is the stock of reproducible physical capital, $L_i(t)$ is the labor force, and $A_i(t)$ represents the aggregate level of technology.

The evolution of $A_i(t)$ is modeled as:

$$A_i(t) = \Omega(t)k_i^\phi(t) \prod_{j \neq i} A_j^{\gamma w_{ij}}(t), \quad (2.32)$$

where $k_i(t) = K_i(t)/L_i(t)$ is the capital intensity, and w_{ij} denotes the weight capturing the connectivity between country i and country j in the spatial network.

This formulation reflects three distinct sources of technological progress:

- (a) Exogenous Technological Progress: The term $\Omega(t)$ captures the common component of technological change across all countries, assumed to grow at a constant rate μ , i.e., $\Omega(t) = \Omega(0)e^{\mu t}$.
- (b) Domestic Knowledge Spillovers: The second term, $k_i^\phi(t)$, reflects the idea that technology in a country improves with its own capital deepening. The parameter $\phi \in [0, 1)$ measures the strength of internal (or home) externalities from physical capital accumulation.
- (c) International Knowledge Spillovers: The final term captures cross-country technological interdependence. It is modeled as a spatially weighted geometric mean of the technological levels of neighboring countries, where $\gamma \in [0, 1)$ measures the strength of international spillovers. While γ is assumed to be common across countries, the impact of spatial externalities on a specific country i depends on its relative position and connectivity in the spatial structure, as encoded by the weights w_{ij} . The more a given country i is connected to its neighbors, the higher w_{ij} is, and the more country i benefits from spatial externalities.

In logarithmic and matrix form, Equation (2.32) becomes:

$$\mathbf{a} = \boldsymbol{\Omega} + \phi \mathbf{k} + \gamma \mathbf{W} \mathbf{a},$$

where \mathbf{a} is the $n \times 1$ vector of the logarithms of technology levels, \mathbf{k} is the $n \times 1$ vector of the logarithms of capital per worker, and \mathbf{W} is the $n \times n$ spatial weight matrix with elements w_{ij} . Solving for \mathbf{a} —assuming $\gamma \neq 0$ and $1/\gamma$ is not an eigenvalue of \mathbf{W} —yields:

$$\mathbf{a} = (\mathbf{I} - \gamma \mathbf{W})^{-1} \boldsymbol{\Omega} + \phi (\mathbf{I} - \gamma \mathbf{W})^{-1} \mathbf{k}.$$

If $|\gamma| < 1$, then Equation (2.32) can be rewritten as:

$$A_i(t) = \Omega^{\frac{1}{1-\gamma}} k_i^\phi \prod_{j=1}^n k_j^{\phi \sum_{r=1}^{\infty} \gamma^r w_{ij}^{(r)}}(t), \quad (2.33)$$

where $w_{ij}^{(r)}$ denotes the r -th order spatial weight between countries i and j .

Substituting (2.33) into the per-worker form of the production function (2.31) gives:

$$y_i(t) = \Omega^{\frac{1}{1-\gamma}} k_i^{u_{ii}}(t) \prod_{j \neq i}^n k_j^{u_{ij}}(t), \quad (2.34)$$

where $u_{ii} = \alpha + \phi \left(1 + \sum_{r=1}^{\infty} \gamma^r w_{ii}^{(r)}\right)$ and $u_{ij} = \phi \sum_{r=1}^{\infty} \gamma^r w_{ij}^{(r)}$, and $y_i(t) = Y_i(t)/L_i(t)$.

This expression implies spatial heterogeneity in the production function parameters. Specifically, if there are no externalities from physical capital (i.e., $\phi = 0$), then $u_{ii} = \alpha$ and $u_{ij} = 0$, and the model collapses to the standard neoclassical production function.

The model also allows for an analysis of the *social elasticity* of income per worker with respect to physical capital. According to Equation (2.34), when country i increases its own capital per worker, it benefits from a marginal social return of u_{ii} . If instead all countries simultaneously raise their capital per worker, the total return for country i increases to:

$$u_{ii} + \sum_{j \neq i} u_{ij} = \alpha + \frac{\phi}{1-\gamma}.$$

To ensure local convergence and avoid explosive or endogenous growth dynamics, it is assumed that social returns are decreasing; that is, $\alpha + \frac{\phi}{1-\gamma} < 1$.

Assume that a constant fraction s_i of output is saved, labor grows exogenously at rate n_i , and the depreciation rate of capital δ is identical across countries. Then, the dynamics of capital per worker follow the Solow equation:

$$\dot{k}_i(t) = s_i y_i(t) - (n_i + \delta) k_i(t),$$

where the dot denotes a time derivative.

Under these assumptions, [Ertur and Koch \(2007\)](#) derive the steady-state income per worker in country i as:

$$\begin{aligned} \log y_{it}^* &= \frac{1}{1-\alpha-\phi} \log \Omega_t + \frac{\alpha+\phi}{1-\alpha-\phi} \log s_i - \frac{\alpha+\phi}{1-\alpha-\phi} \log(n_i + g + \delta) \\ &\quad - \frac{\alpha\gamma}{1-\alpha-\phi} \sum_{j \neq i}^N w_{ij} \ln s_j + \frac{\alpha\gamma}{1-\alpha-\phi} \sum_{j \neq i}^N w_{ij} \ln(n_j + g + \delta) \\ &\quad + \frac{\gamma(1-\alpha)}{1-\alpha-\phi} \sum_{j \neq i}^N w_{ij} \log y_{jt}^* \end{aligned}$$

This steady-state equation is operationalized econometrically as the following SDM (for $t = 0$):

$$\log \left(\frac{Y_i}{L_i} \right) = \beta_0 + \beta_1 \log s_i + \beta_2 \log(n_i + g + \delta) + \theta_1 \sum_{j \neq i}^N w_{ij} \ln s_j + \theta_2 \sum_{j \neq i}^N w_{ij} \ln(n_j + g + \delta) + \rho \sum_{j \neq i}^N w_{ij} \log \left(\frac{Y_j}{L_j} \right) + \epsilon_i,$$

where:

Mauricio Sarrias

- $\beta_0 + \epsilon_i = \frac{1}{1-\alpha-\phi} \log \Omega_0$,
- $\beta_1 = \frac{\alpha+\phi}{1-\alpha-\phi}$,
- $\beta_2 = -\frac{\alpha+\phi}{1-\alpha-\phi}$,
- $\theta_1 = -\frac{\alpha\gamma}{1-\alpha-\phi}$,
- $\theta_2 = \frac{\alpha\gamma}{1-\alpha-\phi}$,
- $\rho = \frac{\gamma(1-\alpha)}{1-\alpha-\phi}$

with the following theoretical constraints:

- $\beta_1 = -\beta_2$,
- $\theta_2 = -\theta_1$

Ertur and Koch (2007) argue that the textbook Solow model is misspecified because it omits spatially correlated variables related to technological interdependence and physical capital externalities. In particular, the error term from the Solow model can be written as:

$$\varepsilon_{\text{Solow}} = \frac{\phi}{1-\alpha} (\mathbf{I} - \gamma \mathbf{W})^{-1} k^* + (\mathbf{I} - \gamma \mathbf{W})^{-1} \varepsilon.$$

Hence, even if $\phi = 0$ (i.e., no domestic capital externalities), omitting spatially correlated variables induces spatial autocorrelation in the residuals, leading to model misspecification.

To test this hypothesis empirically, we can estimate the following standard Solow model using OLS:

$$\log \left(\frac{Y_i}{L_i} \right) = \beta_0 + \beta_1 \ln(s_i) + \beta_2 \ln(n_i + g + \delta) + \epsilon_i,$$

and then use the Moran's I test on the residuals.

2.7 Exercises

Exercise 2.1 Assume three regions with row-normalized spatial weight matrix given in Equation (2.20). Derive the total, direct and indirect effects for the following models:

(a) Spatial Durbin Model given by:

$$\mathbf{y} = \rho \mathbf{W} \mathbf{y} + \alpha \mathbf{1}_n + \mathbf{X} \boldsymbol{\beta} + \mathbf{W} \mathbf{X} \boldsymbol{\theta} + \boldsymbol{\varepsilon} \quad (2.35)$$

(b) Spatial Lag Model given by:

$$\mathbf{y} = \rho \mathbf{W} \mathbf{y} + \alpha \mathbf{1}_n + \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2.36)$$

(c) Spatial Durbin Error Model given by:

$$\mathbf{y} = \alpha \mathbf{1}_n + \mathbf{X} \boldsymbol{\beta} + \mathbf{W} \mathbf{X} \boldsymbol{\theta} + \mathbf{u} \quad (2.37)$$

$$\mathbf{u} = \lambda \mathbf{W} \mathbf{u} + \boldsymbol{\varepsilon} \quad (2.38)$$

(d) OLS given by:

$$\mathbf{y} = \alpha \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2.39)$$

(e) Spatial Error model given by:

$$\begin{aligned} \mathbf{y} &= \alpha \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \\ \mathbf{u} &= \lambda \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon} \end{aligned} \quad (2.40)$$

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:

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

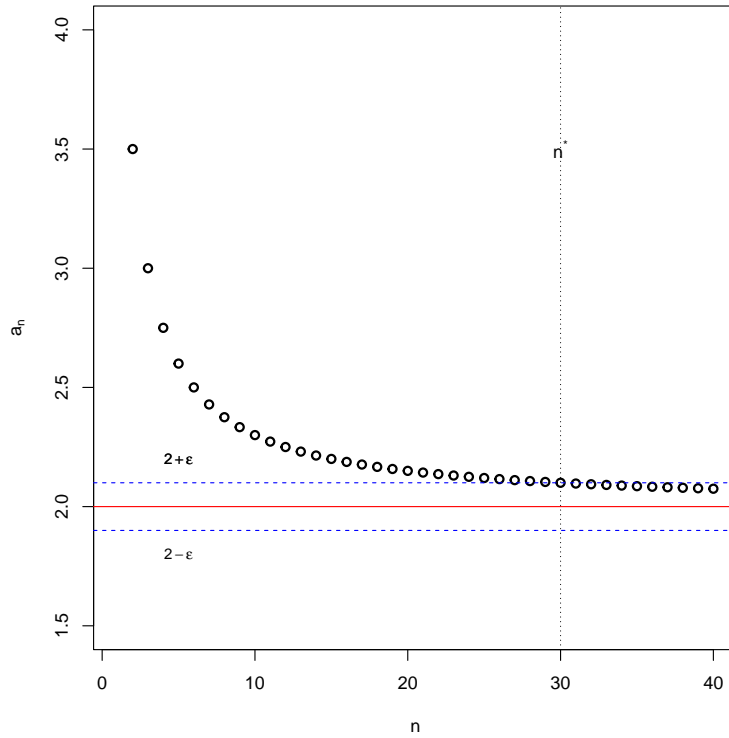
In Definition 3.1.1 by choosing a very small ϵ , we ensure that b_n gets arbitrarily close to its limit b for all n that is sufficiently large. In fact, the smaller ϵ is, the larger $n(\epsilon)$ will be. So, ϵ can be interpreted as a prespecified tolerance level for the discrepancy between b_n and b . 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 non-singular 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.

R 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

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

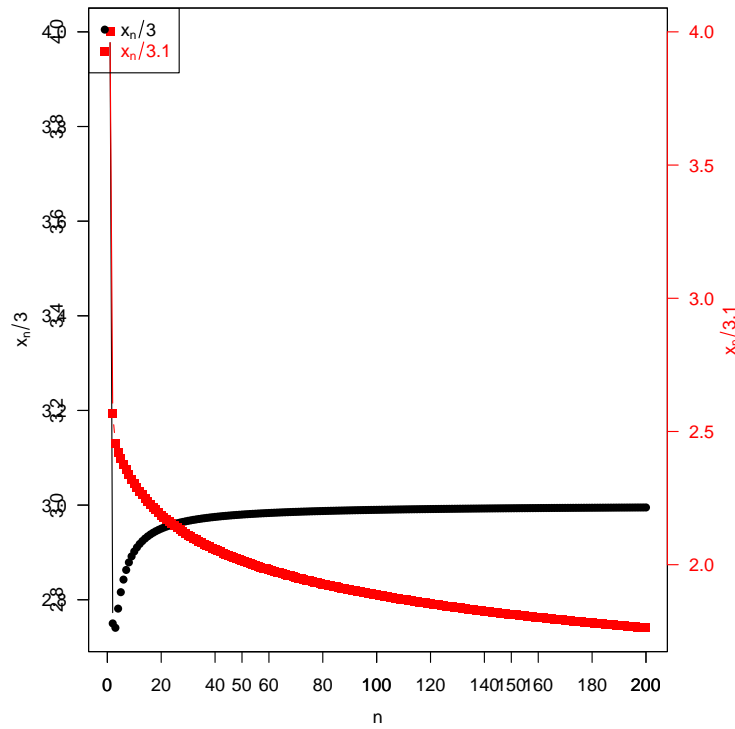
Then the vector sequence $\{\mathbf{x}_n\}$ is $o(1)$ and $O(1)$, since

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

■

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

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,

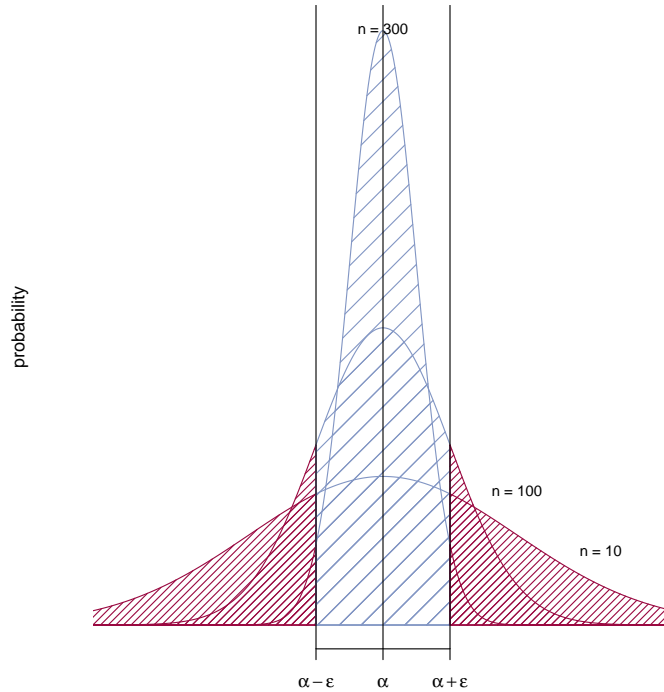
$$X_n \xrightarrow{p} \alpha \quad \text{is the same as} \quad X_n - \alpha \xrightarrow{p} 0$$

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

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

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 $\mathbb{E}(\mathbf{x} \mathbf{x}')$ exists for all n if

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

A special case of convergence in quadratic mean occurs when \mathbf{x} , 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:

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

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.B.4 we obtain:

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

If $\lim_{n \rightarrow \infty} \mathbb{V}(\hat{\theta}_n) = 0$, then $\Pr \left[\left| \hat{\theta}_n - \theta \right| \geq \delta \right] \rightarrow 0$, 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 $n \rightarrow \infty$.

R 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:

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

or equivalently,

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

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:

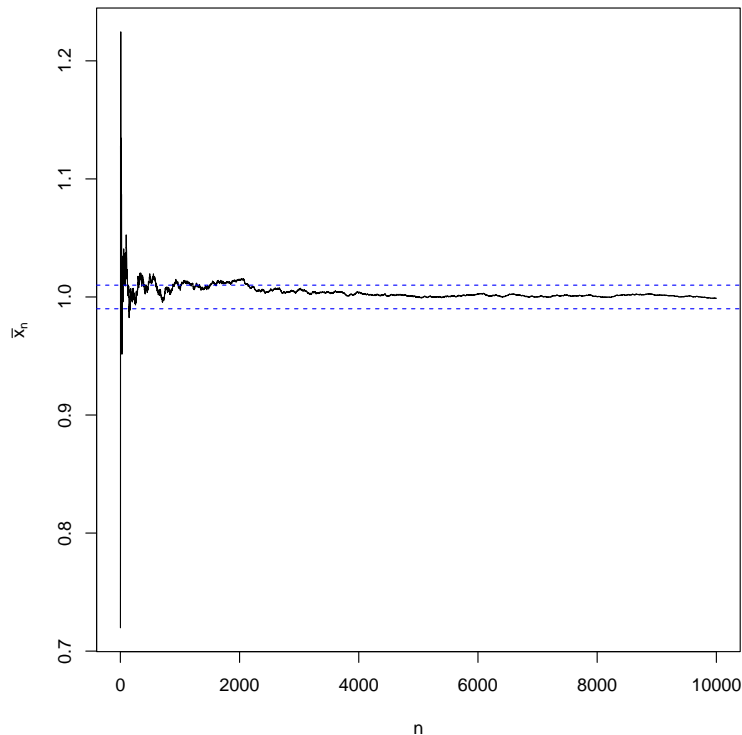
$$\begin{aligned}\bar{X}_1 &= g(X_1) = X_1 \\ \bar{X}_2 &= g(X_1, X_2) = \frac{1}{2} \sum_{i=1}^2 X_i \\ \bar{X}_3 &= g(X_1, X_2, X_3) = \frac{1}{3} \sum_{i=1}^3 X_i \\ &\vdots \\ \bar{X}_n &= g(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n X_i\end{aligned}$$

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.

■

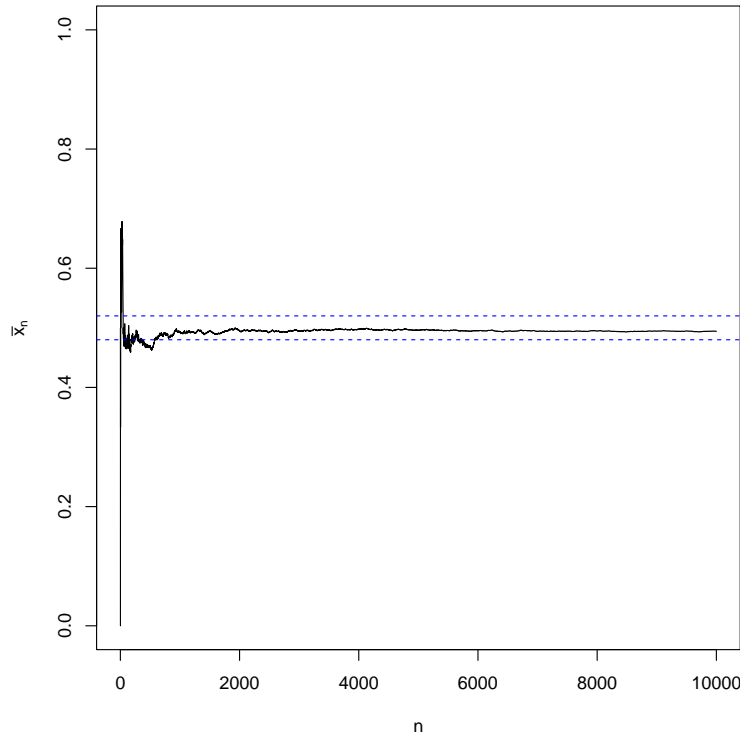
■ **Example 3.14 — Tossing a fair coin.** Now we simulate $n = 1000$ coin tosses. After each simulated toss, we plot the proportion X_n of heads obtained so far against the number n of tosses so far. The LLN says we should see a trace that gets very close to $1/2$ as n increases.

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 $\{(\mathbf{Z}_i^\top, \mathbf{X}_i, \boldsymbol{\epsilon}_i)\}$ is an i.i.d random sequence, then $\{\mathbf{X}_i \mathbf{X}_i^\top\}$, $\{\mathbf{X}_i \boldsymbol{\epsilon}_i\}$, $\{\mathbf{Z}_i \mathbf{X}_i^\top\}$, $\{\mathbf{Z}_i \boldsymbol{\epsilon}_i\}$, 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 $(\mathbf{X}_i, \boldsymbol{\epsilon}_i)$ 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 $\hat{\boldsymbol{\beta}}_n \xrightarrow{a.s.} \boldsymbol{\beta}_0$. Assume the following:

(a) $y_i = \mathbf{x}_i^\top \boldsymbol{\beta}_0 + \epsilon_i$, $i = 1, \dots, n$; $\boldsymbol{\beta}_0 \in \mathbb{R}^k$;

- (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.B.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.B.6:

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

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

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

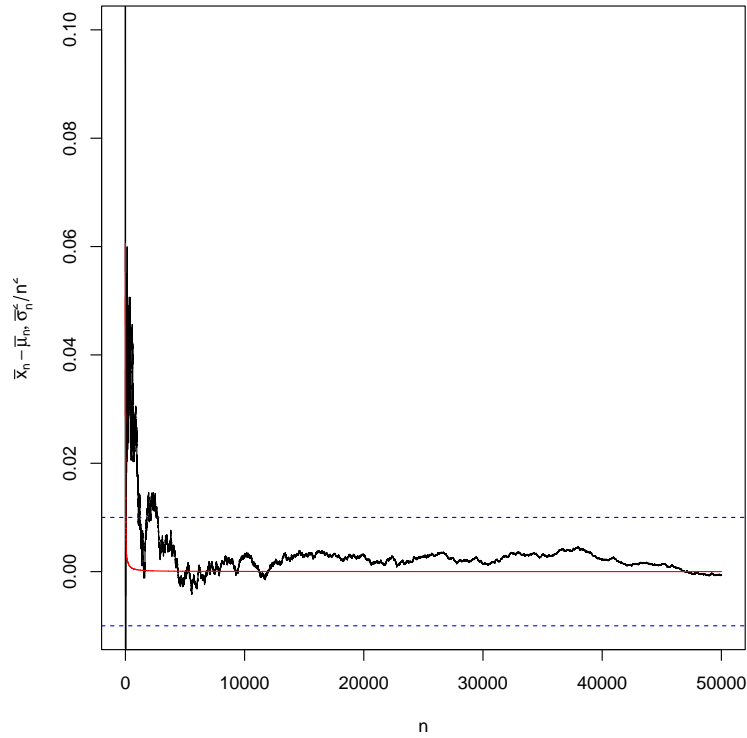
The Chebychev's theorem does not state that \bar{X}_n converges to $\bar{\mu}_n$, or even that it converges to a constant at all. The theorem states that as n increases without bound, these two quantities will be arbitrarily close to each other. In other words, the difference between them converges to 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 if 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 $\mathbf{Y}_n \xrightarrow{(2 \times 1)} \mathbf{Y} \sim N(\mathbf{0}, \mathbf{I})$. 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:

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

That is, the limiting distribution of $\hat{\boldsymbol{\theta}}_n$ is a spike (the asymptotic distribution of $\hat{\boldsymbol{\theta}}_j$ 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 FGSL.

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 .



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 $\mathbb{E} \left[|X_i - \mu_i|^{2+\delta} \right]$ if 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 unboundedness 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:

$$\Pr [n^{-\lambda} |X_n| \geq n_0] \leq \epsilon \quad (3.19)$$

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 ∞ .

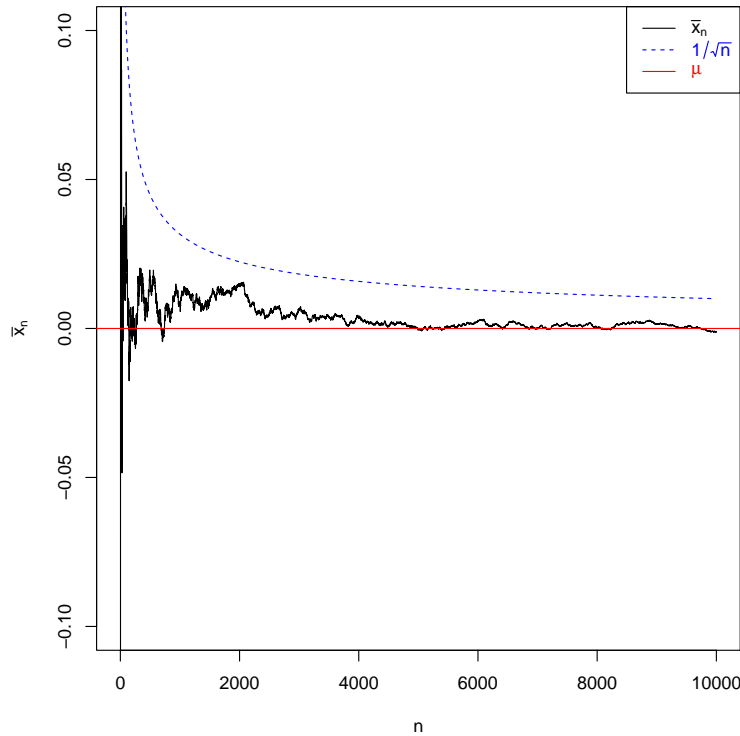
Note also that:

$$\begin{aligned}\sqrt{n}(\hat{\beta}_n - \beta_0) &= n^{1/2}O_p(n^{-1})O_p(n^{1/2}) \\ &= n^{1/2}O_p(n^{-1/2}) \\ &= O_p(1) \quad \because X_n = O_p(1/\sqrt{n}) \implies X_n/(1/\sqrt{n}) = \sqrt{n}X_n = O_p(1)\end{aligned}$$

■

■ **Example 3.21 — Rate of convergence for sample mean.** Consider and 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 graphs 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 does $n \rightarrow \infty$ mean in a spatial context?* Does it imply an increase in the geographical area, or does it refer to an increase in the number of spatial units within a given geographical area?

For spatial data, two distinct asymptotic frameworks have been studied: **increasing domain** and **infill asymptotic**. Increasing domain refers to a sampling structure where new observations (spatial units) are added at the edges (boundary points), similar to the

underlying asymptotics in time series analysis. In other words, increasing domain asymptotics involve more and more observations being sampled over an expanding domain. The key issue here is defining what constitutes the boundary.

In contrast, infill asymptotics are appropriate when the spatial domain is bounded, and new observations (points) are added between existing ones, resulting in a denser surface. In most applications of spatial econometrics, the underlying structure aligns more with the increasing domain framework.

The increasing domain framework necessitates an understanding of **triangular arrays**. The following section provides a straightforward 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 $\{X_{ni}\}$.

Central limit theorems (CLTs) applied to triangular arrays of random variables focus on the limiting distributions of appropriately defined functions of row averages. Consider the row average $S_n = n^{-1} \sum_{i=1}^n X_{ni}$ for the n -th row of the triangular array. For instance, when $n = 3$ (the third row), we have $S_3 = \frac{1}{3}(X_{31} + X_{32} + X_{33})$.

Traditional CLTs address averages of the form $n^{-1} \sum_{i=1}^n X_i$, where X_i are elements of a sequence $\{X_n\}$. However, triangular arrays, denoted by $\{X_{ni}\}$, are more general structures. The random variables in a row need not be identical to those in other rows, making triangular arrays more flexible for modeling.

This flexibility introduces certain statistical challenges, particularly regarding the appropriate CLT to apply. In triangular arrays, the random variables across rows may exhibit different distributions or dependencies, which complicates the derivation of limiting distributions. As a result, both the Law of Large Numbers (LLN) and the CLT for triangular arrays require slightly stronger conditions than those for independent and identically distributed (i.i.d.) sequences of random variables.

In order for S_n to converge to a normal distribution as $n \rightarrow \infty$, the following conditions must be met:

- 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 $n \rightarrow \infty$.

You may be wondering: Why are triangular arrays important in the spatial context? When adopting the increasing domain approach, it becomes evident that as n increases, the spatial weight matrix \mathbf{W} changes as observations are added. To illustrate, consider the true parameter vector $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^\top, \rho_0, \sigma_0^2)^\top$, where variables and estimates depend on the sample size n . This dependency allows us to study their behavior as $n \rightarrow \infty$.

Let $\mathbf{A}_n(\rho) = \mathbf{I}_n - \rho \mathbf{W}_n$ for any value of ρ . The “equilibrium” vector is given by

$$\mathbf{y}_n = \mathbf{A}_n^{-1}(\mathbf{X}_n \boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}_n), \quad (3.22)$$

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 $(\mathbf{I}_n - \rho \mathbf{W}_n)^{-1}$ generally depend on the sample size n , the vectors \mathbf{y}_n and $\boldsymbol{\varepsilon}_n$ will also depend upon n , forming a **triangular arrays**. This arises because new spatial units—or data points—alter the structure of the existing spatial units within the weight matrix \mathbf{W}_n (see for example [Kelejian and Prucha, 1999, 2001](#); [Anselin, 2007](#)). For example, the outcome for the first spatial unit, $y_{1,n}$, will differ depending on whether the total number of observations is $n = 10$ or $n = 15$, due to the evolving structure of \mathbf{W}_n as n changes.

This dependence implies that the elements of \mathbf{y} should be indexed by n :

$$\mathbf{y}_n = (y_{11}, y_{21}, y_{22}, \dots, y_{nn}).$$

For example, for $n = 1, 2, 3$, we have:

$$\begin{aligned} n = 1 &\implies y_{11} \\ n = 2 &\implies y_{12} \ y_{22} \\ n = 3 &\implies y_{13} \ y_{23} \ y_{33} \\ &\vdots \\ n = n &\implies y_{13} \ y_{23} \ y_{33} \dots y_{3n} \end{aligned}$$

where $y_{11} \neq y_{12} \neq y_{13}$ and $y_{22} \neq y_{23}$. Note that the dependent variables within the same row are mutually independent (spatial units are independent) and have the same distribution. However, the distribution of the dependent variables across different rows can vary.

The triangular array structure of \mathbf{y}_n partially arises from allowing the disturbances in the model to have a triangular array structure. More fundamentally, this structure also applies to the regressors in \mathbf{X} . By allowing the elements of \mathbf{X}_n to depend on n , we explicitly account for spatial lags among some of the regressors.

In regularly observed time series settings, indices correspond to equidistant points on the real line, making the notion of increasing n straightforward. In contrast, spatial settings involve ambiguity. For example, consider n points on a two-dimensional regularly spaced lattice, where both the number of rows (n_1) and columns (n_2) increase with $n = n_1 \cdot n_2$. Listing these points in lexicographic order (e.g., left to right, top to bottom) requires re-labeling as n increases, which the triangular array structure accommodates ([Anselin, 2021](#)).

Another consequence of this listing is that dependence between spatial locations i and j is not necessarily a function of the difference $i - j$, particularly when the dependence is isotropic.

3.8 Bounded Matrices and Useful Lemmas for Spatial Econometrics

In asymptotic it is important to establish whether the matrices are bounded as $n \rightarrow \infty$. As explained before, this implies that the impact of spatial interactions does not grow unbounded as the sample size increases. For example, if \mathbf{W}_n is not bounded, the effect of spatial dependence could explode, making the process unstable. Similarly, for asymptotic results such as the LLN and CLTs to hold in spatial context, regularity conditions on the spatial weight matrix.

In theoretical literature, we often encounter the following definition of bounded matrices.

Definition 3.8.1 — 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.

Note that for the typical row-standardized \mathbf{W} , row-boundedness is guaranteed by construction since the row sum is 1. The column boundedness is in most cases also guaranteed as long as the number of neighbors is constrained. We also need $(\mathbf{I}_n - \rho \mathbf{W}_n)^{-1}$ to be uniformly bounded in row and column sum and consider also the parameter space of ρ .

An important characteristic of uniformly bounded matrices is that this property is preserved under matrix multiplication.

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 , suppose that $\sum_{i=1}^n |a_{ij}| \leq c_a$, $\sum_{j=1}^n |a_{ij}| \leq c_a$, $\sum_{i=1}^n |b_{ij}| \leq c_b$, and $\sum_{j=1}^n |b_{ij}| \leq c_b$. Let $\mathbf{D} = \mathbf{AB}$, then

$$d_{ij} = \sum_{r=1}^n a_{ir} b_{rj}.$$

Let r_i be the i th row sum, then

$$\begin{aligned}
 r_i &= \sum_{j=1}^n |d_{ij}|, \\
 &= \sum_{j=1}^n \left| \sum_{r=1}^n a_{ir} b_{rj} \right|, \\
 &\leq \sum_{j=1}^n \sum_{r=1}^n |a_{ir} b_{rj}| \text{ by triangle inequality 3.B.2,} \\
 &= \sum_{j=1}^n \sum_{r=1}^n |a_{ir}| |b_{rj}| \text{ by multiplicativity 3.B.1,} \\
 &= \sum_{r=1}^n \sum_{j=1}^n |a_{ir}| |b_{rj}| \text{ by property of summation,} \\
 &= \sum_{r=1}^n |a_{ir}| \sum_{j=1}^n |b_{rj}|, \\
 &\leq c_a c_b, \text{ for all } i = 1, \dots, n \text{ and } n \geq 1 \text{ by Def. 3.8.1.}
 \end{aligned}$$

Similarly, we can show that

$$\sum_{i=1}^n |d_{ij}| \leq c_a c_b, \text{ for all } j = 1, \dots, n \text{ and } n \geq 1.$$

■

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

$$\begin{aligned}
 \delta_{ij} &= \sum_{r=1}^n \sum_{s=1}^n Z_{si} a_{sr} Z_{rj}, \\
 |\delta_{ij}| &= \left| \sum_{r=1}^n \sum_{s=1}^n Z_{si} a_{sr} Z_{rj} \right|, \\
 |\delta_{ij}| &\leq \sum_{r=1}^n \sum_{s=1}^n |Z_{si}| |a_{sr}| |Z_{rj}| \text{ by 3.B.2,} \\
 &\leq c_z^2 \sum_{r=1}^n \sum_{s=1}^n |a_{sr}| \\
 &\leq c_z^2 \sum_{r=1}^n c_a \\
 &\leq c_z^2 c_a n \\
 &= O(n)
 \end{aligned}$$

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

$$\begin{aligned}
 \text{tr}(\mathbf{A}_n \mathbf{A}_n^\top) &= \sum_{i=1}^n \sum_{j=1}^n a_{ij}^2 \\
 |\text{tr}(\mathbf{A}_n \mathbf{A}_n^\top)| &= \left| \sum_{i=1}^n \sum_{j=1}^n a_{ij}^2 \right| \\
 &\leq \sum_{i=1}^n \sum_{j=1}^n |a_{ij}^2| \\
 &\leq \sum_{i=1}^n \left(\sum_{j=1}^n |a_{ij}| \right)^2 \\
 &\leq nc_1^2 \quad \text{if } \mathbf{A}_n \text{ is uniformly bounded in row sums} \\
 |\text{tr}(\mathbf{A}_n \mathbf{A}_n^\top)| &\leq \sum_{j=1}^n \left(\sum_{i=1}^n |a_{ij}| \right)^2 \\
 &\leq nc_2^2 \quad \text{if } \mathbf{A}_n \text{ is uniformly bounded in column sums}
 \end{aligned}$$

3.9 Linear and Quadratic Forms

3.9.1 Moments

Linear and quadratic forms of the error terms will often appear in the analysis of spatial models. Thus, it is important to review and analyze some of its characteristics.

First, we define a quadratic form:

Definition 3.9.1 — Quadratic form. For a $n \times n$ symmetric matrix $\mathbf{A}_n = [a_{n,ij}]$ the quadratic function of n variables $\boldsymbol{\varepsilon}$ defined by:

$$\boldsymbol{\varepsilon}^\top \mathbf{A}_n \boldsymbol{\varepsilon} = \sum_{i=1}^n \sum_{j=1}^n a_{ij} \varepsilon_i \varepsilon_j$$

is called the quadratic form with matrix \mathbf{A}_n . If \mathbf{A}_n is not symmetric, we can replace \mathbf{A}_n by $\mathbf{A}^s = (\mathbf{A}_n + \mathbf{A}_n^\top)/2$.

The following Lemma is based on (Lee, 2004).

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

$$(a) \quad \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) = \sigma_0^2 \text{tr}(\mathbf{A}_n),$$

$$(b) \quad \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n)^2 = (\mu_4 - 3\sigma_0^4) \sum_{i=1}^n a_{ii}^2 + \sigma_0^4 [\text{tr}^2(\mathbf{A}_n) + \text{tr}(\mathbf{A}_n \mathbf{A}_n^\top) + \text{tr}(\mathbf{A}_n^2)], \text{ and}$$

$$(c) \quad \mathbb{V}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) = (\mu_4 - 3\sigma_0^4) \sum_{i=1}^n a_{ii}^2 + \sigma_0^4 [\text{tr}(\mathbf{A}_n \mathbf{A}_n^\top) + \text{tr}(\mathbf{A}_n^2)].$$

For the moment assume that \mathbf{A}_n 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}_n is not symmetric, then:

$$\text{Cov}(\boldsymbol{\varepsilon}^\top \mathbf{A}_1 \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}^\top \mathbf{A}_2 \boldsymbol{\varepsilon}) = 2 \text{tr} \left[\frac{1}{2} (\mathbf{A}_1 + \mathbf{A}_1^\top) \boldsymbol{\Sigma} \frac{1}{2} (\mathbf{A}_2 + \mathbf{A}_2^\top) \boldsymbol{\Sigma} \right] + 4 \boldsymbol{\mu}^\top \frac{1}{2} (\mathbf{A}_1 + \mathbf{A}_1^\top) \boldsymbol{\Sigma} \frac{1}{2} (\mathbf{A}_2 + \mathbf{A}_2^\top) \boldsymbol{\mu} \quad (3.23)$$

In particular, if $\boldsymbol{\varepsilon}$'s are normally distributed with mean 0 and variance σ_0^2 , then

- $\mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n)^2 = \sigma_0^4 [\text{tr}^2(\mathbf{A}_n) + \text{tr}(\mathbf{A}_n \mathbf{A}_n^\top) + \text{tr}(\mathbf{A}_n^2)] = \sigma_0^4 [\text{tr}^2(\mathbf{A}_n) + \text{tr}(\mathbf{A}_n \mathbf{A}_n^s)],$ and
- $\mathbb{V}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) = \sigma_0^4 \text{tr}(\mathbf{A}_n \mathbf{A}_n^s) = \frac{\sigma_0^4}{2} \text{tr}(\mathbf{A}_n^s) = \frac{\sigma_0^4}{2} \text{diag}(\mathbf{A}_n^s)' \text{diag}(\mathbf{A}_n^s).$

Proof. For (a), and using the Definition of quadratic form 3.9.1 We can write:

$$\begin{aligned}
 \mathbb{E}(\boldsymbol{\varepsilon}^\top \mathbf{A} \boldsymbol{\varepsilon}) &= \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}(\epsilon_i a_{ij} \epsilon_j) \\
 &= \sum_{i=1}^n \sum_{j=1}^n a_{ij} \mathbb{E}(\epsilon_i \epsilon_j) \\
 &= \sum_{i=1}^n \sum_{j=i}^n a_{ij} \sigma^2 \text{ since } \mathbb{E}(\epsilon_i \epsilon_j) = 0 \text{ for } i \neq j \\
 &= \sigma^2 \sum_{i=1}^n a_{ii} \\
 &= \sigma^2 \text{tr}(\mathbf{A}).
 \end{aligned}$$

We provide the proof for (b). Using the Definition of quadratic form 3.9.1 We can write:

$$\begin{aligned}
 \mathbb{E}(\boldsymbol{\varepsilon}^\top \mathbf{A}_n \boldsymbol{\varepsilon})^2 &= \mathbb{E} \left(\sum_{i=1}^n \sum_{j=1}^n a_{ij} \epsilon_i \epsilon_j \right)^2, \\
 &= \mathbb{E} \left(\sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \sum_{l=1}^n a_{ij} a_{kl} \epsilon_i \epsilon_j \epsilon_k \epsilon_l \right), \\
 &= \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \sum_{l=1}^n a_{ij} a_{kl} \mathbb{E}(\epsilon_i \epsilon_j \epsilon_k \epsilon_l).
 \end{aligned}$$

Because the error terms are i.i.d with zero mean, $\mathbb{E}(\epsilon_i \epsilon_j \epsilon_k \epsilon_l) \neq 0$ only when $i = j = k = l$, $(i = j) \neq (k = l)$, $(i = k) \neq (j \neq l)$, and $(i \neq l) \neq (j = k)$. Thus

$$\begin{aligned}
 \mathbb{E}(\boldsymbol{\varepsilon}^\top \mathbf{A}_n \boldsymbol{\varepsilon})^2 &= \sum_{i=1}^n a_{ii}^2 \mathbb{E}(\epsilon_i^4) + \sum_{i=1}^n \sum_{j \neq i}^n a_{ii} a_{jj} \mathbb{E}(\epsilon_i^2 \epsilon_j^2) \\
 &\quad + \sum_{i=1}^n \sum_{j \neq i}^n a_{ij}^2 \mathbb{E}(\epsilon_i^2 \epsilon_j^2) + \sum_{i=1}^n \sum_{j \neq i}^n a_{ij} a_{ji} \mathbb{E}(\epsilon_i^2 \epsilon_j^2), \\
 &= \mu_4 \sum_{i=1}^n a_{ii}^2 + \sigma^4 \sum_{i=1}^n \sum_{j \neq i}^n a_{ii} a_{jj} + \sigma^4 \sum_{i=1}^n \sum_{j \neq i}^n a_{ij}^2 + \sigma^4 \sum_{i=1}^n \sum_{j \neq i}^n a_{ij} a_{ji}, \\
 &= (\mu_4 - 3\sigma^4) \sum_{i=1}^n a_{ii}^2 + \sigma^4 \left[\sum_{i=1}^n \sum_{j=1}^n a_{ii} a_{jj} + \sum_{i=1}^n \sum_{j=1}^n a_{ij}^2 + \sum_{i=1}^n \sum_{j=1}^n a_{ij} a_{ji} \right], \\
 &= (\mu_4 - 3\sigma^4) \sum_{i=1}^n a_{ii}^2 + \sigma^4 [\text{tr}^2(\mathbf{A}) + \text{tr}(\mathbf{A} \mathbf{A}^\top) + \text{tr}(\mathbf{A}^2)], \\
 &= (\mu_4 - 3\sigma^4) \sum_{i=1}^n a_{ii}^2 + \sigma^4 [\text{tr}^2(\mathbf{A}) + \text{tr}(\mathbf{A} \mathbf{A}^s)].
 \end{aligned}$$

For (c), note that

$$\begin{aligned}
 \mathbb{V}(\boldsymbol{\varepsilon}^\top \mathbf{A} \boldsymbol{\varepsilon}) &= \mathbb{E}(\boldsymbol{\varepsilon}^\top \mathbf{A}_n \boldsymbol{\varepsilon})^2 - [\mathbb{E}(\boldsymbol{\varepsilon}^\top \mathbf{A} \boldsymbol{\varepsilon})]^2, \\
 &= (\mu_4 - 3\sigma^4) \sum_{i=1}^n a_{ii}^2 + \sigma^4 [\text{tr}^2(\mathbf{A}) + \text{tr}(\mathbf{A} \mathbf{A}^s)] - [\sigma^2 \text{tr}(\mathbf{A})]^2, \\
 &= (\mu_4 - 3\sigma^4) \sum_{i=1}^n a_{ii}^2 + \sigma^4 \text{tr}(\mathbf{A} \mathbf{A}^s), \\
 &= (\mu_4 - 3\sigma^4) \sum_{i=1}^n a_{ii}^2 + \frac{\sigma^2}{2} \text{tr}(\mathbf{A}^s).
 \end{aligned}$$

When $\boldsymbol{\varepsilon}$'s are normally distributed, $\mu_4 = 3\sigma^2$. ■

The following Lemma provides the moments for the product of moment conditions.

Lemma 3.26 For any n -dimensional (column) vector \mathbf{q}_n and $n \times n$ square matrices $\mathbf{A}_n = [a_{n,ij}]$ and $\mathbf{B}_n = [b_{n,ij}]$. Suppose that ϵ_{ni} 's in $\boldsymbol{\varepsilon}_n$ are i.i.d $(0, \sigma^2)$ and have finite third and fourth moments μ_3 and μ_4 . Then

- (a) $\mathbb{E}(\mathbf{q}_n^\top \boldsymbol{\varepsilon}_n \cdot \boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) = \mathbf{q}_n^\top \text{diag}(\mathbf{A}_n) \mu_3$,
- (b) $\mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n \cdot \boldsymbol{\varepsilon}_n^\top \mathbf{B}_n \boldsymbol{\varepsilon}_n) = (\mu_4 - 3\sigma^4) \text{diag}(\mathbf{A}_n)^\top \text{diag}(\mathbf{B}_n) + \sigma^4 [\text{tr}(\mathbf{A}_n) \text{tr}(\mathbf{B}_n) + \text{tr}(\mathbf{A}_n \mathbf{B}_n^s)]$,
- (c) and

$$\begin{aligned}
 \text{Cov}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n \cdot \boldsymbol{\varepsilon}_n^\top \mathbf{B}_n \boldsymbol{\varepsilon}_n) &= (\mu_4 - 3\sigma^4) \text{diag}(\mathbf{A}_n)^\top \text{diag}(\mathbf{B}_n) + \sigma^4 \text{tr}(\mathbf{A}_n \mathbf{B}_n^s) \\
 &= (\mu_4 - 3\sigma^4) \text{diag}(\mathbf{A}_n)^\top \text{diag}(\mathbf{B}_n) + \frac{\sigma^4}{2} \text{tr}(\mathbf{A}_n^s \mathbf{B}_n^s).
 \end{aligned}$$

Proof. As previously noted, $\mathbb{E}(\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_j) = \mathbf{0}$ whenever $i \neq j$, and $\mathbb{E}(\boldsymbol{\varepsilon}_i^2) = \mu_3 \mathbf{e}_i$ where \mathbf{e}_i is the n th unit vector with 1 at its i th place and zero elsewhere, we have:

$$\begin{aligned}
 \mathbb{E}(\mathbf{q}^\top \boldsymbol{\varepsilon} \cdot \boldsymbol{\varepsilon}^\top \mathbf{A} \boldsymbol{\varepsilon}) &= \mathbf{q}^\top \sum_{i=1}^n \sum_{j=1}^n a_{ij} \mathbb{E}(\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_j), \\
 &= \mu_3 \mathbf{q}^\top \begin{pmatrix} a_{11} \\ a_{22} \\ \vdots \\ a_{nn} \end{pmatrix}, \\
 &= \mu_3 \mathbf{q}^\top \text{diag}(\mathbf{A}).
 \end{aligned}$$

Since:

$$\boldsymbol{\varepsilon}^\top \mathbf{A}_n \boldsymbol{\varepsilon} \cdot \boldsymbol{\varepsilon}_n^\top \mathbf{B}_n \boldsymbol{\varepsilon}_n = \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \sum_{l=1}^n a_{ij} b_{kl} \epsilon_i \epsilon_j \epsilon_k \epsilon_l,$$

the mutual independence implies that $\mathbb{E}(\epsilon_i \epsilon_j \epsilon_k \epsilon_l) \neq 0$ only if $i = j = k = l$, $(i = j) \neq (k = l)$,

$(i = k) \neq (j \neq l)$, and $(i \neq l) \neq (j = k)$. It follows that

$$\begin{aligned} \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n \cdot \boldsymbol{\varepsilon}_n^\top \mathbf{B}_n \boldsymbol{\varepsilon}_n) &= \sum_{i=1}^n a_{ii} b_{ii} \mathbb{E}(\epsilon_i^4) + \sum_{i=1}^n \sum_{j \neq i}^n (a_{ii} b_{jj} + a_{ij} b_{ij} + a_{ij} b_{ji}) \mathbb{E}(\epsilon_i^2 \epsilon_j^2), \\ &= (\mu_4 - 3\sigma^4) \sum_{i=1}^n a_{ii} b_{ii} + \sigma^4 \sum_{i=1}^n \sum_{j=1}^n (a_{ii} b_{jj} + a_{ij} b_{ij} + a_{ij} b_{ji}) \mathbb{E}(\epsilon_i^2 \epsilon_j^2) \\ &= (\mu_4 - 3\sigma^4) \text{diag}(\mathbf{A})' \text{diag}(\mathbf{B}) + \sigma^4 [\text{tr}(\mathbf{A}) \text{tr}(\mathbf{B}) + \text{tr}(\mathbf{A} \mathbf{B}^\top) + \text{tr}(\mathbf{A} \mathbf{B})]. \end{aligned}$$

■

3.9.2 Law of Large Numbers

The following Lemma provides the asymptotic behavior for quadratic forms under homoskedasticity (see for example Lemma 2 in [Kelejian and Prucha, 1999](#)).

Lemma 3.27 — Consistency of quadratic forms in spatial models. Suppose that $\{\mathbf{A}_n\}$ is uniformly bounded in either row and column sums, and $\boldsymbol{\varepsilon}_n$ is i.i.d with $\mathbb{E}(\epsilon_i) = 0$ and $\mathbb{V}(\epsilon_i) = \sigma_0^2$ for all $i = 1, \dots, n$. 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) $\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n = O_p(n)$,
- (d) $\frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n - \frac{1}{n} \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) = o_p(1)$.

This Lemma states that both $\mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n)$ and $\mathbb{V}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n)$ are both well behaved sequence of nonrandom variables. It further states that $\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n$ is a well behaved (bounded) random sequence such that

$$\frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n - \frac{1}{n} \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) \xrightarrow{p} 0.$$

Thus, 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. 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

$$\mathbb{V}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) = (\mu_4 - 3\sigma_0^4) \sum_{i=1}^n a_{ii}^2 + \sigma_0^4 [\text{tr}(\mathbf{A}_n \mathbf{A}_n^\top) + \text{tr}(\mathbf{A}_n^2)].$$

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

$$\begin{aligned} \mathbb{V}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) &= (\mu_4 - 3\sigma_0^4)O(n) + \sigma_0^4 [O(n) + O(n)] \\ &= O(n) \end{aligned}$$

Proof of (d) and (c). By Chebyshev inequality 3.B.4

$$\begin{aligned} \Pr \left[\frac{1}{n} \left| \boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n - \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) \right| \geq \delta \right] &\leq \frac{\mathbb{V}(\frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n)}{\delta^2}, \\ \Pr \left[\left| \frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n - \sigma_0^2 \frac{\text{tr}(\mathbf{A}_n)}{n} \right| \geq \delta \right] &\leq \frac{\mathbb{V}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n)}{n^2 \delta^2}, \\ &\leq \frac{1}{\delta^2} \frac{1}{n^2} \mathbb{V}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n). \end{aligned}$$

Therefore, $n^{-2} \mathbb{V}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) = n^{-2} O(n) = O(1/n) = o(1)$.² 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

$$\frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n - \frac{1}{n} \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) = o_p(1).$$

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.B.4 implies that

$$\Pr \left(\frac{1}{n} \left| \boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n \right| \geq \delta \right) \leq \frac{1}{\delta^2 n^2} \mathbb{E}[(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n)^2] = \frac{1}{\delta^2} O(1).$$

Then $\frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n = O_p(1)$, and $n^{-1} \boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n$ is bounded. Furthermore, $\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n = O_p(n)$. ■

Lemma 3.28 Suppose that $\{\mathbf{A}_n\}$ consists of nonstochastic $n \times n$ matrices uniformly bounded in spectral matrix norm. Let \mathbf{c}_n be a column vector of constants. If, $\frac{1}{n} \mathbf{c}_n^\top \mathbf{c}_n$, then $\frac{1}{\sqrt{n}} \mathbf{c}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n = o_p(1)$. On the other hand, if $\frac{1}{n} \mathbf{c}_n^\top \mathbf{c}_n = O(1)$, then $\frac{1}{\sqrt{n}} \mathbf{c}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n = O_p(1)$.

3.10 CLT for Spatial Models

Theorem 3.29 — Liapounov's CLT theorem for double arrays. Let $\{X_{ni} : i = 1, \dots, n\}$ be an array of independent random variables. Let $\mathbb{E}(X_{ni}) = \mu_n$ and $\mathbb{E}(X_{ni} - \mu_n)^2 = \sigma_{ni}^2 \neq 0$. If

$$\frac{\sum_{k=1}^n \mathbb{E}|X_{nk} - \mu_{nk}|^2}{(\sum_{i=1}^n \sigma_{ni}^2)^{1+\delta/2}} \rightarrow 0,$$

² $O(n^{-1})$ implies $o(1)$. If $a_n = O(n^{-1})$ means that there exists a constant $c > 0$ such that for all sufficiently large n , $|a_n| \leq c/n$. $a_n = o(1)$ means that $\lim_{n \rightarrow \infty} a_n = 0$. Since $a_n = O(n^{-1})$ implies that $|a_n| \leq c/n$ for large n , it follows that $\lim_{n \rightarrow \infty} |a_n| \leq \lim_{n \rightarrow \infty} c/n = 0$. Thus, $a_n \rightarrow 0$ as $n \rightarrow \infty$, which is exactly the definition of $o(1)$.

for a positive $\delta > 0$, then

$$\sqrt{n} \frac{\sum_{i=1}^n (X_{ni} - \mu_{ni})}{(\sum_{i=1}^n \sigma_{ni}^2)^{1/2}} \xrightarrow{d} N(0, 1).$$

The following theorem states the limiting distribution for triangular arrays with homoskedastic errors in linear forms:

Theorem 3.30 — CLT for triangular arrays with homoskedastic 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

$$\frac{1}{\sqrt{n}} \mathbf{A}_n^\top \mathbf{v}_n \xrightarrow{d} N(\mathbf{0}, \sigma^2 \mathbf{Q}_{AA})$$

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).

Theorem 3.31 — CLT for Vectors of Linear Quadratic Forms with Heteroskedastic 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:

$$Q_{rn} = \boldsymbol{\varepsilon}' \mathbf{A}_r \boldsymbol{\varepsilon} + \mathbf{a}_r' \boldsymbol{\varepsilon} = \sum_{i=1}^n \sum_{j=1}^n a_{ijr} \epsilon_i \epsilon_j + \sum_{i=1}^n a_{ir} \epsilon_i.$$

Let $\mu_v = \mathbb{E}[\mathbf{v}_n] = [\mu_{Q_1}, \dots, \mu_{Q_2}]^\top$ 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:

$$\begin{aligned}\mu_{Q_r} &= \sum_{i=1}^n a_{iir} \sigma_i^2 \\ \sigma_{Q_{rs}} &= 2 \sum_{i=1}^n \sum_{j=1}^n a_{ijr} a_{ijs} \sigma_i^2 \sigma_j^2 + \sum_{i=1}^n a_{ir} a_{is} \sigma_i^2 \\ &\quad + \sum_{i=1}^n a_{iir} a_{iis} \left[\mu_i^{(4)} - 3\mu_i^4 \right] + \sum_{i=1}^n (a_{ir} a_{iis} + a_{is} a_{iir}) \mu_i^{(3)}\end{aligned}$$

with $\mu_i^{(3)} = \mathbb{E}(\epsilon_i^3)$ and $\mu_i^{(4)} = \mathbb{E}(\epsilon_i^4)$. Furthermore, given that $n^{-1} \lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{v}_n}) \geq c$ for some $c > 0$, then

$$\boldsymbol{\Sigma}_{\mathbf{v}_n}^{-1/2}(\mathbf{v}_n - \mu_{\mathbf{v}_n}) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_m),$$

and thus:

$$n^{-1/2}(\mathbf{v}_n - \mu_{\mathbf{v}_n}) \overset{a}{\sim} N(\mathbf{0}, n^{-1} \boldsymbol{\Sigma}_{\mathbf{v}_n})$$

In case that $\mathbf{A}_{r,n}$ is not symmetric, it can be replaced by $(\mathbf{A}_{r,n} + \mathbf{A}_{r,n}^\top)/2$. See our Definition 3.9.1. Also note that the matrix $\mathbf{A}_{r,n}$ is allowed to have non-zero diagonal elements. However, for most spatial models this is not needed, since, by convention, the diagonals of \mathbf{W} and \mathbf{M} are zero. If the diagonal elements of $\mathbf{A}_{r,n}$ are zero, then the mean $\mu_{Q_r} = \sum_{i=1}^n a_{iir} \sigma_i^2 = 0$. Similarly, if the ϵ_i are homoskedastic, then $\text{tr}(\mathbf{A}_r) = \sum_{i=1}^n a_{iir} = 0$ and, as a result, $\mu_{Q_r} = 0$.

In addition, the covariance $\sigma_{Q_{rs}}$ between Q_{rn} and Q_{sn} can be written more compactly as

$$2 \text{tr}(\mathbf{A}_r \boldsymbol{\Sigma} \mathbf{A}_s \boldsymbol{\Sigma}) + \mathbf{a}_r^\top \boldsymbol{\Sigma} \mathbf{a}_s,$$

with $\boldsymbol{\Sigma} = \text{Diag}(\sigma_i^2)$. Finally, note that if $a_{iir} = a_{iis} = 0$, then the last two terms drop out from the expression for covariance. Under normality, $\mu_i^{(3)} = 0$ and $\mu_i^{(4)} = 3$ the last two terms are always equal to zero.

This Theorem would be useful in, for example, Chapter 6. For an univariate application see Annex 5.A. Consider the following quadratic moment:

$$\begin{aligned}g_n(\rho_0, \tilde{\boldsymbol{\delta}}_n) &= \tilde{\mathbf{u}}_n^\top (\mathbf{I}_n - \rho_0 \mathbf{M}_n)^\top \mathbf{A}_n (\mathbf{I}_n - \rho_0 \mathbf{M}_n) \tilde{\mathbf{u}}_n, \\ &= \tilde{\mathbf{u}}_n^\top \mathbf{C}_0 \tilde{\mathbf{u}}_n,\end{aligned}$$

where $\mathbf{C}_0 = (1/2) (\mathbf{I}_n - \rho_0 \mathbf{M}_n)^\top (\mathbf{A}_n + \mathbf{A}_n^\top) (\mathbf{I}_n - \rho_0 \mathbf{M}_n)$. A Taylor expansion of $g_n(\rho_0, \tilde{\boldsymbol{\delta}}_n)$ around $\boldsymbol{\delta}_0$, gives:

$$\frac{1}{n} g_n(\rho_0, \tilde{\boldsymbol{\delta}}_n) = \frac{1}{n} g_n(\rho_0, \boldsymbol{\delta}_0) + \frac{1}{n} \frac{\partial g_n(\rho_0, \boldsymbol{\delta}_0)}{\partial \boldsymbol{\delta}} (\tilde{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0).$$

If $\mathbf{u}_n = \mathbf{y} - \mathbf{Z}_n \boldsymbol{\delta}_0$, with $\mathbf{Z}_n = [\mathbf{W}_n \mathbf{y}_n, \mathbf{X}_n]$, then

$$\begin{aligned}\frac{1}{n} \frac{\partial g_n(\rho_0, \boldsymbol{\delta}_0)}{\partial \boldsymbol{\delta}} &= 2 \frac{1}{n} \mathbb{E} \left(\mathbf{u}_n^\top \mathbf{C}_0 \frac{\partial \mathbf{u}_n}{\partial \boldsymbol{\delta}_0} \right), \\ &= -2n^{-1} \mathbb{E} (\mathbf{u}_n^\top \mathbf{C}_0 \mathbf{Z}_n).\end{aligned}$$

If we assume that $(\tilde{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0) = \mathbf{T}_n^\top \boldsymbol{\varepsilon}_n + o_p(1)$, and $\mathbf{u}_n = (\mathbf{I}_n - \rho_0 \mathbf{M}_n)^{-1} \boldsymbol{\varepsilon}_n$, then

$$\begin{aligned} \frac{1}{n} g_n(\rho_0, \tilde{\boldsymbol{\delta}}_n) &= \frac{1}{2} n^{-1} \boldsymbol{\varepsilon}_n^\top (\mathbf{A}_n + \mathbf{A}_n^\top) \boldsymbol{\varepsilon}_n - 2n^{-1} \mathbb{E}(\mathbf{u}_n^\top \mathbf{C}_0 \mathbf{Z}_n) \mathbf{T}_n^\top \boldsymbol{\varepsilon}_n + o_p(1) \\ &= \frac{1}{2} n^{-1} \boldsymbol{\varepsilon}_n^\top (\mathbf{A}_n + \mathbf{A}_n^\top) \boldsymbol{\varepsilon}_n + \mathbf{a}_n^\top \boldsymbol{\varepsilon}_n + o_p(1), \end{aligned}$$

where

$$\mathbf{a}_n = -2n^{-1} \mathbf{T}_n \mathbb{E}(\mathbf{Z}_n^\top \mathbf{C}_0 \mathbf{u}_n) = -n^{-1} \mathbf{T}_n \mathbb{E}[\mathbf{Z}_n^\top (\mathbf{I}_n - \rho_0 \mathbf{M}_n^\top) (\mathbf{A}_n + \mathbf{A}_n^\top) (\mathbf{I}_n - \rho_0 \mathbf{M}_n) \mathbf{u}_n].$$

Since \mathbf{Z}_n is stochastic, $\mathbf{a}_n \neq \mathbf{0}$.

3.11 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) $\hat{\mathbf{P}}_n \xrightarrow{p} \mathbf{P}$, finite, symmetric, and positive definite. Then $\hat{\boldsymbol{\beta}}_n$ exists in probability, and $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$.

Appendix

3.A Matrix Norm

Definition 3.A.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 propertie

- (a) $\|\mathbf{A}\| > 0$ when $\mathbf{A} \neq \mathbf{0}$ and $\|\mathbf{A}\| = 0$ if and only if $\mathbf{A} = \mathbf{0}$,
- (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

$$\|\mathbf{A}\|_1 = \max_j \sum_{i=1}^n |a_{ij}|$$

The spectral norm $\|\mathbf{A}\|_2$ is defined as

$$\|\mathbf{A}\|_2 = \sigma_{\max}(\mathbf{A}) = \sqrt{\lambda_{\max}(\mathbf{A}^\top \mathbf{A})}$$

The maximum absolute row sum norm is defined by

$$\|\mathbf{A}\|_{\infty} = \max_i \sum_{j=1}^n |a_{ij}|$$

$\|\mathbf{A}\|_1$, $\|\mathbf{A}\|_2$ and $\|\mathbf{A}\|_{\infty}$ satisfy the inequality $\|\mathbf{A}\|_2^2 \leq \|\mathbf{A}\|_1 \|\mathbf{A}\|_{\infty}$.

3.B Inequalities

Definition 3.B.1 — Multiplicativity of absolute value. For any two random variables a and b

$$|ab| = |a| |b| \quad (3.24)$$

Definition 3.B.2 — Triangle Inequality. For any real numbers x_j ,

$$\left| \sum_{j=1}^n x_j \right| \leq \sum_{j=1}^n |x_j| \quad (3.25)$$

Definition 3.B.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.B.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.B.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.B.6 — Cauchy-Schwarz Inequality. For two random variables

$$\mathbb{E} [|X_n Y_n|] \leq \{\mathbb{E} [X_n^2]\}^{1/2} \{\mathbb{E} [Y_n^2]\}^{1/2}$$

Maximum Likelihood Estimation

In this chapter, we begin the study of estimation methods for spatial models. In particular, we focus in the maximum likelihood (ML) estimation method. The ML procedure is one of the oldest and most widely used approaches in the literature for estimating spatial models. It provides a robust framework for obtaining consistent and efficient parameters estimates, even in the presence of spatial dependence.

4.1 What Are The Consequences of Applying OLS?

In this section, we examine the implications of using the Ordinary Least Squares (OLS) estimator on data generated by a Spatial Autoregressive (SAR) process. The key result is that OLS estimates of the coefficients in such models are biased and inconsistent. This means that, even with a large dataset, the estimated parameters will not converge to the true population parameters, even if we have a very large data set, which is a serious problem.

4.1.1 Finite and Asymptotic Properties

Let's begin by demonstrating that the OLS estimate of ρ becomes biased within the framework of the Spatial Lag Model (SLM). Consider the simplified first-order spatial autoregressive model:

$$\underset{(n \times 1)}{\mathbf{y}} = \rho_0 \underset{(n \times 1)}{\mathbf{W}} \underset{(n \times 1)}{\mathbf{y}} + \underset{(n \times 1)}{\boldsymbol{\varepsilon}}, \quad (4.1)$$

where ρ_0 represents the true population parameter. The reduced form for the **pure SLM** in Equation (4.1) is:

$$\mathbf{y} = (\mathbf{I}_n - \rho_0 \mathbf{W})^{-1} \boldsymbol{\varepsilon}. \quad (4.2)$$

Subsequently, the spatial lag term can be then expressed as:

$$\mathbf{W} \mathbf{y} = \mathbf{W} (\mathbf{I}_n - \rho_0 \mathbf{W})^{-1} \boldsymbol{\varepsilon}. \quad (4.3)$$

Now, recalling 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 Equation (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) yields the sampling error equation:

$$\begin{aligned}\hat{\rho}_{OLS} &= \rho_0 + \left[(\mathbf{W}\mathbf{y})^\top (\mathbf{W}\mathbf{y}) \right]^{-1} (\mathbf{W}\mathbf{y})^\top \boldsymbol{\varepsilon}, \\ &= \rho_0 + \left(\sum_{i=1}^n \mathbf{y}_{Li}^2 \right)^{-1} \left(\sum_{i=1}^n \mathbf{y}_{Li} \epsilon_i \right),\end{aligned}$$

where \mathbf{y}_{Li} is the i th element of the spatial lag operator $\mathbf{W}\mathbf{y} = \mathbf{y}_L$. Assuming a nonstochastic \mathbf{W} , the mathematical expectation of $\hat{\rho}_{OLS}$ is expressed as:

$$\begin{aligned}\mathbb{E}(\hat{\rho}_{OLS} | \mathbf{W}) &= \rho_0 + \mathbb{E} \left(\left[(\mathbf{W}\mathbf{y})^\top (\mathbf{W}\mathbf{y}) \right]^{-1} (\mathbf{W}\mathbf{y})^\top \boldsymbol{\varepsilon} \middle| \mathbf{W} \right), \\ &= \rho_0 + \left(\sum_{i=1}^n \mathbf{y}_{Li}^2 \right)^{-1} \mathbb{E} \left(\sum_{i=1}^n \mathbf{y}_{Li} \epsilon_i \middle| \mathbf{W} \right).\end{aligned}\tag{4.5}$$

Examining (4.5), it is evident that if the expectation of the last term is zero, $\hat{\rho}_{OLS}$ is unbiased. However, as shown in (4.6), this condition is not met, leading to bias unless $\rho_0 = 0$:

$$\begin{aligned}\mathbb{E} \left(\sum_{i=1}^n \mathbf{y}_{Li} \epsilon_i \middle| \mathbf{W} \right) &= \mathbb{E} \left[(\mathbf{W}\mathbf{y})^\top \boldsymbol{\varepsilon} \middle| \mathbf{W} \right], \\ &= \mathbb{E} \left[\boldsymbol{\varepsilon}^\top (\mathbf{I} - \rho_0 \mathbf{W}^\top)^{-1} \mathbf{W}^\top \boldsymbol{\varepsilon} \middle| \mathbf{W} \right] \quad \text{using (4.3),} \\ &= \mathbb{E} \left[\boldsymbol{\varepsilon}^\top \mathbf{C}^\top \boldsymbol{\varepsilon} \middle| \mathbf{W} \right], \\ &= \mathbb{E} \left[\text{tr} \boldsymbol{\varepsilon}^\top \mathbf{C}^\top \boldsymbol{\varepsilon} \middle| \mathbf{W} \right], \\ &= \mathbb{E} \left[\text{tr} \mathbf{C}^\top \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top \middle| \mathbf{W} \right], \\ &= \text{tr}(\mathbf{C}) \mathbb{E}(\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top | \mathbf{W}) \quad \text{since } \text{tr}(\mathbf{A}) = \text{tr}(\mathbf{A}^\top), \\ &\neq 0,\end{aligned}\tag{4.6}$$

where $\mathbf{C} = \mathbf{W}(\mathbf{I} - \rho_0 \mathbf{W})^{-1}$. 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} :

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^n a_{ii} = a_{11} + a_{22} + \dots + a_{nn}\tag{4.7}$$

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 $\text{tr}(\mathbf{A}) = \text{rank}(\mathbf{A})$.

Regarding consistency, we can express $\hat{\rho}_{OLS}$ as:

$$\hat{\rho}_{OLS} = \rho_0 + \left(\frac{1}{n} \sum_{i=1}^N \mathbf{y}_{Li}^2 \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{y}_{Li} \epsilon_i \right). \quad (4.10)$$


Under certain 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) demonstrates that, in some cases, the OLS estimator may still be consistent and asymptotically efficient relative to certain other estimators.

4.1.2 Illustration of Bias

To examine the properties of the OLS estimator when the data generating process follows a SAR process, we will conduct a straightforward simulation experiment. The fundamental design of this experiment involves generating simulated observations based on a known data generating process, specifically, a Spatial Lag Model (SLM). Subsequently, we will estimate the parameters for each simulated sample. If the estimator is biased, the average estimated parameters should deviate significantly from the true parameter.

In our simulation experiment, we assume the true Data Generating Process (DGP) to be:

$$\mathbf{y} = \rho_0 \mathbf{W}\mathbf{y} + \epsilon,$$

where the true value $\rho_0 = 0.7$, and the sample size for each sample is $n = 225$. The error term ε follows a normal distribution $N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, and \mathbf{W} is an artificial $n \times n$ weight matrix. The matrix \mathbf{W} is constructed from a neighbor list for rook contiguity on a 500×500 regular lattice.

In R, the syntax for creating the global parameters for the simulation is as follows:

```
# 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 $1/\min \omega$ and $1/\max \omega$, where $\omega = \text{eigen}(\mathbf{W})$, and returns the $n \times n$ inverted matrix $(\mathbf{I}_n - \rho \mathbf{W})^{-1}$.

The simulation loop is structured as follows:

```
# 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
}
```

It is important to note that \mathbf{W} is treated as fixed (nonstochastic), and therefore, it is created outside the simulation loop.

The summary of the estimated ρ is presented below:

```
# 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
```

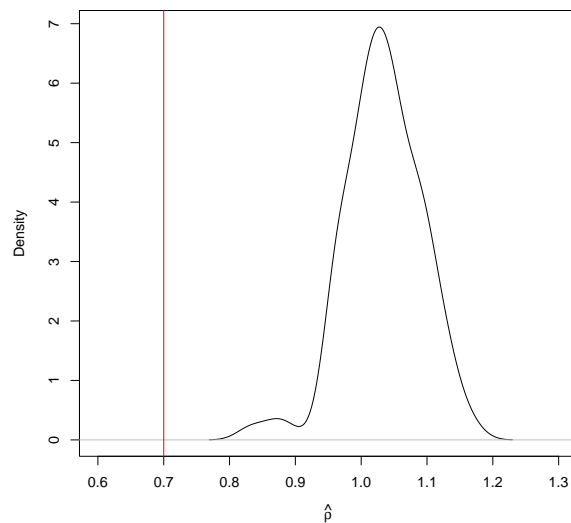
Upon examination, it is evident that the estimated ρ spans the range from 0.8 to 1.2, with this range failing to include the true parameter $\rho_0 = 0.7$. Furthermore, the mean of the estimated parameters is 1, a considerable distance from the true value of 0.7. This discrepancy suggests a substantial bias in the OLS estimator for the pure Spatial Lag Model (SLM) when ρ is high.

To visually explore the sampling distribution of the estimated parameters, a plot can be generated using the following approach:


```
# Plot density of estimated rho_hat.
plot(density(rho_hat),
     xlab = expression(hat(rho)),
     main = "")
abline(v = rho, col = "red")
```

Figure 4.1 illustrates the sampling distribution of ρ estimated by OLS for each sample in the Monte Carlo simulation study. Consistent with our earlier observations, the depicted pattern remains unchanged: the empirical distribution excludes the true parameter $\rho_0 = 0.7$.

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

The Maximum Likelihood (ML) estimation of spatial lag and spatial error regression models was initially formulated by [Ord \(1975\)](#). The foundation of this approach lies in assuming normality for the error terms. The joint likelihood, in turn, emerges from the multivariate normal distribution for the dependent variable \mathbf{y} . Unlike the conventional MLE estimator, the joint log likelihood for a spatial regression model does not simply equate to the sum of log likelihoods associated with individual observations. This departure arises from the inherent spatial simultaneity within the spatial 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 is given by the following structural model:

$$\begin{aligned} \mathbf{y} &= \rho_0 \mathbf{W} \mathbf{y} + \mathbf{X} \boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}, \\ \boldsymbol{\varepsilon} &\sim N(\mathbf{0}_n, \sigma_0^2 \mathbf{I}_n), \end{aligned} \quad (4.13)$$

where \mathbf{y} is an $n \times 1$ vector of 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 and exogenous variables; $\boldsymbol{\beta}_0$ is a k -dimensional vector of parameters; ρ_0 measures represents the spatial autoregressive parameter; and $\boldsymbol{\varepsilon}$ is an n -dimensional vector of error terms.

We assume the error terms are normally distributed with mean zero and variance-covariance $\sigma_0^2 \mathbf{I}_n$, implying homoskedasticity across spatial units. These distributional assumptions enable us to apply ML estimation. While the MLE exhibit desirable asymptotic properties such as consistency and efficiency under correct model specification, they can be sensitive to violations of the underlying assumption

The MLE seeks parameter values $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^\top, \hat{\rho}, \hat{\sigma}^2)^\top$ that maximizes the probability of observing the sample at hand.

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 and $\left| \frac{\partial \boldsymbol{\varepsilon}}{\partial \mathbf{y}} \right|$ is known as the **Jacobian**. The error vector is a function of \mathbf{y} as $\boldsymbol{\varepsilon} = \mathbf{A} \mathbf{y} - \mathbf{X} \boldsymbol{\beta}$, with $\mathbf{A} = \mathbf{I}_n - \rho \mathbf{W}$.¹ Here $\mathbf{A} \mathbf{y}$ is the **spatially filtered dependent variable**, i.e., with the effect of spatial autocorrelation taken out.

The Jacobian term is:

$$\det \left(\frac{\partial \boldsymbol{\varepsilon}}{\partial \mathbf{y}} \right) = \det(\mathbf{J}) = \det(\mathbf{A}) = \det(\mathbf{I}_n - \rho \mathbf{W}),$$

where $\mathbf{J} = \left(\frac{\partial \boldsymbol{\varepsilon}}{\partial \mathbf{y}} \right)$ is an $n \times n$ matrix, and $\det(\mathbf{I}_n - \rho \mathbf{W})$ is the determinant of an $n \times n$ 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 complicates its computation considerably. Note that 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:

$$f(\boldsymbol{\varepsilon} | \mathbf{X}) = (2\pi \cdot \sigma^2)^{-n/2} \exp \left[-\frac{1}{2\sigma^2} \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} \right].$$

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

$$(2\pi)^{-n/2} \det(\boldsymbol{\Sigma})^{-1/2} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right).$$

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 maximized, takes the form³

$$\begin{aligned} \ell(\boldsymbol{\theta}) &= \log |\mathbf{A}| - \frac{n \log(2\pi)}{2} - \frac{n \log(\sigma^2)}{2} - \frac{1}{2\sigma^2} (\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \\ &= \log |\mathbf{A}| - \frac{n \log(2\pi)}{2} - \frac{n \log(\sigma^2)}{2} - \frac{1}{2\sigma^2} \left[\mathbf{y}^\top \mathbf{A}^\top \mathbf{A} \mathbf{y} - 2 (\mathbf{A}\mathbf{y})^\top \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} \right], \end{aligned} \quad (4.14)$$

in which we use 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 (4.14) 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

To find the MLE for the SLM model, we need to maximize $\ell(\boldsymbol{\theta})$ in Equation (4.14) 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:

$$\frac{\partial(\rho \mathbf{W})}{\partial \rho} = \mathbf{W}. \quad (4.15)$$

$$\begin{aligned} \frac{\partial \mathbf{A}}{\partial \rho} &= \frac{\partial(\mathbf{I}_n - \rho \mathbf{W})}{\partial \rho}, \\ &= \frac{\partial \mathbf{I}_n}{\partial \rho} - \frac{\partial \rho \mathbf{W}}{\partial \rho}, \\ &= -\mathbf{W}. \end{aligned} \quad (4.16)$$

$$\frac{\partial \log |\mathbf{A}|}{\partial \rho} = \text{tr}(\mathbf{A}^{-1} \partial \mathbf{A} / \partial \rho) = \text{tr} [\mathbf{A}^{-1} (-\mathbf{W})]. \quad (4.17)$$

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).

Let $\boldsymbol{\varepsilon} = \mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$, then:

$$\frac{\partial \boldsymbol{\varepsilon}}{\partial \rho} = \frac{\partial (\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\partial \rho} = -\mathbf{W}\mathbf{y}. \quad (4.18)$$

$$\frac{\partial \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}}{\partial \rho} = \boldsymbol{\varepsilon}^\top (\partial \boldsymbol{\varepsilon} / \partial \rho) + (\partial \boldsymbol{\varepsilon}^\top / \partial \rho) \boldsymbol{\varepsilon} = 2\boldsymbol{\varepsilon}^\top (\partial \boldsymbol{\varepsilon} / \partial \rho) = 2\boldsymbol{\varepsilon}^\top (-\mathbf{W})\mathbf{y}. \quad (4.19)$$

$$\frac{\partial \mathbf{A}^{-1}}{\partial \rho} = -\mathbf{A}^{-1} (\partial \mathbf{A} / \partial \rho) \mathbf{A}^{-1} = \mathbf{A}^{-1} \mathbf{W} \mathbf{A}^{-1}. \quad (4.20)$$

$$\frac{\partial \text{tr}(\mathbf{A}^{-1} \mathbf{W})}{\partial \rho} = \text{tr}(\partial \mathbf{A}^{-1} \mathbf{W} / \partial \rho). \quad (4.21)$$

Taking the derivative of Equation (4.14) with respect to $\boldsymbol{\beta}$ yields

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = -\frac{1}{2\sigma^2} \left[-2 \left((\mathbf{A}\mathbf{y})^\top \mathbf{X} \right)^\top + 2\mathbf{X}^\top \mathbf{X} \boldsymbol{\beta} \right] = \frac{1}{\sigma^2} \mathbf{X}^\top (\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \quad (4.22)$$

and with respect to σ^2 yields

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \quad (4.23)$$

Solving both Equation (4.22) and (4.23), we obtain:

$$\hat{\boldsymbol{\beta}}_{ML}(\rho) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{A}\mathbf{y}, \quad (4.24)$$

$$\hat{\sigma}_{ML}^2(\rho) = \frac{(\mathbf{A}\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML})^\top (\mathbf{A}\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{ML})}{n}. \quad (4.25)$$

Note that conditional on ρ (assuming we know ρ), these estimators are simply OLS estimators applied to the *spatial filtered* dependent variable $\mathbf{A}\mathbf{y}$ and the exploratory variables \mathbf{X} . Moreover, after some manipulation, Equation (4.24) can be rewritten as

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{ML}(\rho) &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} - \rho (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}\mathbf{y}, \\ &= \hat{\boldsymbol{\beta}}_O - \rho \hat{\boldsymbol{\beta}}_L. \end{aligned} \quad (4.26)$$

Note that the first term in (4.26) 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 identities:

$$\mathbf{e}_O \equiv \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_O \text{ and } \mathbf{e}_L \equiv \mathbf{W}\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_L. \quad (4.27)$$

Then, plugging (4.26) into (4.25) yields

$$\tilde{\sigma}^2(\rho) = \frac{(\mathbf{e}_O - \rho \mathbf{e}_L)^\top (\mathbf{e}_O - \rho \mathbf{e}_L)}{n}. \quad (4.28)$$

Note that both (4.26) and (4.28) rely only on observables, except for ρ , and so are readily calculable given some estimate of ρ . Therefore, plugging (4.26) and (4.28) back into the likelihood (4.14) we obtain the **concentrated log-likelihood function**:

$$\ell(\rho) = -\frac{n}{2} - \frac{n}{2} \log(2\pi) - \frac{n}{2} \log \left[\frac{(\mathbf{e}_O - \rho \mathbf{e}_L)^\top (\mathbf{e}_O - \rho \mathbf{e}_L)}{n} \right] + \log |\mathbf{I}_n - \rho \mathbf{W}|, \quad (4.29)$$

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 (4.29). Once we obtain $\hat{\rho}$, we can easily obtain $\hat{\beta}$ and $\hat{\sigma}^2$. 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 (4.26).
- (b) Use $\hat{\beta}_O$ and $\hat{\beta}_L$ to compute the residuals in Equation (4.27).
- (c) Maximize the concentrated likelihood given in Equation (4.29) 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 4.24) and σ^2 (Equation 4.25).

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 (4.14) with respect to ρ , we obtain:

$$\begin{aligned}\frac{\partial\ell(\theta)}{\partial\rho} &= \left(\frac{\partial}{\partial\rho}\right) \log|\mathbf{A}| - \frac{1}{2\sigma^2} \left(\frac{\partial}{\partial\rho}\right) \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}, \\ &= -\text{tr}(\mathbf{A}^{-1}\mathbf{W}) + \frac{1}{2\sigma^2} 2\boldsymbol{\varepsilon}^\top \mathbf{W}\mathbf{y} \quad \text{using (4.17) and (4.19),} \\ &= -\text{tr}(\mathbf{A}^{-1}\mathbf{W}) + \frac{1}{2\sigma^2} 2\boldsymbol{\varepsilon}^\top \mathbf{W}\mathbf{y}, \\ &= -\text{tr}(\mathbf{A}^{-1}\mathbf{W}) + \frac{1}{\sigma^2} \boldsymbol{\varepsilon}^\top \mathbf{W}\mathbf{y}.\end{aligned}\tag{4.30}$$

Thus the complete gradient (or score function) is:

$$\nabla_{\theta} = \frac{\partial\ell(\theta)}{\partial\theta} = \begin{pmatrix} \frac{\partial\log L(\theta)}{\partial\beta} \\ \frac{\partial\log L(\theta)}{\partial\sigma^2} \\ \frac{\partial\log L(\theta)}{\partial\rho} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2} \mathbf{X}^\top \boldsymbol{\varepsilon} \\ \frac{1}{2\sigma^4} (\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} - n\sigma^2) \\ -\text{tr}(\mathbf{A}^{-1}\mathbf{W}) + \frac{1}{\sigma^2} \boldsymbol{\varepsilon}^\top \mathbf{W}\mathbf{y} \end{pmatrix}, \tag{4.31}$$

where $\boldsymbol{\varepsilon} = \mathbf{A}\mathbf{y} - \mathbf{X}\beta$.

4.2.3 Hessian

The Hessian matrix plays a critical role in subsequent sections, particularly for deriving the asymptotic variance-covariance matrix. To facilitate this, we dedicate this section to deriving the Hessian matrix for the SLM.

The Hessian is a $(k+2) \times (k+2)$ matrix of second derivatives, expressed as:

$$\mathbf{H}(\beta, \sigma^2, \rho) = \begin{pmatrix} \frac{\ell(\beta, \sigma^2, \rho)}{\partial\beta\partial\beta^\top} & \frac{\ell(\beta, \sigma^2, \rho)}{\partial\beta\partial\sigma^2} & \frac{\ell(\beta, \sigma^2, \rho)}{\partial\beta\partial\rho} \\ \frac{\ell(\beta, \sigma^2, \rho)}{\partial\sigma^2\partial\beta^\top} & \frac{\ell(\beta, \sigma^2, \rho)}{\partial(\sigma^2)^2} & \frac{\ell(\beta, \sigma^2, \rho)}{\partial\sigma^2\partial\rho} \\ \frac{\ell(\beta, \sigma^2, \rho)}{\partial\rho\partial\beta^\top} & \frac{\ell(\beta, \sigma^2, \rho)}{\partial\rho\partial\sigma^2} & \frac{\ell(\beta, \sigma^2, \rho)}{\partial\rho^2} \end{pmatrix}.$$

Using the first-order condition for β from (4.22), the second derivatives are:

$$\frac{\partial^2 \ell(\theta)}{\partial \beta \partial \beta^\top} = -\frac{1}{\sigma^2} (\mathbf{X}^\top \mathbf{X}), \quad (4.32)$$

$$\frac{\partial^2 \ell(\theta)}{\partial \beta \partial \sigma^2} = -\frac{1}{(\sigma^2)^2} \mathbf{X}^\top \boldsymbol{\varepsilon}, \quad (4.33)$$

$$\frac{\partial^2 \ell(\theta)}{\partial \beta \partial \rho} = -\frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{W} \mathbf{y}. \quad (4.34)$$

Using the first-order condition for σ^2 from (4.23), we obtain

$$\frac{\partial^2 \ell(\theta)}{\partial (\sigma^2)^2} = \frac{n}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3} \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}, \quad (4.35)$$

and:

$$\begin{aligned} \frac{\partial^2 \ell(\theta)}{\partial \sigma^2 \partial \rho} &= \frac{1}{2\sigma^4} \left[2\boldsymbol{\varepsilon}^\top \left(\frac{\partial \boldsymbol{\varepsilon}}{\partial \rho} \right) \right] \quad \text{using Equation (4.19),} \\ &= -\frac{1}{\sigma^4} \boldsymbol{\varepsilon}^\top \mathbf{W} \mathbf{y}, \\ &= -\frac{\boldsymbol{\varepsilon}^\top \mathbf{W} \mathbf{y}}{\sigma^4}. \end{aligned} \quad (4.36)$$

Finally, working in the second derivatives for ρ , and using (4.30), we obtain

$$\begin{aligned} \frac{\partial^2 \ell(\theta)}{\partial \rho^2} &= -\left(\frac{\partial}{\partial \rho} \right) \text{tr}(\mathbf{A}^{-1} \mathbf{W}) + \frac{1}{\sigma^2} \left(\frac{\partial}{\partial \rho} \right) \boldsymbol{\varepsilon}^\top \mathbf{W} \mathbf{y}, \\ &= -\text{tr} \left(\frac{\partial \mathbf{A}^{-1} \mathbf{W}}{\partial \rho} \right) + \frac{1}{\sigma^2} \left(\frac{\partial}{\partial \rho} \right) (\mathbf{A} \mathbf{y})^\top \mathbf{W} \mathbf{y}, \\ &= -\text{tr} (\mathbf{A}^{-1} \mathbf{W} \mathbf{A}^{-1} \mathbf{W}) + \frac{1}{\sigma^2} (-\mathbf{y}^\top \mathbf{W}^\top \mathbf{W} \mathbf{y}), \\ &= -\text{tr} [(\mathbf{W} \mathbf{A}^{-1})^2] - \frac{1}{\sigma^2} (\mathbf{y}^\top \mathbf{W}^\top \mathbf{W} \mathbf{y}). \end{aligned} \quad (4.37)$$

Combining the results, the Hessian matrix for the SLM is:

$$\mathbf{H}(\beta, \sigma^2, \rho) = \begin{pmatrix} -\frac{1}{\sigma^2} (\mathbf{X}^\top \mathbf{X}) & -\frac{1}{(\sigma^2)^2} \mathbf{X}^\top \boldsymbol{\varepsilon} & -\frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{W} \mathbf{y} \\ \cdot & \frac{n}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3} \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} & -\frac{\boldsymbol{\varepsilon}^\top \mathbf{W} \mathbf{y}}{\sigma^4} \\ \cdot & \cdot & -\text{tr} [(\mathbf{C})^2] - \frac{1}{\sigma^2} (\mathbf{y}^\top \mathbf{W}^\top \mathbf{W} \mathbf{y}) \end{pmatrix} \quad (4.38)$$

which is symmetric and $\mathbf{C} = \mathbf{W} \mathbf{A}^{-1}$.

4.2.4 Ord's Jacobian

A key feature of the concentrated log-likelihood function in Equation (4.29) is the Jacobian term $|\mathbf{I}_n - \rho \mathbf{W}|$. This term presents computational challenges, as estimating $\hat{\rho}$ requires evaluating the determinant of the $n \times n$ matrix $|\mathbf{I}_n - \rho \mathbf{W}|$ at each iteration. However, Ord (1975) provided a significant simplification based on the eigen values of \mathbf{W} . Specifically, Ord (1975) noted that:

$$|\omega \mathbf{I}_n - \mathbf{W}| = \prod_{i=1}^n (\omega - \omega_i),$$

where ω_i are the eigenvalues of \mathbf{W} . Consequently, for $\mathbf{I}_n - \rho\mathbf{W}$, the determinant can be expressed as:

$$|\mathbf{I}_n - \rho\mathbf{W}| = \prod_{i=1}^n (1 - \rho\omega_i).$$

The corresponding log-determinant term follows as:

$$\log |\mathbf{I}_n - \rho\mathbf{W}| = \sum_{i=1}^n \log(1 - \rho\omega_i). \quad (4.39)$$

This formulation offers a significant computational advantage: the eigenvalues of \mathbf{W} only need to be computed once. While the initial computation has some overhead, it drastically reduces the computational burden during iterative evaluations of the log-likelihood. For datasets with more than 4,000 observations, Ord's method is typically much faster than direct determinant computation.

This eigenvalue-based approach also delineates the admissible domain for 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:

$$\ell(\rho) = -\frac{n}{2} - \frac{n}{2} \log(2\pi) - \frac{n}{2} \log \left[\frac{(\mathbf{e}_O - \rho\mathbf{e}_L)^\top (\mathbf{e}_O - \rho\mathbf{e}_L)}{n} \right] + \sum_{i=1}^n \log(1 - \rho\omega_i). \quad (4.40)$$

An alternative to Ord's approach is the characteristic root method outlined by [Smirnov and Anselin \(2001\)](#). This method facilitates the estimation of spatial lag models for extremely large datasets ($> 100,000$ observations) in very short computational 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 further approximations and methods, 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 discussed in Section [2.1.3](#), one approach to account for spatial autocorrelation in regression models is to model the error term as a spatial process. The SEM model is given by

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta_0 + \mathbf{u}, \\ \mathbf{u} &= \lambda_0 \mathbf{W}\mathbf{u} + \boldsymbol{\varepsilon}, \\ \boldsymbol{\varepsilon} &\sim N(\mathbf{0}, \sigma_0^2 \mathbf{I}_n), \end{aligned} \quad (4.41)$$

where λ_0 is the spatial autoregressive coefficient for the error lag term $\mathbf{W}\mathbf{u}$ (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 $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I}_n)$.

The SEM does not require a theoretical justification for the spatial process but is consistent with situations where omitted determinants of the dependent variable are **spatially**

autocorrelated, or with a situation where unobserved shocks follow a spatial pattern (Elhorst, 2014). In essence, SEM treats spatial correlation primarily as a nuisance.

When $\lambda_0 > 0$, the error term exhibits positive spatial correlation, implying clustering of similar values. This means the errors for a spatial unit i systematically vary with the errors of nearby observations j . For example 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 can be expressed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + (\mathbf{I}_n - \lambda_0 \mathbf{W})^{-1} \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{u},$$

where $\mathbf{u} = (\mathbf{I}_n - \lambda_0 \mathbf{W})^{-1} \boldsymbol{\varepsilon}$. It follows that $\mathbb{E}(\mathbf{u}|\mathbf{W}, \mathbf{X}) = \mathbf{0}$, and the variance-covariance matrix of \mathbf{u} is given by:

$$\mathbb{V}(\mathbf{u}|\mathbf{W}, \mathbf{X}) = \mathbb{E}(\mathbf{u}\mathbf{u}^\top|\mathbf{W}, \mathbf{X}) = \sigma_0^2 (\mathbf{I}_n - \lambda_0 \mathbf{W})^{-1} (\mathbf{I}_n - \lambda_0 \mathbf{W}^\top)^{-1} = \sigma_0^2 \boldsymbol{\Omega}_u^{-1}, \quad (4.42)$$

where $\boldsymbol{\Omega}_u = (\mathbf{I}_n - \lambda_0 \mathbf{W})(\mathbf{I}_n - \lambda_0 \mathbf{W}^\top)$. The variance covariance (4.42) 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 (4.42) yields non constant diagonal elements in the error covariance matrix, thus inducing heteroskedasticity in \mathbf{u} , irrespective of the heteroskedasticity of $\boldsymbol{\varepsilon}$. Finally, $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma_0^2 \boldsymbol{\Omega}_u^{-1})$.

 The OLS estimates of model in Equation (4.41) are unbiased, but inefficient if $\lambda_0 \neq 0$.

The inefficiency of OLS in this context arises because the presence of spatially correlated errors invalidates the assumption of spherical error variance. As a result:

- (a) **Inefficient parameter estimation:** OLS does not account for the spatial structure in the errors, leading to less precise estimates of $\boldsymbol{\beta}_0$.
- (b) **Invalid statistical inference:** The inefficiency in OLS estimation biases the variance and standard error estimates for $\boldsymbol{\beta}_0$, rendering significance test unreliable.

To address these issues, generalized least squares (GLS) should be used for more efficient parameter estimation. GLS accounts for the spatial covariance structure in $\boldsymbol{\Omega}_u^{-1}$, producing valid inference for the SEM model.

4.3.2 Log-likelihood function

The SEM model defined in Equation (4.41) implies the following transformation for the residual term:

$$\boldsymbol{\varepsilon} = (\mathbf{I}_n - \lambda_0 \mathbf{W}) \mathbf{y} - (\mathbf{I}_n - \lambda_0 \mathbf{W}) \mathbf{X}\boldsymbol{\beta}_0 = \mathbf{B}_0 \mathbf{y} - \mathbf{B}_0 \mathbf{X}\boldsymbol{\beta}_0,$$

where $\mathbf{B}_0 = (\mathbf{I}_n - \lambda_0 \mathbf{W})$.

To derive the log-likelihood function, we first need the joint density function. Using the Transformation Theorem, the joint density of \mathbf{y} is obtained as:

$$f(y_1, \dots, y_n | \mathbf{X}; \boldsymbol{\theta}) = f(\boldsymbol{\varepsilon}(\mathbf{y}) | \mathbf{X}; \boldsymbol{\theta}) \cdot |\mathbf{J}|,$$

where \mathbf{J} is the Jacobian of the transformation. The Jacobian 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} —is:

$$f(\boldsymbol{\varepsilon}(\mathbf{y})|\mathbf{X};\boldsymbol{\theta}) = (2\pi\sigma^2)^{-n/2} \exp \left[-\frac{[(\mathbf{I}_n - \lambda\mathbf{W})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})]^\top [(\mathbf{I}_n - \lambda\mathbf{W})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})]}{2\sigma^2} \right],$$

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

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Omega}(\lambda)(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2} + \log |\mathbf{I}_n - \lambda\mathbf{W}|, \quad (4.43)$$

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 (4.43) is equivalent to minimizing the sum of the transformed errors, $\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}$, adjusted by the Jacobian term, $\log |\mathbf{I}_n - \lambda\mathbf{W}|$. This adjustment ensures that the maximum likelihood (ML) estimates differ from the ordinary least squares (OLS) estimates. However, the two coincide as $\lambda \rightarrow 0$.

To derive the ML estimates, we apply the first-order necessary conditions (FONC) to the log-likelihood function (4.43). Taking the derivative with respect to $\boldsymbol{\beta}$ yields:

$$\begin{aligned} \boldsymbol{\beta}_{ML}(\lambda) &= [\mathbf{X}^\top \boldsymbol{\Omega}(\lambda) \mathbf{X}]^{-1} \mathbf{X}^\top \boldsymbol{\Omega}(\lambda) \mathbf{y}, \\ &= [(\mathbf{B}\mathbf{X})^\top (\mathbf{B}\mathbf{X})]^{-1} (\mathbf{B}\mathbf{X})^\top \mathbf{B}\mathbf{y}, \\ &= [\mathbf{X}(\lambda)^\top \mathbf{X}(\lambda)]^{-1} \mathbf{X}(\lambda)^\top \mathbf{y}(\lambda), \end{aligned} \quad (4.44)$$

where:

$$\begin{aligned} \mathbf{X}(\lambda) &= \mathbf{B}\mathbf{X} = (\mathbf{I} - \lambda\mathbf{W})\mathbf{X} = (\mathbf{X} - \lambda\mathbf{W}\mathbf{X}), \\ \mathbf{y}(\lambda) &= (\mathbf{y} - \lambda\mathbf{W}\mathbf{y}). \end{aligned}$$

When λ is known, this estimator is equivalent to the generalized least squares (GLS) estimator— $\hat{\boldsymbol{\beta}}_{ML} = \hat{\boldsymbol{\beta}}_{GLS}$ —and it can be interpreted as the OLS estimate obtained from regressing $\mathbf{y}(\lambda)$ 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.

R 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*.

Similarly, the FONC of (4.43) with respect to σ^2 gives the MLE for the error variance:

$$\sigma_{ML}^2(\lambda) = \frac{1}{n} (\hat{\boldsymbol{\varepsilon}}^\top \mathbf{B}^\top \mathbf{B} \hat{\boldsymbol{\varepsilon}}) = \frac{1}{n} \hat{\boldsymbol{\varepsilon}}^\top(\lambda) \hat{\boldsymbol{\varepsilon}}(\lambda), \quad (4.45)$$

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 (4.43) 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:

$$\ell(\lambda) = \text{const} + \frac{n}{2} \log \left[\frac{1}{n} \hat{\boldsymbol{\varepsilon}}^\top \mathbf{B}^\top \mathbf{B} \hat{\boldsymbol{\varepsilon}} \right] + \log |\mathbf{B}|, \quad (4.46)$$

where the residual vector indirectly depends on λ . An iterative optimization procedure is generally required to estimate all parameters.

The iterative procedure for ML estimation of the SEM, based on Anselin (1988), can be summarized as follows:

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:

$$\text{AsyVar}(\boldsymbol{\beta}, \sigma^2, \lambda) = \begin{pmatrix} \frac{\mathbf{X}(\lambda)^\top \mathbf{X}(\lambda)}{\sigma^2} & 0 & 0 \\ 0 & \frac{n}{2\sigma^4} & \frac{\text{tr}(\mathbf{W}_B)}{\sigma^2} \\ 0 & \frac{\text{tr}(\mathbf{W}_B)}{\sigma^2} & \text{tr}(\mathbf{W}_B)^2 + \text{tr}(\mathbf{W}_B^\top \mathbf{W}_B) \end{pmatrix}^{-1}, \quad (4.47)$$

where $\mathbf{W}_B = \mathbf{W}(\mathbf{I} - \lambda \mathbf{W})^{-1}$.

4.4 Asymptotic Properties of SLM

This section reviews the asymptotic properties of Maximum Likelihood (ML) and Quasi-Maximum Likelihood (QML) estimators for the Spatial Lag Model (SLM), following the foundational work of Lee (2004). We focus on consistency and asymptotic normality.

4.4.1 Consistency of QMLE

Consider the true SLM given by:

$$\mathbf{y}_n = \mathbf{X}_n \beta_0 + \lambda_0 \mathbf{W}_n \mathbf{y}_n + \boldsymbol{\varepsilon}_n,$$

where all variables and the weight matrix \mathbf{W}_n are indexed by n to explicitly denote their dependence on the sample size.

Lee (2004) establishes the asymptotic properties (consistency and asymptotic normality) of the ML and QML estimator under specific regularity conditions.

To begin, we impose the following assumption on the error terms $\boldsymbol{\varepsilon}_n$:

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 i.i.d with mean zero and variance σ^2 . Its moment $\mathbb{E}(|\epsilon|^{4+\gamma})$ for some $\gamma > 0$ exists.

Assumption 4.3 ensures homoskedasticity of the errors and guarantees the existence of finite variances for quadratic forms involving $\boldsymbol{\varepsilon}_n$. This is critical for deriving the asymptotic properties of the QMLE, as it enables the application of the Central Limit Theorem (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 .

In spatial econometrics, the spatial weight matrix \mathbf{W}_n is often conceptualized as a triangular array, reflecting the fact that its structure changes as the sample size increases (see Section 3.7). For instance, the element w_{ij} in \mathbf{W}_n may differ when $n = 50$ versus $n = 55$. This indexing scheme explicitly acknowledges such changes, and the elements of \mathbf{W}_n are denoted by $w_{n,ij}$.

A natural question arises regarding the boundedness of the elements of \mathbf{W}_n . Definition 4.4.1 provides a precise characterization of sequences that are bounded or diverging. In this context, we impose the following assumption:

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 if $X_n = O(b_n)$, then

$$\lim_{n \rightarrow \infty} \frac{X_n}{b_n} = -\infty < c < \infty.$$

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 establish a direct connection between the spatial weight matrix and the sample size n . Intuitively, as the sample size n increases, the row sums of the weight matrix \mathbf{W}_n are expected to grow, as regions can potentially have more neighbors (see the discussion in Section 3.7). The rate of growth of the spatial weights $w_{n,ij}$ with n can be either bounded (limited number of neighbors) or divergent (unlimited number of neighbors). 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 \mathbf{W}_n . Specifically, the row and column sums of \mathbf{W}_n , before row-normalization, should not diverge to infinity at a rate equal to or faster than the rate of the sample size n . This contrasts slightly with the conditions in Kelejian and Prucha (1998) and Kelejian and Prucha (1999), which require that the row and column sums of \mathbf{W} and $(\mathbf{I}_n - \rho \mathbf{W})^{-1}$, before row-normalization, remain uniformly bounded in absolute value as $n \rightarrow \infty$. Both sets of conditions aim to ensure that the cross-sectional correlation remains manageable, meaning that the correlation between two spatial units diminishes as the distance separating them increases to infinity.

In practical terms, these assumptions often imply that no spatial unit is assumed to have more than a fixed number, say q , of neighbors. This limitation ensures that the conditions in Lee (2004) and Kelejian and Prucha (1998, 1999) are satisfied.

However, these conditions may not hold if the spatial weight matrix is defined as an inverse distance matrix. To illustrate, consider an infinite number of spatial units arranged linearly. Let the distance between each spatial unit and its nearest left and right neighbors be d , its next-nearest neighbors $2d$, and so on. In such cases, the conditions in Kelejian and Prucha (1998, 1999) may be violated. See Figure 4.2 for an illustration.

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}_n is an inverse distance matrix, its off-diagonal elements are defined as $w_{ij} = 1/d_{ij}$, where d_{ij} represents the distance between two spatial units i and j . The row sum of \mathbf{W}_n for any given i is:

$$\sum_{j=1}^n w_{ij} = 1/d + 1/d + 1/2d + 1/2d + \cdots = 2 \times (1/d + 1/2d + 1/3d + \cdots),$$

which represents a harmonic series that diverges to infinity.

This divergence is one of the main reasons why some empirical applications introduce a cutoff distance d^* , such that $w_{ij} = 0$ if $d_{ij} > d^*$. Imposing this cutoff ensures that the row sums remain finite and avoids potential numerical issues.

However, note that even without a cutoff, the ratio of the row sum to the sample size,

$$\frac{2 \times \left(\frac{1}{d} + \frac{1}{2d} + \frac{1}{3d} + \cdots \right)}{n},$$

converges to zero as $n \rightarrow \infty$. Therefore, the condition proposed by Lee (2004) is satisfied, implying that an inverse distance matrix without a cutoff does not necessarily violate consistency requirements. Assumption 4.5 excludes cases where the row sums, $\sum_{j=1}^n w_{ij}$ for $i = 1, \dots, n$, diverge to infinity at a rate equal to or faster than the sample size n . In such cases, the maximum likelihood (ML) estimator would likely be inconsistent.

A special case where the sequence $\{h_n\}$ is bounded arises when the number of neighbors is fixed, as in the k -nearest neighbors approach. However, the example of inverse distance weights highlights why these matrices can sometimes lead to numerical problems or unexpected outcomes in empirical applications. This is because, in practice, the sample size n is typically finite and does not approach infinity, leading to potential complications with unbounded row sums.

What happens if h_n is unbounded? In 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 as the total number of unit increases to infinity. For example, it excludes scenarios where spatial units correspond to counties, and neighbors are defined as those with contiguous border.

When does $h_n \rightarrow \infty$? This case requires that each unit in the limit has infinitely many neighbors. As noted by Lee (2002), in economic applications where (i) the neighbors of each unit are densely distributed in a relevant space, or (ii) each unit is significantly influenced by a large proportion of the total population units, it is likely that $\sum_{j=1}^n d_{ij}$ diverges and $(1/n) \sum_{j=1}^n d_{ij}$ converges as $n \rightarrow \infty$.

For example, consider $d_{ij} = 1/|r_i - r_j|$, where r_i is the proportion of state i 's population of African descent. Since no state in the U.S. has a zero proportion of African-Americans, $d_{ij} > 0$. Here, $(1/n) \sum_{j=1}^n d_{ij}$ will remain bounded away from zero, while $\sum_{j=1}^n d_{ij}$ will diverge at a rate proportional to n .

Another case arises when all cross-sectional units are assumed to be neighbors of one another, with equal weights assigned. If all off-diagonal elements of the spatial weights matrix are $w_{ij} = 1$, then the row and column sums become $n - 1$, diverging to infinity as $n \rightarrow \infty$. However, in this scenario, $(n - 1)/n \rightarrow 1$ rather than converging to zero. Consequently, a spatial weight matrix with equal weights and subsequent row-normalization, $w_{ij} = 1/(n - 1)$, violates both Lee (2004)'s and Kelejian and Prucha (1998, 1999)'s conditions, making it unsuitable for consistent estimation.

Alternatively, a group interaction matrix, as introduced by Case (1991), satisfies the conditions. In this setup, “neighbors” refer to individuals within the same district. Suppose there are R districts and m individuals per district, with a total sample size $n = mR$. Each neighbor within a district is given equal weight, and the spatial weight matrix is defined as $\mathbf{W}_n = \mathbf{I}_R \otimes \mathbf{B}_m$, where $\mathbf{B}_m = (\mathbf{e}_m \mathbf{e}_m^\top - \mathbf{I}_m)/(m - 1)$.

Here, $h_n = m - 1$, and $h_n/n = (m - 1)/(mR) = O(1/R)$. If the sample size n increases through growth in both R and m , $h_n \rightarrow \infty$ and $h_n/n \rightarrow 0$ as $n \rightarrow \infty$. Thus, this matrix

satisfies Lee (2004)'s condition.

R The boundedness or divergence of $\{h_n\}$ has significant implications for OLS estimation. When $\{h_n\}$ is bounded, the OLS estimators of β and ρ are inconsistent. However, when $\{h_n\}$ diverges, these estimators can be consistent (see Lee, 2002).

In summary, when $\{h_n\}$ is a bounded sequence, it implies that each unit has a fixed, small number of neighbors, typically based on geographical proximity. In contrast, when $\{h_n\}$ diverges, it reflects scenarios where each unit interacts with a large number of neighbors, which frequently occurs in empirical studies of social networks, economic spillovers, 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

$$\mathbf{y}_n = \mathbf{A}_{n0}^{-1}(\mathbf{X}_n\beta_0 + \varepsilon_n),$$

where $\mathbf{A}_{n0}^{-1} = (\mathbf{I}_n - \rho_0\mathbf{W}_n)$ and with the following expectation and variance:

$$\begin{aligned}\mathbb{E}(\mathbf{y}_n) &= (\mathbf{I}_n - \rho_0\mathbf{W}_n)^{-1} \mathbf{X}_n\beta_0 = \mathbf{A}_{n0}^{-1} \mathbf{X}_n\beta_0, \\ \mathbb{V}(\mathbf{y}_n) &= \mathbb{E}(\mathbf{y}_n\mathbf{y}_n^\top) = \sigma_0^2 (\mathbf{I}_n - \rho_0\mathbf{W}_n)^{-1} [(\mathbf{I}_n - \rho_0\mathbf{W}_n)^{-1}]^\top = \sigma_0^2 \mathbf{A}_{n0}^{-1} (\mathbf{A}_{n0}^{-1})^\top.\end{aligned}$$

For further reference, note that we can write the reduced-form equation as follows:

$$\begin{aligned}\mathbf{y}_n &= \mathbf{X}_n\beta_0 + \rho_0\mathbf{W}_n\mathbf{y}_n + \varepsilon_n, \\ &= \mathbf{X}_n\beta_0 + \rho_0\mathbf{W}_n [\mathbf{A}_{n0}^{-1} \mathbf{X}_n\beta_0 + \mathbf{A}_{n0}^{-1} \varepsilon_n] + \varepsilon_n, \\ &= \mathbf{X}_n\beta_0 + \rho_0\mathbf{W}_n\mathbf{A}_{n0}^{-1} \mathbf{X}_n\beta_0 + \rho_0\mathbf{W}_n\mathbf{A}_{n0}^{-1} \varepsilon_n + \varepsilon_n, \\ &= \mathbf{X}_n\beta_0 + \rho_0\mathbf{W}_n\mathbf{A}_{n0}^{-1} \mathbf{X}_n\beta_0 + (\mathbf{I}_n + \rho_0\mathbf{W}_n\mathbf{A}_{n0}^{-1}) \varepsilon_n, \\ &= \mathbf{X}_n\beta_0 + \rho_0\mathbf{C}_{n0}\mathbf{X}_n\beta_0 + (\mathbf{I}_n + \rho_0\mathbf{C}_{n0}) \varepsilon_n, \\ &= \mathbf{X}_n\beta_0 + \rho_0\mathbf{C}_{n0}\mathbf{X}_n\beta_0 + \mathbf{A}_{n0}^{-1} \varepsilon_n,\end{aligned}$$

because $\mathbf{I}_n + \rho_0\mathbf{C}_{n0} = \mathbf{A}_{n0}^{-1}$ (see Exercise 4.6), where $\mathbf{C}_{n0} = \mathbf{W}_n\mathbf{A}_{n0}^{-1}$.

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 3.8.

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

$$w_{n,ij} = \frac{d_{ij}}{\sum_{l=1}^n d_{il}}$$

⁴Equivalently, this assumption rules out the unit root case in time series.

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 $O(h_n)$ 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 $\|\rho \mathbf{W}\| < 1$, then $\mathbf{I}_n - \rho \mathbf{W}_n$ is invertible for all n . Then if $\|\rho \mathbf{W}\| < 1$, then the sequence of matrices $\|(\mathbf{I}_n - \rho \mathbf{W}_n)^{-1}\|$ 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 $(\mathbf{I}_n - \rho \mathbf{W})^{-1}$ 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 $|(\mathbf{I}_n - \rho \mathbf{W})^{-1}|$ 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

$$\lim_{n \rightarrow \infty} \frac{1}{n} (\mathbf{X}_n, \mathbf{C}_n \mathbf{X}_n \boldsymbol{\beta}_0)' (\mathbf{X}_n, \mathbf{C}_n \mathbf{X}_n \boldsymbol{\beta}_0)$$

exists and is nonsingular.

This is a sufficient condition for global identification of $\boldsymbol{\theta}_0$.

Theorem 4.12 — Consistency. Let $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^\top, \rho_0, \sigma_0^2)^\top$. Under assumption 4.3-4.11, $\boldsymbol{\theta}_0$ is globally identifiable and $\hat{\boldsymbol{\theta}}_n$ is a consistent estimator of $\boldsymbol{\theta}_0$.

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 $\hat{\boldsymbol{\theta}}_n$ 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

$$\frac{\partial \ell_n(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}} = \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + \frac{\partial^2 \ell_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0),$$

where $\tilde{\boldsymbol{\theta}}_n = \alpha_n \hat{\boldsymbol{\theta}}_n + (1 - \alpha_n) \boldsymbol{\theta}_0$ and $\alpha_n \in [0, 1]$, therefore:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = - \left[\frac{1}{n} \frac{\partial^2 \ell_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right]^{-1} \frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}. \quad (4.48)$$

As standard in asymptotic theory of MLE, we need to show that the first element of the rhs of (4.48) converges to some finite matrix. 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):

$$\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma_0^2 \sqrt{n}} \mathbf{X}_n^\top \boldsymbol{\varepsilon}_n \quad (4.49)$$

$$\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \sigma^2} = \frac{1}{2\sigma_0^4 \sqrt{n}} (\boldsymbol{\varepsilon}_n' \boldsymbol{\varepsilon}_n - n\sigma_0^2) \quad (4.50)$$

$$\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \rho} = \frac{1}{\sigma_0^2 \sqrt{n}} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \boldsymbol{\varepsilon}_n + \frac{1}{\sigma_0^2 \sqrt{n}} (\boldsymbol{\varepsilon}_n^\top \mathbf{C}_{n0} \boldsymbol{\varepsilon}_n - \sigma_0^2 \text{tr}(\mathbf{C}_{n0})) \quad (4.51)$$

As explained by Lee (2004, pag. 1905), these are linear and quadratic functions of $\boldsymbol{\varepsilon}_n$. In particular, the asymptotic distribution of (4.51) 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:

$$\mathbb{E} \left[\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \cdot \frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^\top} \right] = -\mathbb{E} \left(\frac{1}{n} \frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right) + \boldsymbol{\Omega}_{\boldsymbol{\theta},n},$$

where

$$-\mathbb{E} \left(\frac{1}{n} \frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right) = \begin{pmatrix} \frac{1}{n\sigma_0^2} (\mathbf{X}_n^\top \mathbf{X}_n) & \frac{1}{n\sigma_0^2} \mathbf{X}_n^\top (\mathbf{C}_n \mathbf{X}_n \boldsymbol{\beta}_0) & \mathbf{0}^\top \\ \frac{1}{n} \text{tr}(\mathbf{C}_{n0} \mathbf{C}_{n0}) + \frac{1}{n\sigma_0^2} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0) & \frac{1}{n\sigma_0^2} \text{tr}(\mathbf{C}_{n0}) & \frac{1}{2\sigma_0^4} \end{pmatrix} \quad (4.52)$$

and $\mathbf{C}_{n0}^s = \mathbf{C}_{n0} + \mathbf{C}_{n0}^\top$. Equation (4.52) represents the average Hessian matrix (or information matrix when ε 's are **normal**). The matrix $\mathbf{\Omega}_{\theta,n}$ is

$$\mathbf{\Omega}_{\theta,n} = \begin{pmatrix} \mathbf{0} & * & * \\ \frac{\mu_3}{n\sigma_0^4} \sum_{i=1}^n \mathbf{C}_{n,ii} \mathbf{x}_{i,n} & \frac{2\mu_3}{n\sigma_0^4} \sum_{i=1}^n \mathbf{C}_{n,ii} \mathbf{C}_{n,ii} \mathbf{X}_m \boldsymbol{\beta}_0 + \frac{(\mu_4 - 3\sigma_0^4)}{n\sigma_0^4} \sum_{i=1}^n \mathbf{C}_{n,ii}^2 & * \\ \frac{1}{n2\sigma_0^6} [\mu_3 \mathbf{t}_n^\top \mathbf{G}_n \mathbf{X}_n \boldsymbol{\beta}_0 + (\mu_4 - 3\sigma_0^4) \text{tr}(\mathbf{G}_n)] & \frac{(\mu - 3\sigma_0^4)}{4\sigma_0^8} & * \end{pmatrix} \quad (4.53)$$

which is a symmetric matrix with the second, third, and fourth moments of ε . If ε_n is normally distributed, then $\mathbf{\Omega}_{\theta,n} = \mathbf{0}$.

Derivation of (4.52) 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,

$$\sqrt{n} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \text{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{\Omega}_{\boldsymbol{\theta}} \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}), \quad (4.54)$$

where $\mathbf{\Omega}_{\boldsymbol{\theta}} = \lim_{n \rightarrow \infty} \mathbf{\Omega}_{\theta,n}$ and

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = - \lim_{n \rightarrow \infty} \mathbb{E} \left[\frac{1}{n} \frac{\partial^2 \log L_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right], \quad (4.55)$$

which are assumed to exist. If the ε_i 's are **normally distributed**, then:

$$\sqrt{n} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \text{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}). \quad (4.56)$$

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 2.4.2, 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:

$$\begin{aligned} \bar{M}(\boldsymbol{\theta})_{\text{direct}} &= n^{-1} \text{tr}[\mathbf{S}_r(\boldsymbol{\theta})], \\ \bar{M}(\boldsymbol{\theta})_{\text{total}} &= n^{-1} \mathbf{t}_n^\top \mathbf{S}_r(\boldsymbol{\theta}) \mathbf{t}_n, \\ \bar{M}(\boldsymbol{\theta})_{\text{indirect}} &= \bar{M}(r)_{\text{total}} - \bar{M}(r)_{\text{direct}}, \end{aligned}$$

which are highly nonlinear due to $\mathbf{S}_r(\boldsymbol{\theta})$.⁵ Therefore, a procedure such as the Delta Method is difficult to perform. Instead, we can use a Monte Carlo approximation which takes into account the sampling distribution of $\boldsymbol{\theta}$. To show this procedure, consider the SDM where:

$$\mathbf{S}(\boldsymbol{\theta})_r = (\mathbf{I}_n - \rho \mathbf{W})^{-1} (\mathbf{I}_n \boldsymbol{\beta}_r + \mathbf{W} \boldsymbol{\gamma}_r).$$

⁵Note that we have replaced the parameter for the spatially lagged independent variable to let $\boldsymbol{\theta}$ be the vector parameters of the model.

Let $g(\boldsymbol{\theta}) = \bar{M}(\boldsymbol{\theta})$ be a function representing the marginal (direct, indirect or total) effect that depends on the population parameters $\boldsymbol{\theta}$. If $N(\boldsymbol{\theta}|\bar{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\theta})$ denotes the multivariate normal density of $\boldsymbol{\theta}$ with mean $\bar{\boldsymbol{\theta}}$ and asymptotic variance-covariance matrix $\boldsymbol{\Sigma}_{\theta}$, then the expected value of the marginal effects conditional on the population parameters $\bar{\boldsymbol{\theta}}$ and $\boldsymbol{\Sigma}_{\theta}$ is:

$$\mathbb{E} [g(\boldsymbol{\theta})|\bar{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\theta}] = \int_{\boldsymbol{\theta}} \mathbb{E} [g(\boldsymbol{\theta})|\mathbf{y}, \mathbf{X}, \boldsymbol{\theta}] N(\boldsymbol{\theta}|\bar{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\theta}) d\boldsymbol{\theta}.$$

A Monte Carlo approximation to this expectation is obtained by calculation of the empirical marginal effects evaluated at pseudo draws of $\boldsymbol{\theta}$ from the asymptotic distribution of the estimator. The algorithm is the following:

Algorithm 4.14 — Standard Errors of the Marginal Effects. Estimate the model using MLE. Consider $s = 1, \dots, S$, and start with $s = 1$

- (a) Take a random draw of $\boldsymbol{\theta}^s$ from $N(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\Sigma}}_{\theta})$, which is the estimated asymptotic distribution of $\hat{\boldsymbol{\theta}}$.
- (b) Compute the marginal effect, but substituting $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}^s$.
- (c) Update $s = s + 1$, and go back to step 1.
- (d) Repeat for a large number of repetitions S (e.g., $S = 1000$).
- (e) Calculate the empirical mean of the marginal effects. The standard error of the marginal effect across the S draws is the standard error.

4.6 Spillover Effects on Crime: An Application in R

4.6.1 Estimation of Spatial Models in R

This example uses the dataset from [Anselin \(1988\)](#), which provides a cross-sectional view of 49 neighborhoods in Columbus, Ohio. The goal is to explain the crime rate as a function of household income and housing values. The dataset includes the following variables:

- CRIME: residential burglaries and vehicle thefts per thousand household in the neighborhood.
- HOVAL: housing value in US\$1,000.
- INC: household income in US\$1,000.

We begin the analysis by loading the necessary R packages into the workspace:

```
# Load packages
library("spdep")
library("spatialreg")
library("memisc")           # Package for tables
library("maptools")
```

```
## Error in library("maptools"): there is no package called 'maptools'

library("RColorBrewer")
library("classInt")
source("getSummary.sarlm.R") # Function for spdep models
```

The dataset is available in the **spdep** package, and we load it as follows:

```
# Load data
columbus <- readShapePoly("columbus.shp")

## Error in readShapePoly("columbus.shp"): could not find function "readShapePoly"

col.gal.nb <- read.gal(system.file("etc/weights/columbus.gal",
                                   package = "spdep")[1])
```

Before estimating models, it is essential to determine whether there is a spatial pattern in the crime variable. To visualize this, we create a quantile choropleth map for **CRIME**:

```
# 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 = "")
```

The map in Figure 4.3 suggests a positive spatial autocorrelation in crime rates. To confirm this, we perform Moran's I test using a row-normalized binary contiguity matrix, `col.gal.nb`, based on the Queen criterion. The Moran test is performed using Monte Carlo simulations with 99 replications:

Figure 4.3: Spatial Distribution of Crime in Columbus, Ohio Neighborhoods

```
## Error: unable to find an inherited method for function 'spplot'
for signature 'obj = "data.frame"'
```

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 Moran's I statistic is 0.51 with a p-value of 0.01, providing evidence of positive spatial autocorrelation. This indicates that neighborhoods with high (low) crime rates tend to be surrounded by neighborhoods with similarly high (low) crime rates.

Next, we estimate various spatial models using functions from the **spatialreg** package. These models include:

- **OLS**: Estimated using the `lm` function.
- **SLX**: Estimated with the `lm` function, using the `lag.listw` function from **spdep** to create WX . Alternatively, the `lmSLX` function from **spatialreg** can be used.
- **SLM**: Estimated using the `lagsarlm` function from **spatialreg**.
- **SDM**: Estimated with `lagsarlm` using the `type = "mixed"` argument. Alternatively, `type = "Durbin"` may be specified.
- **SEM**: Estimated using the `errorsarlm` function. The Spatial Durbin Error Model (SDEM) can be estimated with `type = "emixed"`.
- **SAC**: Estimated with the `sacsarlm` function.

These models are estimated using Maximum Likelihood (ML) methods, as outlined in the previous section. To compute the determinant of the Jacobian, we follow the approach of Ord (1975) and explicitly set `method = "eigen"` in the spatial model functions, ensuring consistency with Equation (4.39).

```
# 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 results of the estimations are presented in Table 4.1. Column 1 reports the OLS estimates. The findings suggest that, on average, an increase of one thousand dollars in neighborhood income is associated with a reduction of 1.6 crimes per thousand households. Similarly, an increase of one thousand dollars in housing value is associated with a reduction of 0.3 crimes per thousand households. Both coefficients are statistically significant.⁶ These results imply that residential burglaries and vehicle thefts are less prevalent in wealthier neighborhoods.

Column 2 shows the results for the SLX model, specified as

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

where $\mathbf{W}\mathbf{X}$ is a 49×2 matrix representing the spatial lags of `INC` and `HOVAL`. The coefficient for the spatial lag of income (`W.INC`) is negative and statistically significant, indicating that crime in a given neighborhood is inversely related to the income levels of its neighboring areas. This suggests that higher income among neighbors is associated with lower crime levels in the focal neighborhood. Conversely, the coefficient for the spatial lag of housing value (`W.HOVAL`) is positive but not statistically significant, implying no clear relationship between the housing values of neighboring areas and crime.

Column 3 presents the results for the Spatial Lag Model (SLM). The spatial autoregressive parameter ρ is positive and significant, providing strong evidence of spatial autocorrelation and spillover effects in crime. The coefficients for the explanatory variables are qualitatively similar to the OLS results, though smaller in absolute magnitude, reflecting the influence of incorporating spatial dependence.

The Spatial Durbin Model (SDM) results are reported in column 4. The estimated ρ parameter remains positive and significant, confirming the presence of endogenous spatial interaction effects. However, the coefficients for the spatially lagged explanatory variables are

⁶We use the term "associated" because potential endogeneity issues in either variable could affect causal interpretation.

not statistically significant. This suggests that once endogenous interaction effects in crime are accounted for, the socio-economic characteristics of neighboring areas do not significantly influence crime in the focal neighborhood. Furthermore, the coefficient for the spatial lag of income exhibits an unexpected positive sign, which contradicts the common factor hypothesis. This discrepancy suggests that the primary spatial effect may stem from an omitted spatial lag variable rather than spatial dependence in the error term.

Column 5 contains the results for the Spatial Error Model (SEM). The autoregressive parameter for $\mathbf{W}u$ is positive and significant, indicating substantial spatial transmission of random shocks. This finding may reflect the omission of spatially correlated variables, which contributes to the propagation of unobserved shocks across neighboring areas.

Finally, the Spatial Autoregressive Combined (SAC) model results are shown in column 6. This model accounts for both endogenous interaction effects and interactions among error terms. The coefficients for $\mathbf{W}y$ and $\mathbf{W}u$ are not statistically significant when estimated jointly. However, when these interaction effects are separated, both coefficients become significant. This suggests that the SAC model may be overparameterized, leading to reduced statistical power and diminished significance levels for all variables.

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 extend the analysis from Section 2.5 by integrating the estimation of marginal effects using a real-world application in R.

We begin by addressing the following question: What would happen to crime rates across all regions if income increased from 13.906 to 14.906 in the 30th region ($\Delta \text{INC} = 1$)? This is analogous to the question posed in the commuting-time example in the previous chapter.

Following the approach in Section 2.5, we use the reduced-form predictor defined by the formula:

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

to estimate the predicted values before and after the change in the income variable. Using the observed values of the exogenous variables and the reduced-form predictor, we compute the predicted values for CRIME, denoted as $\hat{\mathbf{y}}^1$, based on the previously estimated SLM model.

```
# 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, $\hat{\mathbf{y}}^2$, as follows:

```
# 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

## Error in 'tmp'@data: no applicable method for '@' applied to an object of
class "data.frame"

# 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.      :0
## 1st Qu.:0
## Median :0
## Mean     :0
## 3rd Qu.:0
## Max.     :0

sum(delta_y)

## [1] 0
```

According to the results from `sum(delta_y)`, the predicted effect of this income increase is a decrease of 1.65 crimes per thousand households, accounting for both direct and indirect effects. In other words, increasing income by \$1,000 in region 30 leads to a system-wide adjustment, resulting in a new equilibrium where total crime decreases by approximately 1.7 crimes per thousand households.

Sometimes it is useful to visualize these effects. For instance, we might want to identify regions with high and low impacts due to the increase in INC. Let us define “highly impacted regions” as those where the crime rate decreases by more than 0.05. The following code generates Figure 4.4, which illustrates these regions:

```
# 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

```
## Error in findInterval(col_new$delta_y, breaks): 'vec' must be
sorted non-decreasingly and not contain NAs
## Error: object 'np' not found
## Error in (function (s, units = "user", cex = NULL, font = NULL,
vfont = NULL, : plot.new has not been called yet
## Error in plot.xy(xy.coords(x, y), type = type, ...): plot.new
has not been called yet
```

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 following code produces the map, with the resulting visualization shown 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")
```

Next, we use the `impacts()` function from the **spatialreg** package to decompose the total effects of a unit change in each predictor variable into direct (local) effects, indirect

Figure 4.5: Effects of a Change in Region 30: Magnitude

```
## Error in classIntervals(col_new$delta_y, n = 5, style = "jenks"):
single unique value
## Error: object 'cats5' not found
## Error: object 'colors5' not found
## Error: object 'cats5' not found
```

Notes: This graph shows the spatial distribution of the changes caused by altering INC in region 30.

(spillover) effects, and total effects. The `impacts()` function computes these measures using the reduced-form representation:

$$\mathbf{y} = \sum_{r=1}^K \mathbf{A}(\mathbf{W})^{-1} (\mathbf{I}_n \beta_r) + \mathbf{A}(\mathbf{W})^{-1} \boldsymbol{\epsilon}$$

$$\mathbf{A}(\mathbf{W})^{-1} = \mathbf{I}_n + \rho \mathbf{W} + \rho^2 \mathbf{W}^2 + \dots$$

If the spatial weights object (`listw`) is provided, the exact $\mathbf{A}(\mathbf{W})^{-1}$ is calculated. When traces are computed by powering sparse matrices, the approximation $\mathbf{I}_n + \rho \mathbf{W} + \rho^2 \mathbf{W}^2 + \dots$ is used. Both methods yield similar results, except when the number of powers used is very small or when the spatial coefficient ρ is close to its bounds.

```
spatialreg::impacts.Sarlm(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 indicates that a \$1,000 increase in income results in a total crime reduction of 1.8 crimes per thousand households. The direct effect of the income variable in the SLM model is -1.123, while the estimated coefficient is -1.074. This implies a feedback effect of: $-1.123 - (-1.074) = -0.049$, which accounts for 4.5% of the coefficient estimate. To corroborate these results, we can compute the impacts manually 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 obtained here are consistent with those computed using the `impact()` function. Additionally, we can calculate the p-values of the impacts by specifying the argument `R`, which determines the number of simulations used to generate distributions for the impact measures. This is possible when the fitted model object includes a coefficient covariance matrix. Below, we compute the impacts with p-values:

```
# Compute standard errors of impacts
im_obj <- spatialreg::impacts.Sarlm(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 indicate that the variable with the largest negative direct impact is `INC`, implying that `INC` has the strongest effect in reducing its own region's crime rate. The second column of the output presents the indirect effects, which measure the spatial spillovers caused by changes in each variable. Negative indirect effects can be interpreted as spatial benefits, as they reflect reductions in neighboring regions' crime rates. Conversely, positive indirect effects represent negative externalities, where increases in a variable result in higher crime

rates in neighboring regions. From the results, we observe that **INC** has the largest and most significant negative indirect effects, highlighting its substantial spatial spillover benefits.

On the other hand, the indirect effect for **HOVAL** is not statistically significant. This weak result may be attributed to the inherent rigidity of the spatial lag model (SLM), which assumes that the ratio of spillover effects to direct effects is the same for all explanatory variables. Such an assumption may limit the model's ability to accurately capture spillover dynamics.

The total effect combines both the direct and indirect impacts, offering a comprehensive view of each variable's importance in reducing crime rates. From the results, it is evident that **INC** also has the largest total effect, underscoring its overall significance.

To further investigate impacts, we follow an approach that converts the spatial weight matrix into a sparse format and computes its powers using the `trW` function:

```
# Impacts using traces.
W <- as(nb2listw(col.gal.nb, style = "W"), "CsparseMatrix")
trMC <- trW(W, type = "MC")
im <- spatialreg::impacts.Sarlm(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
```

Additionally, we can examine cumulative impacts by specifying the `Q` argument. When both `Q` and `tr` are provided in the `impacts()` function, the output includes impact components for each step in the traces of the powers of the weight matrix, up to and including the Q th power:

```

# Cumulative impacts
im2    <- spatialreg::impacts.Sarlm(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  0.0000000000
## 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 demonstrate how to create a custom function to estimate a Spatial Lag Model (SLM) using maximum likelihood (ML) estimation. Two approaches are considered. The first involves a constrained optimization procedure that directly uses the log-likelihood function in Equation (4.14). The second approach employs the concentrated log-likelihood function, following the steps outlined in Algorithm (4.1).

4.7.1 First approach

To estimate the SLM via maximum likelihood, we first define a function that computes the log-likelihood, its gradient, and its Hessian. We then use the `maxLik` function from the `maxLik` package (Henningsen and Toomet, 2011) to optimize these functions. Below is the implementation of 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(ll, 'hessian') <- H
}
return(ll)
}

```

The function `sml_ll` has the following arguments: `theta` is a vector log length $k + 2$, 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 of `matrix` class; the arguments `gradient` and `hessian` indicate whether the analytical gradient and Hessian, respectively, should be use in the numerical optimization algorithm.

Note that the function does not approximate the Jacobian matrix during computation. This may not be the most efficient method for large sample sizes. The log-likelihood object `ll` corresponds to Equation (4.14). The gradient and Hessian are implemented following Equations (4.31) and (4.38), respectively.

Now we define the main function, `sml.m1`, which estimates the SLM using MLE with constrained optimization via the `maxLik` function.

```

library("maxLik")
slm.ml <- function(formula, data, W,
                   gradient = TRUE,
                   hessian = TRUE, ...){
  require("maxLik")
  # 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
  K <- ncol(X)

  # Starting values
  ols.init  <- lm(y ~ X - 1)
  b.init    <- coef(ols.init)
  sigma2.init <- sum(residuals(ols.init)^2) / ols.init$df.residual
  rho.init  <- cor(W %*% y, y)
  start     <- c(b.init, rho.init, sigma2.init)
  names(start) <- c(colnames(X), "rho", "sig.sq")

  # Optimization default controls if not added by user
  if (is.null(callT$method)) callT$method <- 'bfgs'
  if (is.null(callT$iterlim)) callT$iterlim <- 100000

  # Restricted optimization if BFGS: A %*% theta + B >= 0: Constraint rho and sigma2
  if (callT$method == "bfgs"){
    sym      <- all(W == t(W))
    omega    <- eigen(W, only.values = TRUE, symmetric = sym)
    lambda_space <- if (is.complex(omega$values)) 1 / range(Re(omega$values)) else 1 / r
    A <- rbind(c(rep(0, K), 1, 0),
               c(rep(0, K), -1, 0),
               c(rep(0, K), 0, 1))
    B <- c(-1L * (lambda_space[1] + sqrt(.Machine$double.eps)),
           lambda_space[2] - sqrt(.Machine$double.eps),
           -1L * sqrt(.Machine$double.eps))
    callT$constraints <- list(ineqA = A, ineqB = B)
  }
}

```



```

# Optimization
opt <- callT
m <- match(c('method', 'print.level', 'iterlim',
            'tol', 'ftol', 'steptol', 'fixed', 'constraints',
            'control', 'finalHessian', 'reltol', 'rho',
            'outer.iterations', 'outer.eps'),
          names(opt), OL)
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)
}

```

The procedure uses numerical optimization via the `maxLik` function. Initial values for $\hat{\beta}$ and $\hat{\sigma}^2$ are derived from the OLS estimator, while the starting value for $\hat{\rho}$ is based on the simple correlation between $\mathbf{W}\mathbf{y}$ and \mathbf{y} . By default, the optimization procedure is based on BFGS algorithm with inequality constraints. These constraints ensure that $\rho < 1/\omega_{\max} - e$, $\rho > 1/\omega_{\min} + e$, and $\sigma^2 > e$. The value e is a very small number, defined as `sqrt(.Machine$double.eps)`.

In matrix form, the inequality constraints are expressed as:

$$\mathbf{A}\boldsymbol{\theta} + \mathbf{b} \geq \mathbf{0},$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \rho, \sigma^2)^\top$, and

$$\mathbf{A} = \begin{pmatrix} \mathbf{0}^\top & 1 & 0 \\ \mathbf{0}^\top & -1 & 0 \\ \mathbf{0}^\top & 0 & 1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} -(1/\omega_{\min} + e) \\ 1/\omega_{\max} - e \\ -e \end{pmatrix},$$

where $\mathbf{0}$ is an $k \times 1$ vector of zeros. This implies:

$$\begin{aligned} \rho &\geq 1/\omega_{\min} + e, \\ -\rho &\geq -(1/\omega_{\max} - e), \\ \sigma^2 &\geq e. \end{aligned}$$

Currently, only the BFGS optimizer supports inequality constraints. If an alternative optimization method such as Newton-Raphson (specified via `method = "nr"`) is used, the procedure will not enforce these constraints, and the optimization will proceed globally.

The function outputs an object of class `maxLik`, making it compatible with existing methods in the `maxLik` package without requiring additional S3 methods.

To evaluate the function, we create an artificial dataset based on a data-generating process (DGP) inspired by [Lee \(2007\)](#). The DGP generates spatially lagged data for testing the ML implementation. The following lines demonstrate the function's application:

```
# Generate DGP
set.seed(1)
n      <- 529
rho    <- 0.6
W.nb2  <- cell2nb(sqrt(n), sqrt(n))
W      <- nb2mat(W.nb2)

# Exogenous variables
x1     <- rnorm(n)
x2     <- rnorm(n)
x3     <- rnorm(n)

# DGP parameters
b0 <- 0 ; b1 <- -1; b2 <- 0; b3 <- 1
sigma2 <- 2
epsilon <- rnorm(n, mean = 0, sd = sqrt(sigma2))

# Simulate the dependent variable
y <- solve(diag(n) - rho * W) %*% (b0 + b1*x1 + b2*x2 + b3*x3 + epsilon)

# Data as data.frame
data <- as.data.frame(cbind(y, x1, x2, x3))
names(data) <- c("y", "x1", "x2", "x3")
```

Two models are tested: one using the MLE approach with inequality constraints, and the other using the Newton-Raphson algorithm without constraints:

```
# Use our function
start <- Sys.time()
sml.mle <- slm.ml(y ~ x1 + x2 + x3, data = data, W = W)
summary(sml.mle)

## -----
## Maximum Likelihood estimation
## BFGS maximization, 48 iterations
## Return code 0: successful convergence
## Log-Likelihood: -999.8243
## 6 free parameters
## Estimates:
##           Estimate Std. error t value Pr(> t)
## (Intercept) -0.08499    0.06709  -1.267   0.205
## x1          -1.07243    0.06643 -16.143 <2e-16 ***
## x2           0.01593    0.06321   0.252   0.801
```

```
## x3          0.99461      0.06664    14.926    <2e-16 ***
## rho         0.56068      0.03737    15.004    <2e-16 ***
## sig.sq      2.34032      0.14742    15.875    <2e-16 ***
## ---
## 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 0.0001031368
## -----

print(Sys.time()- start)

## Time difference of 24.56653 secs

start <- Sys.time()
sml.mle.nr <- slm.ml(y ~ x1 + x2 + x3, data = data, W = W, method = "nr")

## Warning in log(2 * pi * sig.sq): NaNs produced

summary(sml.mle.nr)

## -----
## Maximum Likelihood estimation
## Newton-Raphson maximisation, 5 iterations
## Return code 1: gradient close to zero (gradtol)
## Log-Likelihood: -999.8243
## 6 free parameters
## Estimates:
##          Estimate Std. error t value Pr(> t)
## (Intercept) -0.08495    0.06709  -1.266   0.205
## x1          -1.07245    0.06643 -16.143 <2e-16 ***
## x2           0.01591    0.06321   0.252   0.801
## x3           0.99457    0.06664  14.925 <2e-16 ***
## rho          0.56068    0.03737  15.004 <2e-16 ***
## sig.sq       2.34036    0.14742  15.875 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## -----

print(Sys.time()- start)

## Time difference of 3.850282 secs
```

Both optimization procedures yield similar estimates and standard errors, but the Newton-Raphson algorithm is faster. However, the second approach produces a warning due to a negative value of σ^2 in one of the iterations.

4.7.2 Second approach

Now, we create a function that estimates the parameters of the SLM using the concentrated log-likelihood and the steps in Algorithm (4.1).

The concentrated log-likelihood function is created as follows:

```
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)
}
```

The `logLik_sar` function uses the Ord's approximation for the Jacobian.

The main function is the following:

```
sar.mle.con <- function(formula, data, W)
{
  # 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 <- W %*% y                     # Spatial lag

  # 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
sym      <- all(W == t(W))
omega    <- eigen(W, only.values = TRUE, symmetric = sym)

# Maximize concentrated log-likelihood
rho_space <- if (is.complex(omega$values)) 1 / range(Re(omega$values)) else 1 / range(Im(omega$values))
opt_lc <- optimize(f = logLik_sar, # This function is below
                  lower = rho_space[1] + .Machine$double.eps,
                  upper = rho_space[2] - .Machine$double.eps,
                  maximum = TRUE,
                  e_0 = e_0, e_L = e_L, omega = omega$values, n = n,
                  tol = .Machine$double.eps)

# 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)
error  <- Ay - crossprod(t(X), beta_hat)
sigma2_hat <- crossprod(error) / n

# Save results
out <- structure(
  list(
    callT = callT,
    rho_hat = rho_hat,
    beta_hat = beta_hat,
    sigma2_hat = sigma2_hat,
    A = A,
    W = W,
    X = X,
    omega = omega
  ),
  class = "slmc.mle"
)

return(out)
}

```

The following code creates the S3 method

```

vcov.slmc.mle <- function(object, ...){
  rho_hat <- object$rho_hat
  beta_hat <- object$beta_hat
  sigma2_hat <- object$sigma2_hat

```

```

A      <- object$A
W      <- object$W
X      <- object$X
omega  <- object$omega$values
k      <- ncol(X)

# 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)

return(solve(Hessian))
}

```

```

# S3 methods for summary
summary.slmc.mle <- function(object,
                             table = TRUE,
                             digits = max(3, .Options$digits - 3),
                             ...){
  X      <- object$X
  n      <- nrow(X)
  k      <- ncol(X)
  df     <- n - (k + 1)
  b      <- c(object$beta_hat, object$sigma2_hat, object$rho_hat)
  names(b) <- c(colnames(X), "sigma2", "Wy")
  std.err <- sqrt(diag(vcov(object)))
  z      <- b / std.err
  p      <- 2 * pt(-abs(z), df = df)
  CoefTable <- cbind(b, std.err, z, p)
  colnames(CoefTable) <- c("Estimate", "Std.Error", "t-value", "Pr(>|t|)")
  result <- structure(
    list(
      CoefTable = CoefTable,
      digits    = digits,
      call      = object$call),
    class = 'summary.slmc.mle'
  )
}

```

```

    )
    return(result)
}

print.summary.slmc.mle <- function(x,
                                   digits = x$digits,
                                   na.print = "",
                                   symbolic.cor = p > 4,
                                   signif.stars = getOption("show.signif.stars"),
                                   ...)
{
  cat("\nCall:\n")
  cat(paste(deparse(x$call), sep = "\n", collapse = "\n"), "\n\n", sep = "")

  cat("\nCoefficients:\n")
  printCoefmat(x$CoefTable, digit = digits, P.value = TRUE, has.Pvalue = TRUE)
  invisible(NULL)
}

```

```

slm2 <- sar.mle.con(y ~ x1 + x2 + x3, data = data, W = W)
summary(slm2)

##
## Call:
## sar.mle.con(formula = y ~ x1 + x2 + x3, data = data, W = W)
##
##
## Coefficients:
##              Estimate Std. Error t-value Pr(>|t|)
## (Intercept) -0.08495    0.06708  -1.266    0.206
## x1          -1.07245    0.06660 -16.103 <2e-16 ***
## x2           0.01591    0.06321   0.252    0.801
## x3           0.99457    0.06626  15.011 <2e-16 ***
## sigma2       2.34036    0.14736  15.881 <2e-16 ***
## Wy           0.56068    0.03706  15.129 <2e-16 ***
## ---
## 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 (4.29). Find the first and second derivative respect to ρ .

Exercise 4.2 Consider the Spatial Lag Model:

Table 4.2: Comparing coefficients for SLM.

	F-BFGS	F-NR	CMLE	R	F-BFGS	F-NR	CMLE	R
b0	-0.08499	-0.08495	-0.08495	-0.08495	0.06709	0.06709	0.06708	0.06708
b1	-1.07243	-1.07245	-1.07245	-1.07245	0.06643	0.06643	0.06660	0.06660
b2	0.01593	0.01591	0.01591	0.01591	0.06321	0.06321	0.06321	0.06321
b3	0.99461	0.99457	0.99457	0.99457	0.06664	0.06664	0.06626	0.06626
rho	0.56068	0.56068	0.56068	0.56068	0.03737	0.03737	0.03706	0.03706
sigma2	2.34032	2.34036	2.34036	2.34036	0.14742	0.14742	0.14736	0.00000

$$\begin{aligned}\mathbf{y} &= \rho \mathbf{W} \mathbf{y} + \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon} &\sim \text{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)\end{aligned}$$

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:

$$\begin{aligned}\mathbf{y} &= \mathbf{X} \boldsymbol{\beta} + \mathbf{u} \\ \mathbf{u} &= \lambda \mathbf{W} \mathbf{u} + \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon} &\sim \text{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)\end{aligned}$$

- Show that the OLS estimates $\hat{\boldsymbol{\beta}}$ is unbiased, but inefficient.
- Derived the ML estimates.
- Derived the concentrated log-likelihood function.
- Derive the asymptotic variance-covariance matrix of the estimates given in Equation (4.47).

Exercise 4.4 Consider the following SAC model with heteroskedastic errors:

$$\mathbf{y} = \rho \mathbf{W}_1 \mathbf{y} + \mathbf{X} \boldsymbol{\beta} + \mathbf{u} \quad (4.57)$$

$$\mathbf{u} = \lambda \mathbf{W}_2 \mathbf{u} + \boldsymbol{\varepsilon} \quad (4.58)$$

$$\boldsymbol{\varepsilon} \sim \text{N}(\mathbf{0}, \boldsymbol{\Omega}) \quad (4.59)$$

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:

$$\mathbb{V}(\epsilon_i) = \sigma_i^2 = \mathbf{z}_i^\top \boldsymbol{\alpha} \quad (4.60)$$

or

$$\mathbb{V}(\epsilon_i) = \sigma_i^2 = \exp(\mathbf{z}_i^\top \boldsymbol{\alpha}) \quad (4.61)$$

or more general,

$$\mathbb{V}(\epsilon_i) = \sigma_i^2 = \mathbf{h}(\mathbf{z}_i^\top \boldsymbol{\alpha}) \quad (4.62)$$

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:

$$\boldsymbol{\Omega}_{ii} = \sigma_i^2 = \mathbf{h}_i(\mathbf{z}_i^\top \boldsymbol{\alpha}), \quad \mathbf{h}_i > 0 \quad (4.63)$$

Note that the model has $2 + K + P$ unknown parameters:

$$\boldsymbol{\theta} = (\rho, \boldsymbol{\beta}^\top, \lambda, \boldsymbol{\alpha}^\top)^\top. \quad (4.64)$$

- (a) Find the Log-likelihood function.
- (b) Find the first order conditions

Exercise 4.5 Consider the model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \rho_1 \mathbf{W}_1 \mathbf{y} + \rho_2 \mathbf{W}_2 \mathbf{y} + \mathbf{u}, \quad (4.65)$$

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:

$$\begin{aligned} y_i &= \alpha + \beta x_i + u_i \\ u_i &= \lambda \sum_{j=1}^n w_{ij} u_j + \epsilon_i \\ \epsilon_i &\sim N(0, 1) \end{aligned} \quad (4.66)$$

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

The consistency $\widehat{\boldsymbol{\theta}}_n$ can be established from the uniform convergence of the concentrated log-likelihood of ρ with the QMLE of $\boldsymbol{\beta}_0$ and σ_0^2 concentrated out, because one can have the QMLEs of $\boldsymbol{\beta}$ and σ^2 once ρ is given. Thus, we 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 $\boldsymbol{\beta}$ and σ^2 , $Q_n(\rho) = \max_{\boldsymbol{\beta}, \sigma^2} \mathbb{E}[\ell_n(\boldsymbol{\theta})]$.

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

$$\ell_n(\rho) = -\frac{n}{2} (\ln(2\pi) + 1) - \frac{n}{2} \ln [\tilde{\sigma}_n^2(\rho)] + \ln |\mathbf{A}_n|, \quad (4.67)$$

where

$$\begin{aligned} \tilde{\sigma}_n^2(\rho) &= \frac{1}{n} \left[\mathbf{A}_n \mathbf{y}_n - \mathbf{X}_n \widehat{\boldsymbol{\beta}}_n(\rho) \right]^\top \left[\mathbf{A}_n \mathbf{y}_n - \mathbf{X}_n \widehat{\boldsymbol{\beta}}_n(\rho) \right], \\ &= \frac{1}{n} \mathbf{y}_n^\top \mathbf{A}_n^\top (\mathbf{I}_n - \mathbf{P}_{nX}) \mathbf{A}_n \mathbf{y}_n, \\ &= \frac{1}{n} \mathbf{y}_n^\top \mathbf{A}_n^\top \mathbf{M}_n \mathbf{A}_n \mathbf{y}_n, \end{aligned} \quad (4.68)$$

with $\widehat{\boldsymbol{\beta}}_n(\rho)$ being the MLE of $\boldsymbol{\beta}_0$ which 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 and annihilator matrix, respectively.

The QMLE $\widehat{\rho}_n$ of ρ_0 is the maximizer of the concentrated likelihood in Equation (4.67). If $\widehat{\rho}_n$ is consistent, the consistency of $\widehat{\boldsymbol{\beta}}_n(\widehat{\rho}_n)$ and $\tilde{\sigma}_n^2(\widehat{\rho}_n)$ follows from their closed form formula. It is important to highlight that we do not need to have a compact parameter space for $\boldsymbol{\beta}_0$ and σ_0^2 .

The expectation of $\ell_n(\boldsymbol{\theta})$ is:

$$\mathbb{E}[\ell_n(\boldsymbol{\theta})] = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \boldsymbol{\varepsilon}_n) + \ln |\mathbf{A}_n|. \quad (4.69)$$

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

$$\begin{aligned} \mathbb{E}(\mathbf{y}_n) &= \boldsymbol{\mu}_y = \mathbf{A}_{n0}^{-1} \mathbf{X}_n \boldsymbol{\beta}_0, \\ \mathbb{V}(\mathbf{y}_n) &= \boldsymbol{\Sigma}_y = \sigma_0^2 \mathbf{A}_{n0}^{-1} (\mathbf{A}_{n0}^{-1})^\top, \\ \text{tr}(\mathbf{A}_n^\top(\rho) \mathbf{A}_n(\rho) \boldsymbol{\Sigma}_y) &= \sigma_0^2 \text{tr}[\mathbf{A}_n^\top \mathbf{A}_n \mathbf{A}_{n0}^{-1} (\mathbf{A}_{n0}^{-1})^\top], \\ &= \sigma_0^2 \text{tr}[\mathbf{D}_n(\rho_0, \rho)], \\ \boldsymbol{\mu}_y^\top \mathbf{A}_n^\top(\rho) \mathbf{A}_n(\rho) \boldsymbol{\mu}_y &= \boldsymbol{\beta}_0^\top \mathbf{X}_n^\top (\mathbf{A}_{n0}^{-1})^\top \mathbf{A}_n^\top \mathbf{A}_n \mathbf{A}_{n0}^{-1} \mathbf{X}_n \boldsymbol{\beta}_0, \\ &= \boldsymbol{\beta}_0^\top \mathbf{X}_n^\top \mathbf{D}_n(\rho_0, \rho) \mathbf{X}_n \boldsymbol{\beta}_0, \end{aligned} \tag{4.70}$$

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 (4.70)

$$\begin{aligned} \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \boldsymbol{\varepsilon}_n) &= \mathbb{E}[(\mathbf{A}_n \mathbf{y}_n - \mathbf{X}_n \boldsymbol{\beta})^\top (\mathbf{A}_n \mathbf{y}_n - \mathbf{X}_n \boldsymbol{\beta})], \\ &= \mathbb{E}[\mathbf{y}_n^\top \mathbf{A}_n^\top \mathbf{A}_n \mathbf{y}_n - 2\boldsymbol{\beta}^\top \mathbf{X}_n^\top \mathbf{A}_n \mathbf{y}_n + \boldsymbol{\beta}^\top \mathbf{X}_n^\top \mathbf{X}_n \boldsymbol{\beta}], \\ &= \mathbb{E}[\mathbf{y}_n^\top \mathbf{A}_n^\top \mathbf{A}_n \mathbf{y}_n] - \mathbb{E}[2\boldsymbol{\beta}^\top \mathbf{X}_n^\top \mathbf{A}_n \mathbf{y}_n] + \mathbb{E}[\boldsymbol{\beta}^\top \mathbf{X}_n^\top \mathbf{X}_n \boldsymbol{\beta}], \\ &= \text{tr}(\mathbf{A}_n^\top \mathbf{A}_n \boldsymbol{\Sigma}_y) + \boldsymbol{\mu}_y^\top \mathbf{A}_n^\top \mathbf{A}_n \boldsymbol{\mu}_y - 2\boldsymbol{\beta}^\top \mathbf{X}_n^\top \mathbf{A}_n \boldsymbol{\mu}_y + \boldsymbol{\beta}^\top \mathbf{X}_n^\top \mathbf{X}_n \boldsymbol{\beta}, \\ &= \sigma_0^2 \text{tr}[\mathbf{D}_n(\rho_0, \rho)] + \boldsymbol{\beta}_0^\top \mathbf{X}_n^\top \mathbf{D}_n(\rho_0, \rho) \mathbf{X}_n \boldsymbol{\beta}_0 - 2\boldsymbol{\beta}^\top \mathbf{X}_n^\top \mathbf{A}_n \mathbf{A}_{n0}^{-1} \mathbf{X}_n \boldsymbol{\beta}_0 + \boldsymbol{\beta}^\top \mathbf{X}_n^\top \mathbf{X}_n \boldsymbol{\beta}. \end{aligned}$$

Then, the FONC for the optimal solutions of $\mathbb{E}(\ell_n(\boldsymbol{\theta}))$ in Equation (4.69) 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

$$\begin{aligned} \sigma_n^{2*} &= \frac{1}{n} \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \boldsymbol{\varepsilon}_n), \\ &= \frac{1}{n} \mathbb{E}([\mathbf{A}_n \mathbf{y}_n - \mathbf{X}_n \boldsymbol{\beta}_n^*]^\top [\mathbf{A}_n \mathbf{y}_n - \mathbf{X}_n \boldsymbol{\beta}_n^*]), \\ &= \frac{1}{n} \{(\rho_0 - \rho)^2 (\mathbf{C}_n \mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{M}_n (\mathbf{C}_n \mathbf{X}_n \boldsymbol{\beta}_0) + \sigma_0^2 \text{tr}(\mathbf{D}_n(\rho_0, \rho))\}, \\ &= \frac{1}{n} \{(\rho_0 - \rho)^2 (\mathbf{C}_n \mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{M}_n (\mathbf{C}_n \mathbf{X}_n \boldsymbol{\beta}_0)\} + \sigma_n^2(\rho), \end{aligned} \tag{4.71}$$

where $\sigma_n^2(\lambda) = \frac{1}{n} \sigma_0^2 \text{tr}(\mathbf{D}_n(\rho_0, \rho))$.

Since $\boldsymbol{\beta}_n^*$ and σ_n^{2*} represent the maximum values of the expected value of the log-likelihood function, $Q_n(\rho) = \max_{\boldsymbol{\beta}, \sigma^2} \mathbb{E}(\ell_n(\boldsymbol{\theta}))$ is

$$Q_n(\rho) = -\frac{n}{2} (\ln(2\pi) + 1) - \frac{n}{2} \ln(\sigma_n^{2*}(\rho)) + \log |\mathbf{A}_n|.$$

and

$$\frac{1}{n} [\ell_n(\rho) - Q_n(\rho)] = -\frac{1}{2} [\ln(\tilde{\sigma}^2) - \ln(\sigma_n^{2*}(\rho))],$$

To prove that $\frac{1}{n} [\ell_n(\rho) - Q_n(\rho)] \xrightarrow{p} 0$, 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 (4.68), we can write

$$\begin{aligned} \tilde{\sigma}(\rho) &= \frac{1}{n} \mathbf{y}_n^\top \mathbf{A}_n^\top (\mathbf{I}_n - \mathbf{P}_{nX}) \mathbf{A}_n \mathbf{y}_n, \\ &= \frac{1}{n} ((\rho_0 - \rho) \mathbf{M}_n \mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0 + \mathbf{M}_n \mathbf{A}_n \mathbf{A}_{n0}^{-1} \boldsymbol{\varepsilon}_n)^\top ((\rho_0 - \rho) \mathbf{M}_n \mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0 + \mathbf{M}_n \mathbf{A}_n \mathbf{A}_{n0}^{-1} \boldsymbol{\varepsilon}_n), \\ &= \frac{1}{n} (\rho_0 - \rho)^2 (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{M}_n (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0) + 2(\rho_0 - \rho) \underbrace{\frac{1}{n} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{M}_n \mathbf{A}_n \mathbf{A}_{n0}^{-1} \boldsymbol{\varepsilon}_n}_{B_{n1}} + \\ &\quad \underbrace{\frac{1}{n} \boldsymbol{\varepsilon}_n^\top (\mathbf{A}_{n0}^\top)^{-1} \mathbf{A}_n^\top \mathbf{M}_n \mathbf{A}_n \mathbf{A}_{n0}^{-1} \boldsymbol{\varepsilon}_n}_{B_{n2}} \\ &= \sigma_n^{2*} + B_{n1} + (B_{n2} - \sigma_n^2(\lambda)), \end{aligned}$$

where the last equality comes from Equation (4.71).

It can be shown that

$$\begin{aligned} \frac{1}{n} \mathbf{X}_n^\top \mathbf{C}_n^\top \boldsymbol{\varepsilon}_n &\xrightarrow{p} \mathbf{0} \\ \frac{1}{n} \mathbf{X}_n^\top \mathbf{C}_n^\top \mathbf{C}_n \boldsymbol{\varepsilon}_n &\xrightarrow{p} \mathbf{0}. \end{aligned} \tag{4.72}$$

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

$$\begin{aligned} B_{n1} &= \frac{1}{n} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{M}_n \mathbf{A}_n \mathbf{A}_{n0}^{-1} \boldsymbol{\varepsilon}_n \\ &= \frac{1}{n} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{M}_n (\mathbf{I}_n + (\rho_0 - \rho) \mathbf{C}_n(\rho_0)) \boldsymbol{\varepsilon}_n \\ &= \frac{1}{n} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{M}_n \boldsymbol{\varepsilon}_n + (\rho_0 - \rho) \frac{1}{n} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{M}_n \mathbf{C}_n(\rho_0) \boldsymbol{\varepsilon}_n \end{aligned}$$

The first and second moments of the first element of B_{n1} are

$$\begin{aligned} \mathbb{E} \left[\frac{1}{n} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{M}_n \boldsymbol{\varepsilon}_n \right] &= \frac{1}{n} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{M}_n \mathbb{E}(\boldsymbol{\varepsilon}_n) = \mathbf{0} \\ \mathbb{V} \left[\frac{1}{n} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{M}_n \boldsymbol{\varepsilon}_n \right] &= \frac{1}{n^2} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{M}_n \mathbb{E}(\boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^\top) \mathbf{M}_n^\top (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0) = o(1) \end{aligned}$$

Since $(\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{M}_n$ is uniformly bounded, then by Chebyshev's inequality 3.B.4 $n^{-1} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{M}_n \boldsymbol{\varepsilon}_n \xrightarrow{p} \mathbf{0}$. A similar reasoning can be applied to the second element of element of B_{n1} . Then

$$B_{n1} = o_p(1) + o_p(1) = o_p(1) \quad \text{uniformly in } \rho \in \Gamma.$$

B_{n2} can be expanded as

$$\begin{aligned} B_{n2} - \sigma_n^2(\lambda) &= \frac{1}{n} \boldsymbol{\varepsilon}_n^\top (\mathbf{A}_{n0}^\top)^{-1} \mathbf{A}_n^\top \mathbf{M}_n \mathbf{A}_n \mathbf{A}_{n0}^{-1} \boldsymbol{\varepsilon}_n - \frac{1}{n} \sigma_0^2 \text{tr}(\mathbf{D}_n(\rho_0, \rho)) \\ &= \frac{1}{n} \boldsymbol{\varepsilon}_n^\top (\mathbf{A}_{n0}^\top)^{-1} \mathbf{A}_n^\top (\mathbf{I}_n - \mathbf{P}_{nX}) \mathbf{A}_n \mathbf{A}_{n0}^{-1} \boldsymbol{\varepsilon}_n - \frac{1}{n} \sigma_0^2 \text{tr}(\mathbf{D}_n(\rho_0, \rho)) \\ &= \left[\frac{1}{n} \boldsymbol{\varepsilon}_n^\top (\mathbf{A}_{n0}^\top)^{-1} \mathbf{A}_n^\top \mathbf{A}_n \mathbf{A}_{n0}^{-1} \boldsymbol{\varepsilon}_n - \frac{1}{n} \sigma_0^2 \text{tr}(\mathbf{D}_n(\rho_0, \rho)) \right] - \frac{1}{n} \boldsymbol{\varepsilon}_n^\top (\mathbf{A}_{n0}^\top)^{-1} \mathbf{A}_n^\top \mathbf{P}_{nX} \mathbf{A}_n \mathbf{A}_{n0}^{-1} \boldsymbol{\varepsilon}_n \end{aligned}$$

Using Lemma 3.25 to the first element of B_{n2} yields

$$\begin{aligned} \mathbb{E} \left[\frac{1}{n} \boldsymbol{\varepsilon}_n^\top (\mathbf{A}_{n0}^\top)^{-1} \mathbf{A}_n^\top \mathbf{A}_n \mathbf{A}_{n0}^{-1} \boldsymbol{\varepsilon}_n \right] &= \frac{1}{n} \sigma_0^2 \text{tr}(\mathbf{D}_n(\rho_0, \rho)) \\ \mathbb{V} \left[\frac{1}{n} \boldsymbol{\varepsilon}_n^\top (\mathbf{A}_{n0}^\top)^{-1} \mathbf{A}_n^\top \mathbf{A}_n \mathbf{A}_{n0}^{-1} \boldsymbol{\varepsilon}_n \right] &= \frac{1}{n^2} \sigma_0^4 \left[\text{tr} \left[(\mathbf{A}_{n0}^\top)^{-1} \mathbf{A}_n^\top \mathbf{A}_n \mathbf{A}_{n0}^{-1} (\mathbf{A}_{n0}^\top)^{-1} \mathbf{A}_n^\top \mathbf{A}_n \mathbf{A}_{n0}^{-1} \right] + \right. \\ &\quad \left. \text{tr} \left[(\mathbf{A}_{n0}^\top)^{-1} \mathbf{A}_n^\top \mathbf{A}_n \mathbf{A}_{n0}^{-1} \right]^2 \right] = o(1) \end{aligned}$$

Then by Theorem 3.5, $\frac{1}{n} \boldsymbol{\varepsilon}_n^\top (\mathbf{A}_{n0}^\top)^{-1} \mathbf{A}_n^\top \mathbf{A}_n \mathbf{A}_{n0}^{-1} \boldsymbol{\varepsilon}_n \xrightarrow{p} \frac{1}{n} \sigma_0^2 \text{tr}(\mathbf{D}_n(\rho_0, \rho))$. The second element can be written as

$$\frac{1}{n} \left(\underbrace{\frac{1}{\sqrt{n}} \mathbf{X}_n^\top \mathbf{A}_n \mathbf{A}_{n0}^{-1} \boldsymbol{\varepsilon}}_{O_p(1)} \right)^\top \left(\underbrace{\frac{\mathbf{X}_n^\top \mathbf{X}_n}{n}}_{O(1)} \right)^{-1} \left(\underbrace{\frac{1}{\sqrt{n}} \mathbf{X}_n^\top \mathbf{A}_n \mathbf{A}_{n0}^{-1} \boldsymbol{\varepsilon}}_{O_p(1)} \right) = n^{-1} O_p(n^0) = O_p(n^1) = o_p(1)$$

where we use the fact that if $Z_n = O(n^k)$ then $Z_n = o_p(n^{k+\delta})$ for all $\delta > 0$. Then $B_{n2} - \sigma_n^2(\lambda) = o_p(1)$ 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 $\epsilon > 0$, $\limsup_{n \rightarrow \infty} [\max_{\rho \in \bar{N}_\epsilon(\rho_0)} \frac{1}{n} Q_n(\rho) - \frac{1}{n} Q_n(\rho_0)] < 0$, 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:

$$\ln(\sigma_n^2(\rho)) \leq \ln \sigma_n^{2*}(\rho) \quad (4.73)$$

At $\rho_0, \sigma_n^{2*}(\rho) = \sigma_0^2$ (see Equation (4.71)). 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

$$\begin{aligned} \mathbf{y} &= \mathbf{A}_0^{-1} \mathbf{X} \boldsymbol{\beta}_0 + \mathbf{A}_0^{-1} \boldsymbol{\varepsilon}, \\ \mathbb{E}(\mathbf{y}) &= \boldsymbol{\mu}_u = \mathbf{A}_0^{-1} \mathbf{X} \boldsymbol{\beta}_0, \\ \mathbb{E}(\mathbf{y}^\top \mathbf{y}) &= \boldsymbol{\Sigma}_y = \mathbf{A}_0^{-1} \sigma_0^2 \mathbf{I}_n (\mathbf{A}_0^{-1})^\top, \\ \mathbb{E}(\mathbf{y} \mathbf{y}^\top) &= (\mathbf{A}_0^{-1} \mathbf{X} \boldsymbol{\beta}_0) (\mathbf{A}_0^{-1} \mathbf{X} \boldsymbol{\beta}_0)^\top + \mathbf{A}_0^{-1} \sigma_0^2 \mathbf{I}_n (\mathbf{A}_0^{-1})^\top. \end{aligned} \quad (4.74)$$

We now derive the most difficult expectations. From (4.34) and letting $\mathbf{C}_0 = \mathbf{W} \mathbf{A}_0^{-1}$

$$\begin{aligned} \mathbb{E} \left(\frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta} \partial \rho} \right) &= -\frac{1}{\sigma_0^2} \mathbf{X}^\top \mathbb{E}(\mathbf{W} \mathbf{y}) \\ &= -\frac{1}{\sigma_0^2} \mathbf{X}^\top \mathbf{W} \mathbf{A}_0^{-1} \mathbf{X} \boldsymbol{\beta}_0 \\ &= -\frac{1}{\sigma_0^2} \mathbf{X}^\top (\mathbf{C}_0 \mathbf{X} \boldsymbol{\beta}_0) \end{aligned} \quad (4.75)$$

For (4.35) we obtain:

$$\begin{aligned}
 \mathbb{E} \left(\frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial (\sigma^2)^2} \right) &= \mathbb{E} \left[\frac{n}{2(\sigma_0^2)^2} - \frac{1}{(\sigma_0^2)^3} \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} \right] \\
 &= \frac{n}{2\sigma_0^4} - \frac{1}{\sigma_0^6} \mathbb{E} [\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}] \\
 &= \frac{n}{2\sigma_0^4} - \frac{n}{\sigma_0^6} \sigma_0^2 \mathbf{I}_n \\
 &= -\frac{n}{2\sigma_0^4}
 \end{aligned} \tag{4.76}$$

From (4.36):

$$\begin{aligned}
 \mathbb{E} \left[\frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \sigma^2 \partial \rho} \right] &= \mathbb{E} \left[-\frac{\boldsymbol{\varepsilon}^\top \mathbf{W} \mathbf{y}}{\sigma_0^4} \right] \\
 &= -\frac{1}{\sigma_0^4} \mathbb{E} [\boldsymbol{\varepsilon}^\top \mathbf{W} (\mathbf{A}_0^{-1} \mathbf{X} \boldsymbol{\beta}_0 + \mathbf{A}_0^{-1} \boldsymbol{\varepsilon})] \\
 &= -\frac{1}{\sigma_0^4} \mathbb{E} [\boldsymbol{\varepsilon}^\top \mathbf{C}_0 \mathbf{X} \boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}^\top \mathbf{C}_0 \boldsymbol{\varepsilon}] \\
 &= -\frac{1}{\sigma_0^4} \mathbb{E} [\boldsymbol{\varepsilon}^\top \mathbf{C}_0 \boldsymbol{\varepsilon}] \\
 &= -\frac{1}{\sigma_0^4} \text{tr}(\mathbf{C}_0) \mathbb{E} (\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top) \\
 &= -\text{tr}(\mathbf{C}_0) / \sigma_0^2
 \end{aligned} \tag{4.77}$$

From (4.37):

$$\begin{aligned}
 \mathbb{E} \left[\frac{\partial^2 \ell(\boldsymbol{\theta}_0)}{\partial \rho^2} \right] &= \mathbb{E} \left[-\text{tr} [(\mathbf{W} \mathbf{A}_0^{-1})^2] - \frac{1}{\sigma_0^2} (\mathbf{y}^\top \mathbf{W}^\top \mathbf{W} \mathbf{y}) \right] \\
 &= -\text{tr} [(\mathbf{W} \mathbf{A}_0^{-1})^2] - \frac{1}{\sigma_0^2} \mathbb{E} [\mathbf{y}^\top \mathbf{W}^\top \mathbf{W} \mathbf{y}] \\
 &= -\text{tr}(\mathbf{C}_0^2) - \frac{1}{\sigma_0^2} (\text{tr}(\mathbf{W}^\top \mathbf{W} \boldsymbol{\Sigma}_y) + \boldsymbol{\mu}_y^\top \mathbf{W}^\top \mathbf{W} \boldsymbol{\mu}_y) \\
 &= -\text{tr}(\mathbf{C}_0^2) - \frac{1}{\sigma_0^2} (\sigma_0^2 \text{tr}(\mathbf{W}^\top \mathbf{W} \mathbf{A}_0^{-1} (\mathbf{A}_0^{-1})^\top) + (\mathbf{A}_0^{-1} \mathbf{X} \boldsymbol{\beta}_0)^\top \mathbf{W}^\top \mathbf{W} \mathbf{A}_0^{-1} \mathbf{X} \boldsymbol{\beta}_0) \\
 &= -\text{tr}(\mathbf{C}_0^2) - \frac{1}{\sigma_0^2} (\boldsymbol{\beta}_0^\top \mathbf{X}^\top \mathbf{C}_0^\top \mathbf{C}_0 \mathbf{X} \boldsymbol{\beta}_0 + \sigma_0^2 \text{tr}(\mathbf{C}_0^\top \mathbf{C}_0)) \\
 &= -\text{tr}(\mathbf{C}_0^2) - \frac{1}{\sigma_0^2} (\mathbf{C}_0 \mathbf{X} \boldsymbol{\beta}_0)^\top (\mathbf{C}_0 \mathbf{X} \boldsymbol{\beta}_0) - \text{tr}(\mathbf{C}_0^\top \mathbf{C}_0) \\
 &= -\text{tr}(\mathbf{C}_0^s \mathbf{C}_0) - \frac{1}{\sigma_0^2} (\mathbf{C}_0 \mathbf{X} \boldsymbol{\beta}_0)^\top (\mathbf{C}_0 \mathbf{X} \boldsymbol{\beta}_0)
 \end{aligned} \tag{4.78}$$

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:

$$-\mathbb{E} [\mathbf{H}(\boldsymbol{\theta}_0)] = \begin{pmatrix} \frac{1}{n\sigma_0^2} (\mathbf{X}^\top \mathbf{X}) & \frac{1}{\sigma_0^2} \mathbf{X}^\top (\mathbf{C} \mathbf{X} \boldsymbol{\beta}_0) & \mathbf{0}^\top \\ \text{tr}(\mathbf{C}_0^s \mathbf{C}_0) + \frac{1}{\sigma_0^2} (\mathbf{C}_0 \mathbf{X} \boldsymbol{\beta}_0)^\top (\mathbf{C}_0 \mathbf{X} \boldsymbol{\beta}_0) & \frac{1}{\sigma_0^2} \text{tr}(\mathbf{C}_0) & \frac{n}{2\sigma_0^4} \end{pmatrix} \tag{4.79}$$

By multiplying $-\mathbb{E}[\mathbf{H}(\boldsymbol{\theta}_0)]$ by $(1/n)$, we obtain Equation (4.52).

4.C Variance of the Score Function

In this Appendix, we will derive the variance of the score function. That is

$$\mathbb{V} \left[\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right] = \mathbb{E} \left[\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \cdot \frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^\top} \right] = \mathbf{J}_{\theta,n} = \boldsymbol{\Sigma}_{\theta,n} + \boldsymbol{\Omega}_{\theta,n} \quad (4.80)$$

where $\boldsymbol{\Sigma}_{\theta,n} = -\mathbb{E}[(1/n)\mathbf{H}(\boldsymbol{\theta}_0)]$ is given in Equation (4.52). The following results are important. For i.i.d $\boldsymbol{\epsilon}_n = (\epsilon_1, \dots, \epsilon_n)$, it can be shown that

$$\begin{aligned} \mathbb{E}(\epsilon_i \epsilon_j) &= \begin{cases} \sigma_0^2 & \text{for } i = j, \\ 0 & \text{for } i \neq j \end{cases} \\ \mathbb{E}(\epsilon_i \epsilon_j \epsilon_s) &= \begin{cases} \mu_3 & \text{for } i = j = s, \\ 0 & \text{otherwise} \end{cases} \\ \mathbb{E}(\epsilon_i \epsilon_j \epsilon_s \epsilon_t) &= \begin{cases} \mu_4 & \text{for } i = j = s = t, \\ \sigma_0^4 & \text{for } i = j \neq s = t \text{ or } i = s \neq j = t \text{ or } i = t \neq s = t, \\ 0 & \text{otherwise} \end{cases} \end{aligned} \quad (4.81)$$

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

$$\begin{aligned} \mathbb{E}(\boldsymbol{\epsilon}_n^\top \mathbf{A}_n \boldsymbol{\epsilon}_n) &= \mathbb{E} \left[\sum_{i=1}^n \sum_{j=1}^n a_{ij} \epsilon_i \epsilon_j \right] = \sum_{i=1}^n a_{ii} \mathbb{E}(\epsilon_i \epsilon_i) = \sigma_0^2 \sum_{i=1}^n a_{ii} = \sigma_0^2 \text{tr}(\mathbf{A}_n), \\ \mathbb{E}(\boldsymbol{\epsilon}_n \boldsymbol{\epsilon}_n^\top \mathbf{A}_n \boldsymbol{\epsilon}_n) &= \sum_{s=1}^n \sum_{i=1}^n \sum_{j=1}^n a_{ij} \mathbb{E}(\epsilon_i \epsilon_j \epsilon_s) = \sum_i a_{ii} \mathbb{E}(\epsilon_i^3), \\ &= \mu_3 \text{tr}(\mathbf{A}_n) \\ \mathbb{E}(\boldsymbol{\epsilon}_n^\top \mathbf{A}_n^\top \boldsymbol{\epsilon}_n \boldsymbol{\epsilon}_n^\top \mathbf{A}_n \boldsymbol{\epsilon}_n) &= \sum_i \sum_j \sum_s \sum_t a_{ij} a_{st} \mathbb{E}(\epsilon_i \epsilon_j \epsilon_s \epsilon_t), \\ &= (\mu_4 - 3\sigma_0^4) \sum_{i=1}^n a_{ii}^2 + \sigma_0^4 [\text{tr}^2(\mathbf{A}_n) + \text{tr}(\mathbf{A}_n \mathbf{A}_n^\top) + \text{tr}(\mathbf{A}_n^2)], \\ &= (\mu_4 - 3\sigma_0^4) \sum_{i=1}^n a_{ii}^2 + \sigma_0^4 [\text{tr}^2(\mathbf{A}_n) + \text{tr}(\mathbf{A}_n^s \mathbf{A}_n)] \end{aligned} \quad (4.82)$$

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

$$\begin{aligned} \mathbb{E} \left[\left(\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta}} \right) \left(\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta}} \right)^\top \right] &= \frac{1}{\sigma_0^4} \frac{1}{n} \mathbb{E}(\mathbf{X}_n^\top \boldsymbol{\epsilon} \boldsymbol{\epsilon}^\top \mathbf{X}_n), \\ &= \frac{1}{\sigma_0^2} \frac{1}{n} \mathbf{X}_n^\top \mathbf{X}_n. \end{aligned} \quad (4.83)$$

$$\begin{aligned}
 \mathbb{E} \left[\left(\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta}} \right) \left(\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \rho} \right)^\top \right] &= \frac{1}{\sigma_0^4} \frac{1}{n} \left[\mathbb{E}(\mathbf{X}_n^\top \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^\top (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)) + \mathbb{E}(\mathbf{X}_n^\top \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^\top \mathbf{C}_{n0} \boldsymbol{\varepsilon}_n) - \right. \\
 &\quad \left. \sigma_0^2 \mathbb{E}(\mathbf{X}_n^\top \boldsymbol{\varepsilon}_n \text{tr}(\mathbf{C}_{n0})) \right], \\
 &= \frac{1}{\sigma_0^4} \frac{1}{n} \left[\mathbf{X}_n^\top \mathbb{E}(\boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^\top) (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0) + \mathbf{X}_n^\top \mathbb{E}(\boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^\top \mathbf{C}_{n0} \boldsymbol{\varepsilon}_n) \right], \\
 &= \frac{1}{\sigma_0^2} \frac{1}{n} \mathbf{X}_n^\top (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0) + \frac{1}{\sigma_0^4} \frac{\mu_3}{n} \mathbf{X}_n^\top \text{diag}(\mathbf{C}_{n0}).
 \end{aligned} \tag{4.84}$$

$$\begin{aligned}
 \mathbb{E} \left[\left(\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta}} \right) \left(\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \sigma^2} \right)^\top \right] &= \frac{1}{\sigma_0^6} \frac{1}{n} \mathbb{E} \left[\frac{1}{2} \mathbf{X}_n^\top \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^\top \boldsymbol{\varepsilon}_n - \frac{n \sigma_0^2}{2} \mathbf{X}_n^\top \boldsymbol{\varepsilon}_n \right], \\
 &= \frac{\mu_3}{\sigma_0^6} \frac{1}{2n} \mathbf{X}_n^\top \mathbf{1}_n.
 \end{aligned} \tag{4.85}$$

$$\begin{aligned}
 \mathbb{E} \left[\left(\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \sigma^2} \right) \left(\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \rho} \right)^\top \right] &= \frac{1}{n \sigma_0^2} \text{tr}(\mathbf{C}_{n0}) + \frac{1}{2n \sigma_0^6} [(\mu_4 - 3\sigma_0^4) \text{tr}(\mathbf{C}_{n0}) + \mu_3 \mathbf{1}^\top (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)] \\
 \mathbb{E} \left[\left(\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \sigma^2} \right) \left(\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \sigma^2} \right)^\top \right] &= \frac{1}{2\sigma_0^4} + \frac{(\mu_4 - 3\sigma_0^4)}{4\sigma_0^8}
 \end{aligned} \tag{4.86}$$

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 (4.82)

$$\begin{aligned}
 \mathbb{E} [(\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^\top (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)] &= \sigma_0^2 (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0) \\
 \mathbb{E} [(\boldsymbol{\varepsilon}_n^\top \mathbf{C}_{n0} \boldsymbol{\varepsilon}_n)^2] &= (\mu_4 - 3\sigma_0^4) \sum_{i=1}^n c_{n,ii}^2 + \sigma_0^4 [\text{tr}^2(\mathbf{C}_{n0}) + \text{tr}(\mathbf{C}_{n0}^s \mathbf{C}_{n0})] \\
 \mathbb{E} [\sigma_0^4 \text{tr}^2(\mathbf{C}_{n0})] &= \sigma_0^4 \text{tr}^2(\mathbf{C}_{n0}) \\
 \mathbb{E} [2(\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^\top \mathbf{C}_{n0}^\top \boldsymbol{\varepsilon}_n] &= 2\mu_3 (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \text{diag}(\mathbf{C}_{n0}) \\
 \mathbb{E} [2\sigma_0^2 \text{tr}(\mathbf{C}_{n0}) (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \boldsymbol{\varepsilon}_n] &= \mathbf{0} \\
 \mathbb{E} [2\sigma_0^2 \text{tr}(\mathbf{C}_{n0}) \boldsymbol{\varepsilon}_n^\top \mathbf{C}_{n0} \boldsymbol{\varepsilon}_n] &= 2\sigma_0^4 \text{tr}^2(\mathbf{C}_{n0})
 \end{aligned}$$

Using these results, we obtain

$$\begin{aligned}
 \mathbb{E} \left[\left(\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \rho} \right) \left(\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \rho} \right)^\top \right] &= \frac{1}{\sigma_0^4 n} \left[\sigma_0^2 (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0) - \sigma_0^4 \text{tr}^2(\mathbf{C}_{n0}) \right. \\
 &\quad + (\mu_4 - 3\sigma_0^4) \sum_{i=1}^n c_{n,ii}^2 + \sigma_0^4 [\text{tr}^2(\mathbf{C}_{n0}) + \text{tr}(\mathbf{C}_{n0}^s \mathbf{C}_{n0})] \\
 &\quad \left. + 2\mu_3 (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \text{diag}(\mathbf{C}_{n0}) \right] \\
 &= \frac{1}{\sigma_0^2} \frac{1}{n} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0) + \frac{1}{n} \text{tr}(\mathbf{C}_{n0}^s \mathbf{C}_{n0}) \\
 &\quad + \frac{1}{\sigma_0^4} \frac{1}{n} (\mu_4 - 3\sigma_0^4) \sum_{i=1}^n c_{ii,0}^2 + \frac{1}{\sigma_0^4} \frac{2\mu_3}{n} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \text{diag}(\mathbf{C}_{n0})
 \end{aligned} \tag{4.87}$$

Then,

$$\mathbf{J}_{\theta,n} = \begin{pmatrix} \frac{1}{\sigma_0^2} \frac{1}{n} \mathbf{X}_n^\top \mathbf{X}_n & \frac{1}{n\sigma_0^2} \mathbf{X}_n^\top (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0) + \frac{\mu_3}{n\sigma_0^4} \mathbf{X}_n^\top \text{diag}(\mathbf{C}_{n0}) & \frac{\mu_3}{2n\sigma_0^6} \mathbf{X}_n^\top \mathbf{z}_n \\ * & \frac{1}{n\sigma_0^2} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0) + \frac{1}{n} \text{tr}(\mathbf{C}_{n0}^s \mathbf{C}_{n0}) + J_{22,n} & \frac{1}{n\sigma_0^2} \text{tr}(\mathbf{C}_{n0}) + J_{23,n} \\ * & * & \frac{1}{2\sigma_0^4} + \frac{(\mu_4 - 3\sigma_0^4)}{4\sigma_0^8} \end{pmatrix} \tag{4.88}$$

where

$$\begin{aligned}
 J_{22,n} &= \frac{1}{\sigma_0^4} \frac{1}{n} (\mu_4 - 3\sigma_0^4) \sum_{i=1}^n c_{ii,0}^2 + \frac{1}{\sigma_0^4} \frac{2\mu_3}{n} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \text{diag}(\mathbf{C}_{n0}), \\
 J_{23,n} &= \frac{1}{2n\sigma_0^6} [(\mu_4 - 3\sigma_0^4) \text{tr}(\mathbf{C}_{n0}) + \mu_3 \mathbf{z}^\top (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)].
 \end{aligned}$$

Thus, we can write $\mathbf{J}_{\theta,n} = \boldsymbol{\Sigma}_{\theta,n} + \boldsymbol{\Omega}_{\theta,n}$ with

$$\boldsymbol{\Omega}_{\theta,n} = \begin{pmatrix} \mathbf{0} & \frac{\mu_3}{n\sigma_0^4} \mathbf{X}_n^\top \text{diag}(\mathbf{C}_{n0}) & \frac{\mu_3}{2n\sigma_0^6} \mathbf{X}_n^\top \mathbf{z}_n \\ * & \frac{1}{n\sigma_0^4} (\mu_4 - 3\sigma_0^4) \sum_{i=1}^n c_{ii,0}^2 + \frac{2\mu_3}{n\sigma_0^4} (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top \text{diag}(\mathbf{C}_{n0}) & \frac{1}{2n\sigma_0^6} [(\mu_4 - 3\sigma_0^4) \text{tr}(\mathbf{C}_{n0}) + \mu_3 \mathbf{z}^\top (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)] \\ * & * & \frac{(\mu_4 - 3\sigma_0^4)}{4\sigma_0^8} \end{pmatrix}. \tag{4.89}$$

4.D Proof of Asymptotic Normality

From Equation (4.48), 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

$$\frac{1}{n} \frac{\partial^2 \ell_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \xrightarrow{p} \frac{1}{n} \frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}$$

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 $\mathbf{X}_n^\top \mathbf{X}_n / n = O(1)$ and $\tilde{\sigma}_n^2 \xrightarrow{p} \sigma_0^2$ from consistency, then from Equation (4.32), we have:⁷

$$\begin{aligned} \frac{1}{n} \frac{\partial^2 \ell_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} - \frac{1}{n} \frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} &= -\frac{1}{\tilde{\sigma}_n^2} \frac{(\mathbf{X}_n^\top \mathbf{X}_n)}{n} + \frac{1}{\sigma_0^2} \frac{(\mathbf{X}_n^\top \mathbf{X}_n)}{n}, \\ &= \underbrace{\left(\frac{1}{\sigma_0^2} - \frac{1}{\tilde{\sigma}_n^2} \right)}_{o_p(1)} \underbrace{\frac{(\mathbf{X}_n^\top \mathbf{X}_n)}{n}}_{O(1)}, \\ &= o_p(1) O(1), \\ &= o_p(1). \end{aligned}$$

It can also be shown that

$$\begin{aligned} \frac{1}{n} \mathbf{X}_n^\top \mathbf{W}_n^\top \mathbf{y}_n &= \frac{1}{n} \mathbf{X}_n^\top \mathbf{C}_n \mathbf{X}_n \boldsymbol{\beta}_0 + o_p(1) = O_p(1), \\ \frac{1}{n} \mathbf{y}_n^\top \mathbf{W}_n^\top \boldsymbol{\varepsilon}_n &= \frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{C}_n^\top \boldsymbol{\varepsilon}_n + o_p(1) = O_p(1), \\ \frac{1}{n} \mathbf{y}_n^\top \mathbf{W}_n^\top \mathbf{W}_n \mathbf{y}_n &= \frac{1}{n} (\mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{C}_n^\top \mathbf{C}_n \mathbf{X}_n \boldsymbol{\beta}_0 + \frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{C}_n^\top \mathbf{C}_n \boldsymbol{\varepsilon}_n + o_p(1) = O_p(1). \end{aligned} \tag{4.90}$$

It follows from Equation (4.34):

$$\begin{aligned} \frac{1}{n} \frac{\partial^2 \ell_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} - \frac{1}{n} \frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} &= -\frac{1}{\tilde{\sigma}_n^2} \frac{(\mathbf{X}_n^\top \mathbf{W}_n \mathbf{y}_n)}{n} + \frac{1}{\sigma_0^2} \frac{(\mathbf{X}_n^\top \mathbf{W}_n \mathbf{y}_n)}{n}, \\ &= \underbrace{\left(\frac{1}{\sigma_0^2} - \frac{1}{\tilde{\sigma}_n^2} \right)}_{o_p(1)} \underbrace{\frac{(\mathbf{X}_n^\top \mathbf{W}_n \mathbf{y}_n)}{n}}_{O_p(1)}, \\ &= o_p(1) O_p(1) = o_p(1). \end{aligned}$$

The residuals as function of the true error term is:

$$\begin{aligned} \boldsymbol{\varepsilon}(\tilde{\boldsymbol{\delta}}_n) &= \mathbf{y}_n - \mathbf{X}_n \tilde{\boldsymbol{\beta}}_n - \tilde{\rho}_n \mathbf{W}_n \mathbf{y}_n + (\boldsymbol{\varepsilon}(\boldsymbol{\delta}_0) - \boldsymbol{\varepsilon}(\boldsymbol{\delta}_0)), \\ &= \mathbf{y}_n - \mathbf{X}_n \tilde{\boldsymbol{\beta}}_n - \tilde{\rho}_n \mathbf{W}_n \mathbf{y}_n - \mathbf{y}_n + \mathbf{X}_n \boldsymbol{\beta}_0 - \rho_0 \mathbf{W}_n \mathbf{y}_n + \boldsymbol{\varepsilon}_n(\boldsymbol{\delta}_0), \\ &= \mathbf{X}_n (\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}}_n) + (\rho_0 - \tilde{\rho}_n) \mathbf{W}_n \mathbf{y}_n + \boldsymbol{\varepsilon}_n(\boldsymbol{\delta}_0). \end{aligned} \tag{4.91}$$

Then, taking into account Equation (4.33) and using our result in Equation (4.91)

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.

yields:

$$\begin{aligned}
 \frac{1}{n} \frac{\partial^2 \ell_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\beta} \partial \sigma^2} - \frac{1}{n} \frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta} \partial \sigma^2} &= -\frac{1}{\tilde{\sigma}_n^4} \frac{\mathbf{X}_n^\top \boldsymbol{\varepsilon}(\tilde{\boldsymbol{\delta}}_n)}{n} + \frac{1}{\sigma_0^4} \frac{\mathbf{X}_n^\top \boldsymbol{\varepsilon}_n(\boldsymbol{\delta}_0)}{n}, \\
 &= -\frac{1}{\tilde{\sigma}_n^4} \frac{1}{n} \mathbf{X}_n^\top \left[\mathbf{X}_n(\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}}_n) + (\rho_0 - \tilde{\rho}_n) \mathbf{W}_n \mathbf{y}_n + \boldsymbol{\varepsilon}_n(\boldsymbol{\delta}_0) \right] + \\
 &\quad + \frac{1}{\sigma_0^4} \frac{\mathbf{X}_n^\top \boldsymbol{\varepsilon}_n(\boldsymbol{\delta}_0)}{n}, \\
 &= \frac{1}{\tilde{\sigma}_n^4} \frac{1}{n} \mathbf{X}_n^\top \mathbf{X}_n (\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}}_n) - \frac{1}{\tilde{\sigma}_n^4} \frac{1}{n} \mathbf{X}_n^\top \mathbf{W}_n \mathbf{y}_n (\rho_0 - \tilde{\rho}_n) \\
 &\quad - \frac{1}{\tilde{\sigma}_n^4} \frac{1}{n} \mathbf{X}_n^\top \boldsymbol{\varepsilon}_n(\boldsymbol{\delta}_0) + \frac{1}{\sigma_0^4} \frac{\mathbf{X}_n^\top \boldsymbol{\varepsilon}_n(\boldsymbol{\delta}_0)}{n}, \\
 &= \left(\frac{1}{\sigma_0^4} - \frac{1}{\tilde{\sigma}_n^4} \right) \frac{\mathbf{X}_n^\top \boldsymbol{\varepsilon}_n(\boldsymbol{\delta}_0)}{n} + \frac{\mathbf{X}_n^\top \mathbf{X}_n}{n \tilde{\sigma}_n^4} (\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}}_n) \\
 &\quad + \frac{\mathbf{W}_n^\top \mathbf{W}_n \mathbf{y}_n}{n \tilde{\sigma}_n^4} (\rho_0 - \tilde{\rho}_n), \\
 &= o_p(1) O_p(1) + O(1) o_p(1) + O_p(1) o_p(1), \\
 &= o_p(1).
 \end{aligned}$$

From Equation (4.37), we know that:

$$\frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \rho^2} = -\text{tr}[(\mathbf{C}_n(\rho))^2] - \frac{1}{\sigma^2} (\mathbf{y}_n^\top \mathbf{W}_n^\top \mathbf{W}_n \mathbf{y}_n) \quad \text{where} \quad \mathbf{C}_n(\rho) = \mathbf{W}_n \mathbf{A}_n(\rho)^{-1}.$$

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:

$$\begin{aligned}
 \frac{1}{n} \frac{\partial^2 \ell_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \rho^2} - \frac{1}{n} \frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial \rho^2} &= \underbrace{2 \frac{1}{n} \text{tr}[(\mathbf{C}_n(\bar{\rho}))^3]}_{O(1)} \underbrace{(\rho_0 - \tilde{\rho}_n)}_{o_p(1)} + \underbrace{\left(\frac{1}{\sigma_0^2} - \frac{1}{\tilde{\sigma}_n^2} \right)}_{o_p(1)} \underbrace{\frac{\mathbf{y}_n^\top \mathbf{W}_n^\top \mathbf{W}_n \mathbf{y}_n}{n}}_{O_p(1)}, \\
 &= o_p(1).
 \end{aligned}$$

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 (4.36):

$$\begin{aligned}
 \frac{1}{n} \frac{\partial^2 \ell_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \sigma^2 \partial \rho} - \frac{1}{n} \frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial \sigma^2 \partial \rho} &= -\frac{1}{\tilde{\sigma}^4} \frac{1}{n} \mathbf{y}_n^\top \mathbf{W}_n^\top \boldsymbol{\varepsilon}_n(\tilde{\boldsymbol{\delta}}_n) + \frac{1}{\sigma^4} \frac{1}{n} \mathbf{y}_n^\top \mathbf{W}_n^\top \boldsymbol{\varepsilon}_n(\boldsymbol{\delta}_n), \\
 &= -\frac{1}{\tilde{\sigma}^4} \frac{1}{n} \mathbf{y}_n^\top \mathbf{W}_n^\top \left[\mathbf{X}_n(\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}}_n) + (\rho_0 - \tilde{\rho}_n) \mathbf{W}_n \mathbf{y}_n + \boldsymbol{\varepsilon}_n(\boldsymbol{\delta}_0) \right] + \\
 &\quad \frac{1}{\sigma^4} \frac{1}{n} \mathbf{y}_n^\top \mathbf{W}_n^\top \boldsymbol{\varepsilon}_n(\boldsymbol{\delta}_n), \\
 &= -\frac{1}{\tilde{\sigma}^4} \frac{1}{n} \mathbf{y}_n^\top \mathbf{W}_n^\top \mathbf{X}_n(\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}}_n) + \frac{1}{\tilde{\sigma}^4} \frac{1}{n} \mathbf{y}_n^\top \mathbf{W}_n^\top \mathbf{W}_n \mathbf{y}_n (\tilde{\rho}_n - \rho_0) + \\
 &\quad \left(\frac{1}{\sigma^4} - \frac{1}{\tilde{\sigma}^4} \right) \frac{1}{n} \mathbf{y}_n^\top \mathbf{W}_n^\top \boldsymbol{\varepsilon}_n, \\
 &= o_p(1).
 \end{aligned}$$

Note the following:

$$\begin{aligned}
 \frac{1}{n} \boldsymbol{\varepsilon}(\tilde{\boldsymbol{\delta}})^\top \boldsymbol{\varepsilon}(\tilde{\boldsymbol{\delta}}) &= \left(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 \right)^\top \frac{\mathbf{X}_n^\top \mathbf{X}_n}{n} + (\tilde{\rho}_n - \rho_0)^2 \frac{\mathbf{y}_n^\top \mathbf{W}_n^\top \mathbf{W}_n \mathbf{y}_n}{n} + \frac{\boldsymbol{\varepsilon}_n^\top \boldsymbol{\varepsilon}_n}{n} + \\
 &\quad 2(\tilde{\rho}_n - \rho_0) \left(\tilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 \right)^\top \frac{\mathbf{X}_n^\top \mathbf{W}_n \mathbf{y}_n}{n} + 2 \left(\boldsymbol{\beta}_0 - \tilde{\boldsymbol{\beta}}_n \right)^\top \frac{\mathbf{X}_n^\top \boldsymbol{\varepsilon}_n}{n} + 2(\rho_0 - \tilde{\rho}_n) \frac{\mathbf{y}_n^\top \mathbf{W}_n^\top \boldsymbol{\varepsilon}_n}{n}, \\
 &= \frac{\boldsymbol{\varepsilon}_n^\top \boldsymbol{\varepsilon}_n}{n} + o_p(1).
 \end{aligned}$$

Finally, considering the second derivative in Equation (4.35)

$$\begin{aligned}
 \frac{1}{n} \frac{\partial^2 \ell_n(\tilde{\boldsymbol{\theta}})}{\partial (\sigma^2)^2} - \frac{1}{n} \frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial (\sigma^2)^2} &= \frac{1}{2(\tilde{\sigma}^2)^2} - \frac{1}{(\tilde{\sigma}^2)^3} \frac{1}{n} \boldsymbol{\varepsilon}(\tilde{\boldsymbol{\delta}})^\top \boldsymbol{\varepsilon}(\tilde{\boldsymbol{\delta}}) - \frac{1}{2(\sigma^2)^2} + \frac{1}{(\sigma^2)^3} \frac{1}{n} \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}, \\
 &= \frac{1}{2} \left(\frac{1}{(\tilde{\sigma}^2)^2} - \frac{1}{(\sigma^2)^2} \right) - \frac{1}{(\tilde{\sigma}^2)^3} \frac{1}{n} \boldsymbol{\varepsilon}(\tilde{\boldsymbol{\delta}})^\top \boldsymbol{\varepsilon}(\tilde{\boldsymbol{\delta}}) + \frac{1}{(\sigma^2)^3} \frac{1}{n} \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}, \\
 &= \frac{1}{2} \left(\frac{1}{(\tilde{\sigma}^2)^2} - \frac{1}{(\sigma^2)^2} \right) - \frac{1}{(\tilde{\sigma}^2)^3} \left(\frac{\boldsymbol{\varepsilon}_n^\top \boldsymbol{\varepsilon}_n}{n} + o_p(1) \right) + \frac{1}{(\sigma^2)^3} \frac{1}{n} \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon}, \\
 &= \frac{1}{2} \left(\frac{1}{(\tilde{\sigma}^2)^2} - \frac{1}{(\sigma^2)^2} \right) + \left(\frac{1}{(\sigma^2)^3} - \frac{1}{(\tilde{\sigma}^2)^3} \right) \frac{\boldsymbol{\varepsilon}_n^\top \boldsymbol{\varepsilon}_n}{n} + o_p(1), \\
 &= o_p(1).
 \end{aligned}$$

(c) Now, we need to show that:

$$\frac{1}{n} \frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \xrightarrow{p} \mathbb{E} \left[\frac{1}{n} \frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right]$$

From Section 4.2.3, we know that

$$\frac{1}{n} \frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = \begin{pmatrix} -\frac{1}{n\sigma_0^2} (\mathbf{X}_n^\top \mathbf{X}_n) & -\frac{1}{n\sigma_0^2} \mathbf{X}_n^\top \mathbf{W}_n \mathbf{y}_n & -\frac{1}{(n\sigma_0^2)^2} \mathbf{X}_n^\top \boldsymbol{\varepsilon}_n \\ -\frac{1}{n} \text{tr} [(\mathbf{W}_n \mathbf{A}_{n0}^{-1})^2] - \frac{1}{n\sigma_0^2} (\mathbf{y}_n^\top \mathbf{W}_n^\top \mathbf{W}_n \mathbf{y}_n) & -\frac{\boldsymbol{\varepsilon}_n^\top \mathbf{W}_n \mathbf{y}_n}{n\sigma_0^4} & -\frac{\boldsymbol{\varepsilon}_n^\top \mathbf{W}_n \boldsymbol{\varepsilon}_n}{n(\sigma_0^2)^3} \\ \frac{1}{2(\sigma_0^2)^2} - \frac{1}{n(\sigma_0^2)^3} \boldsymbol{\varepsilon}_n^\top \boldsymbol{\varepsilon}_n & & \end{pmatrix} \quad (4.92)$$

Then, using the results in Equation (4.90)

$$\begin{aligned}
 -\frac{1}{n\sigma_0^2}(\mathbf{X}_n^\top \mathbf{X}_n) &\xrightarrow{p} \mathbb{E} \left[\frac{1}{n} \frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \right] = -\frac{1}{\sigma_0^2} \frac{1}{n} (\mathbf{X}_n^\top \mathbf{X}_n) \\
 -\frac{1}{n\sigma_0^2} \mathbf{X}_n^\top \mathbf{W}_n \mathbf{y}_n &\xrightarrow{p} \mathbb{E} \left[\frac{1}{n} \frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta} \partial \rho} \right] = -\frac{1}{\sigma_0^2} \frac{1}{n} \mathbf{X}_n^\top \mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0 \\
 -\frac{1}{(n\sigma_0^2)^2} \mathbf{X}_n^\top \boldsymbol{\varepsilon}_n &\xrightarrow{p} \mathbb{E} \left[\frac{1}{n} \frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta} \partial \sigma^2} \right] = 0 \\
 -\frac{\text{tr}[(\mathbf{W}_n \mathbf{A}_{n0}^{-1})^2]}{n} - \frac{(\mathbf{y}_n^\top \mathbf{W}_n^\top \mathbf{W}_n \mathbf{y}_n)}{n\sigma_0^2} &\xrightarrow{p} \mathbb{E} \left[\frac{1}{n} \frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial \rho^2} \right] = -\frac{\text{tr}(\mathbf{C}_{n0}^s \mathbf{C}_{n0})}{n} - \frac{(\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)^\top (\mathbf{C}_{n0} \mathbf{X}_n \boldsymbol{\beta}_0)}{n\sigma_0^2} \\
 -\frac{\boldsymbol{\varepsilon}_n^\top \mathbf{W}_n \mathbf{y}_n}{n\sigma_0^4} &\xrightarrow{p} \mathbb{E} \left[\frac{1}{n} \frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial \rho \partial \sigma^2} \right] = -\frac{1}{n} \text{tr}(\mathbf{C}_{n0}) / \sigma_0^2 \\
 \frac{1}{2(\sigma_0^2)^2} - \frac{1}{n(\sigma_0^2)^3} \boldsymbol{\varepsilon}_n^\top \boldsymbol{\varepsilon}_n &\xrightarrow{p} \mathbb{E} \left[\frac{1}{n} \frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial (\sigma_0^2)^2} \right] = -\frac{1}{2\sigma_0^4}
 \end{aligned} \tag{4.93}$$

All these expectations exist in the limit by Assumption 4.11 and Lemma 3.27. Then, by nonsingularity we can say that

$$\left[\frac{1}{n} \frac{\partial^2 \ell_n(\tilde{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right]^{-1} \xrightarrow{p} \mathbb{E} \left[\frac{1}{n} \frac{\partial^2 \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right]^{-1}.$$

(d) Recall that the first-order derivatives of the log-likelihood function at $\boldsymbol{\theta}_0$ are given by:

$$\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{1}{\sigma_0^2 \sqrt{n}} \mathbf{X}_n^\top \boldsymbol{\varepsilon}_n \\ \frac{1}{2\sigma_0^4 \sqrt{n}} (\boldsymbol{\varepsilon}_n' \boldsymbol{\varepsilon}_n - n\sigma_0^2) \\ \frac{1}{\sigma_0^2 \sqrt{n}} (\mathbf{C}_n \mathbf{X}_n \boldsymbol{\beta}_0)^\top \boldsymbol{\varepsilon}_n + \frac{1}{\sigma_0^2 \sqrt{n}} (\boldsymbol{\varepsilon}_n^\top \mathbf{C}_n \boldsymbol{\varepsilon}_n - \sigma_0^2 \text{tr}(\mathbf{C}_n)) \end{pmatrix}$$

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.

Then

$$\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \xrightarrow{d} \text{N}(\mathbf{0}, \boldsymbol{\Sigma}_0 + \boldsymbol{\Omega}_0)$$

where

$$\boldsymbol{\Sigma}_0 = \lim_{n \rightarrow \infty} \boldsymbol{\Sigma}_{n,0} \quad \text{and} \quad \boldsymbol{\Omega}_0 = \lim_{n \rightarrow \infty} \boldsymbol{\Omega}_{n,0} \tag{4.94}$$

Then, the proof is complete by using Slutsky's theorem.

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:

$$I = \left(\frac{n}{S_0} \right) \frac{\widehat{\boldsymbol{\varepsilon}}^\top \mathbf{W} \widehat{\boldsymbol{\varepsilon}}}{\widehat{\boldsymbol{\varepsilon}}^\top \widehat{\boldsymbol{\varepsilon}}}$$

where $\widehat{\boldsymbol{\varepsilon}}$ 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:

$$I = \frac{\widehat{\boldsymbol{\varepsilon}}^\top \mathbf{W} \widehat{\boldsymbol{\varepsilon}}}{\widehat{\boldsymbol{\varepsilon}}^\top \widehat{\boldsymbol{\varepsilon}}} \quad (5.1)$$

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:

$$I = \left(\frac{n}{S_0} \right) \frac{\widehat{\boldsymbol{\varepsilon}}^\top \mathbf{W} \widehat{\boldsymbol{\varepsilon}}}{\widehat{\boldsymbol{\varepsilon}}^\top \widehat{\boldsymbol{\varepsilon}}} \quad (5.2)$$

where $\widehat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ is a vector of OLS residuals, $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$, \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:

$$z_I = \frac{I - \mathbb{E}(I)}{\mathbb{V}(I)^{1/2}} \sim N(0, 1) \quad (5.4)$$

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 $\mathbf{W}\widehat{\boldsymbol{\varepsilon}}$ on $\widehat{\boldsymbol{\varepsilon}}$, rather than for $\widehat{\boldsymbol{\varepsilon}}$ on $\mathbf{W}\widehat{\boldsymbol{\varepsilon}}$, 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:

$$\bar{I} = \frac{\widehat{\boldsymbol{\varepsilon}}^\top \mathbf{W} \widehat{\boldsymbol{\varepsilon}}}{\tilde{\sigma}^2}, \quad (5.5)$$

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:

$$\tilde{\sigma}^2 = \frac{\widehat{\boldsymbol{\varepsilon}}^\top \widehat{\boldsymbol{\varepsilon}} \left\{ \text{tr} [(\mathbf{W}^\top + \mathbf{W}) \mathbf{W}] \right\}^{-1/2}}{n}.$$

As a consequence the test statistic can be defined as:

$$\bar{I} = \frac{n \widehat{\boldsymbol{\varepsilon}}^\top \mathbf{W} \widehat{\boldsymbol{\varepsilon}}}{\widehat{\boldsymbol{\varepsilon}}^\top \widehat{\boldsymbol{\varepsilon}} \left\{ \text{tr} [(\mathbf{W}^\top + \mathbf{W}) \mathbf{W}] \right\}^{-1/2}}. \quad (5.6)$$

The two expressions reported in Equations (5.2) and (5.6) coincide if the weight matrix has dichotomous entries in which case $w_{ij} = w_{ij}^2$ and, therefore,

$$\sum_i \sum_j w_{ij} = \left\{ \text{tr} [(\mathbf{W}^\top + \mathbf{W}) \mathbf{W}] \right\}^{-1/2}.$$

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("columbus.shp")

## Error in readShapePoly("columbus.shp"): could not find function "readShapePoly"
```

```
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:

$$\begin{aligned}
 \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + (\mathbf{I}_n - \lambda\mathbf{W})^{-1}\boldsymbol{\varepsilon} \\
 (\mathbf{I}_n - \lambda\mathbf{W})\mathbf{y} &= (\mathbf{I}_n - \lambda\mathbf{W})\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\
 \mathbf{y} - \lambda\mathbf{W}\mathbf{y} &= (\mathbf{X} - \lambda\mathbf{W}\mathbf{X})\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\
 \mathbf{y} &= \lambda\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} - \mathbf{W}\mathbf{X}(\lambda\boldsymbol{\beta}) + \boldsymbol{\varepsilon}
 \end{aligned} \tag{5.7}$$

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 (5.7) is equivalent to the SDM. The unconstrained form of the model—or the SDM model—is

$$\mathbf{y} = \gamma_1 \mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\gamma}_2 + \mathbf{W}\mathbf{X}\boldsymbol{\gamma}_3 + \boldsymbol{\varepsilon}, \tag{5.8}$$

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 $\gamma_3 = -\gamma_1\gamma_2$ is a vector of $K \times 1$ nonlinear constraints of the form:

$$\gamma_{3,k} = -\gamma_1\gamma_{2,k}, \quad \text{for } k = 1, \dots, K. \quad (5.9)$$

These conditions are usually formulated as a null hypothesis, designated as the **Common Factor Hypothesis**, and written as:

$$H_0 : \gamma_3 + \gamma_1\gamma_2 = \mathbf{0}. \quad (5.10)$$

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 $H_0 : \beta_{OLS} = \beta_{SEM}$ 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

$$\mathbb{V}(\hat{q}) = \mathbb{V}(\hat{\beta}_{OLS}) - \mathbb{V}(\hat{\beta}_{SEM}). \quad (5.11)$$

Then the Hausman statistic:

$$H = \hat{q}^\top (\mathbb{V}(\hat{q}))^{-1} \hat{q}, \quad (5.12)$$

is distributed asymptotically chi-square with $\# \beta$ degrees of freedom.

The estimated variance-covariance matrix $\hat{\beta}_{SEM}$ is given by (see Equation 4.47):

$$\mathbb{V}(\hat{\beta}_{SEM}) = \hat{\sigma}^2 \left[\mathbf{X}^\top (\mathbf{I}_n - \lambda \mathbf{W})^\top (\mathbf{I}_n - \lambda \mathbf{W}) \mathbf{X} \right]^{-1}. \quad (5.13)$$

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:

$$\begin{aligned}
\mathbb{V}(\hat{\beta}_{OLS}) &= \mathbb{E} [\hat{\beta} - \mathbb{E}(\hat{\beta})]^2 \\
&= \mathbb{E} [(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{B} \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top \mathbf{B}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}] \\
&= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{B} \mathbb{E} [\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top] \mathbf{B}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \\
&= \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{B} \mathbf{B}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}
\end{aligned} \tag{5.14}$$

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 (5.14) (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:

$$\tilde{\boldsymbol{\theta}}_{ML} = \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{ \frac{1}{n} \sum_{i=1}^n \log L(\boldsymbol{\theta}) \right\} \quad \text{s.t.} \quad \rho = 0 \tag{5.15}$$

or

$$\tilde{\boldsymbol{\theta}}_{ML} = \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{ \frac{1}{n} \sum_{i=1}^n \log L(\boldsymbol{\theta}) \right\} \quad \text{s.t.} \quad \lambda = 0 \quad (5.16)$$

and the unconstrained estimate as:

$$\hat{\boldsymbol{\theta}}_{ML} = \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\{ \frac{1}{n} \sum_{i=1}^n \log L(\boldsymbol{\theta}) \right\} \quad (5.17)$$

Definition 5.4.1 — Likelihood Ratio Test. The Likelihood Ratio (LR) Test is formally defined as:

$$LR = 2 \cdot n \left(\frac{1}{n} \sum_{i=1}^n \log L(\hat{\boldsymbol{\theta}}) - \frac{1}{n} \sum_{i=1}^n \log L(\tilde{\boldsymbol{\theta}}) \right) \xrightarrow{d} \chi^2(r) \quad (5.18)$$

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:

$$\log L(\boldsymbol{\theta}) = \log |\mathbf{A}| - \frac{n \log(2\pi)}{2} - \frac{n \log(\sigma^2)}{2} - \frac{1}{2\sigma^2} (\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (5.19)$$

The log-likelihood for the constrained model is found by setting $\rho = 0$ in Equation (5.19). Recall that if $\rho = 0$, then $\mathbf{A} = \mathbf{I} - \rho\mathbf{W} = \mathbf{I}$, then:

$$\log L(\boldsymbol{\theta}) = -\frac{n \log(2\pi)}{2} - \frac{n \log(\sigma^2)}{2} - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (5.20)$$

Therefore, following our definition in Equation (5.18):

$$\begin{aligned} LR &= 2(\log L(\hat{\boldsymbol{\theta}}) - \log L(\tilde{\boldsymbol{\theta}})) \\ &= 2 \left[\log |\mathbf{A}| + \frac{1}{2\sigma^2} ((\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - (\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta})) \right] \\ &= 2 \log |\mathbf{A}| + \frac{1}{\sigma^2} [(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - (\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{A}\mathbf{y} - \mathbf{X}\boldsymbol{\beta})] \end{aligned} \quad (5.21)$$

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:

$$\log L(\boldsymbol{\theta}) = \log |\mathbf{B}| - \frac{n \log(2\pi)}{2} - \frac{n \log(\sigma^2)}{2} - \frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Omega}(\lambda) (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (5.22)$$

Then the LR for the SEM model is:

$$LR = 2 \log |\mathbf{B}| + \frac{1}{\sigma^2} [(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Omega}(\lambda) (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})] \quad (5.23)$$

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 $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\theta}}$ 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 $(\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}})^2$ 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:

$$W = n \cdot \mathbf{r}(\hat{\boldsymbol{\theta}})^\top \left[\mathbf{R}(\hat{\boldsymbol{\theta}}) \hat{\mathbf{V}} \mathbf{R}(\hat{\boldsymbol{\theta}})^\top \right] \mathbf{r}(\hat{\boldsymbol{\theta}}) \xrightarrow{d} \chi^2(r) \quad (5.24)$$

where r is the number of constraints.

Wald Test for SLM

The W statistic is:

$$W_\rho = \frac{\hat{\rho}^2}{\hat{\mathbb{V}}(\rho)} \quad (5.25)$$

where $\hat{\mathbb{V}}(\rho)$ can be obtained from Equation ?? as:

$$\hat{\mathbb{V}}(\rho) = \left[\text{tr}(\mathbf{C}^s \mathbf{C}) + \frac{1}{\sigma^2} (\mathbf{C} \mathbf{X} \boldsymbol{\beta})^\top (\mathbf{C} \mathbf{X} \boldsymbol{\beta}) \right]^{-1} \quad (5.26)$$

Clearly,

$$\frac{\rho}{se(\rho)} \stackrel{a}{\sim} N(0, 1) \quad (5.27)$$

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:

$$W_\lambda = \frac{\hat{\lambda}^2}{\hat{\mathbb{V}}(\lambda)} \quad (5.28)$$

where $\widehat{\mathbf{V}}(\lambda)$ can be obtained from Equation 4.47 as:

$$\widehat{\mathbf{V}}(\lambda) = \left[-\frac{\text{tr}(\mathbf{W}_B)}{\sigma^2} + \text{tr}(\mathbf{W}_B)^2 + \text{tr}(\mathbf{W}_B^\top \mathbf{W}_B) \right]^{-1} \quad (5.29)$$

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 $\widehat{\boldsymbol{\theta}}$,
- (b) compute an estimator of the variance matrix of the asymptotic distribution of $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$, for example, the information $\mathbf{I}(\widehat{\boldsymbol{\theta}})^{-1}$,
- (c) and finally compute the quadratic form:

$$W = n \cdot \widehat{\boldsymbol{\theta}}^\top \widehat{\mathbf{V}}_w^{-1} \widehat{\boldsymbol{\theta}} \quad (5.30)$$

where $\widehat{\mathbf{V}}_w$ is the $(2, 2)$ block of $\mathbf{I}(\widehat{\boldsymbol{\theta}})^{-1}$ partitioned conformably with $\boldsymbol{\theta}$: that is

$$\widehat{\mathbf{V}}_w = \left\{ \mathbf{I}_{22}(\widehat{\boldsymbol{\theta}}) - \mathbf{I}_{21}(\widehat{\boldsymbol{\theta}}) \left[\mathbf{I}_{11}(\widehat{\boldsymbol{\theta}}) \right]^{-1} \mathbf{I}_{12}(\widehat{\boldsymbol{\theta}}) \right\}^{-1} \quad (5.31)$$

which is a **consistent estimator** of the asymptotic variance of $\widehat{\boldsymbol{\theta}}_2$.

- (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:

$$LM = \left(\frac{\partial \log L(\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}} \right)^\top \left[\mathbf{I}(\tilde{\boldsymbol{\theta}}) \right]^{-1} \left(\frac{\partial \log L(\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}} \right) \xrightarrow{d} \chi(r) \quad (5.32)$$

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

$$\begin{aligned}\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} &= \frac{1}{\sigma^2} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top \mathbf{B}(\lambda)^\top \mathbf{B}(\lambda) \mathbf{X} \\ \frac{\partial \log L(\boldsymbol{\theta})}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top \mathbf{B}(\lambda)^\top \mathbf{B}(\lambda) (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ \frac{\partial \log L(\boldsymbol{\theta})}{\partial \lambda} &= -\text{tr}(\mathbf{B}^{-1} \mathbf{W}) + \frac{1}{\sigma^2} [\boldsymbol{\epsilon}^\top \mathbf{W} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})]\end{aligned}\tag{5.33}$$

Under the null hypothesis $H_0 : \lambda = 0$, we get:

$$\begin{aligned}\left. \frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} \right|_{\lambda=0} &= \frac{1}{\sigma^2} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\top \mathbf{I}_n^\top \mathbf{I}_n \mathbf{X} = \frac{1}{\sigma^2} \hat{\boldsymbol{\epsilon}}_{OLS}^\top \mathbf{X} \\ \left. \frac{\partial \log L(\boldsymbol{\theta})}{\partial \sigma^2} \right|_{\lambda=0} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \hat{\boldsymbol{\epsilon}}_{OLS}^\top \hat{\boldsymbol{\epsilon}}_{OLS} \\ \left. \frac{\partial \log L(\boldsymbol{\theta})}{\partial \lambda} \right|_{\lambda=0} &= \frac{\boldsymbol{\epsilon}^\top \mathbf{W} \boldsymbol{\epsilon}}{\sigma^2}\end{aligned}\tag{5.34}$$

The test is essentially based on the score with respect to λ , i.e., on

$$s_\lambda = \left. \frac{\partial \log L(\boldsymbol{\theta})}{\partial \lambda} \right|_{\lambda=0} = \frac{\boldsymbol{\epsilon}^\top \mathbf{W} \boldsymbol{\epsilon}}{\sigma^2}\tag{5.35}$$

Recall that:

$$\text{AsyVar}(\boldsymbol{\beta}, \sigma^2, \lambda) = \begin{pmatrix} \frac{\mathbf{X}(\lambda)^\top \mathbf{X}(\lambda)}{\sigma^2} & 0 & 0 \\ k \times k & & \\ 0 & \frac{n}{2\sigma^4} & \frac{\text{tr}(\mathbf{W}_B)}{\sigma^2} \\ 0 & \frac{\text{tr}(\mathbf{W}_B)}{\sigma^2} & \text{tr}(\mathbf{W}_B)^2 + \text{tr}(\mathbf{W}_B^\top \mathbf{W}_B) \end{pmatrix}^{-1}\tag{5.36}$$

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 $\mathbb{E}(\boldsymbol{\epsilon}^\top \mathbf{W} \boldsymbol{\epsilon}) = \sigma^2 \text{tr}(\mathbf{W}) = \mathbf{0}$ as \mathbf{W} has a zero diagonal. Furthermore,

$$\mathbb{E}_{H_0} \left(\frac{\partial \log L(\boldsymbol{\theta})}{\partial \lambda^2} \right) = -\text{tr}(\mathbf{W}^2 + \mathbf{W}^\top \mathbf{W})\tag{5.37}$$

Then the expression for the LM test for a SEM specification is:

$$LM_{ERR} = \frac{1}{C} \left(\frac{\hat{\boldsymbol{\varepsilon}}^\top \mathbf{W} \hat{\boldsymbol{\varepsilon}}}{\hat{\sigma}^2} \right)^2 \quad (5.38)$$

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:

$$s_{\rho=0} = \left. \frac{\partial \log L(\boldsymbol{\theta})}{\partial \rho} \right|_{\rho=0} = \frac{1}{\sigma^2} \boldsymbol{\varepsilon}^\top \mathbf{W} \mathbf{y} \quad (5.39)$$

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

$$\mathbb{V}(\boldsymbol{\beta}, \sigma^2, \rho) \big|_{\rho=0} = \begin{pmatrix} \frac{1}{\sigma^2} (\mathbf{X}^\top \mathbf{X}) & \mathbf{0}' & \frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{W} \mathbf{X} \boldsymbol{\beta} \\ \cdot & \frac{n}{2\sigma^4} & \mathbf{0} \\ \cdot & \cdot & \text{tr}(\mathbf{W}^2 + \mathbf{W}^\top \mathbf{W}) + \frac{1}{\sigma^2} (\mathbf{W} \mathbf{X} \boldsymbol{\beta})^\top (\mathbf{W} \mathbf{X} \boldsymbol{\beta}) \end{pmatrix}^{-1}$$

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:

$$LM_{SAR} = \frac{1}{T_1} \left(\frac{\hat{\boldsymbol{\varepsilon}}^\top \mathbf{W} \mathbf{y}}{\hat{\sigma}^2} \right)^2 \quad (5.40)$$

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.

¹This is obtained using partitioned Inversion.

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"))

## Please update scripts to use lm.RStests in place of lm.LMtests

##
## Rao's score (a.k.a Lagrange multiplier) diagnostics for spatial
## dependence
##
## data:
## model: lm(formula = CRIME ~ INC + HOVAL, data = columbus)
## test weights: listw
##
## RSerr = 4.6111, df = 1, p-value = 0.03177
##
## Rao's score (a.k.a Lagrange multiplier) diagnostics for spatial
## dependence
```

```
##
## data:
## model: lm(formula = CRIME ~ INC + HOVAL, data = columbus)
## test weights: listw
##
## adjRSerr = 0.033514, df = 1, p-value = 0.8547
##
##
## Rao's score (a.k.a Lagrange multiplier) diagnostics for spatial
## dependence
##
## data:
## model: lm(formula = CRIME ~ INC + HOVAL, data = columbus)
## test weights: listw
##
## RSlag = 7.8557, df = 1, p-value = 0.005066
##
##
## Rao's score (a.k.a Lagrange multiplier) diagnostics for spatial
## dependence
##
## data:
## model: lm(formula = CRIME ~ INC + HOVAL, data = columbus)
## test weights: listw
##
## adjRSlag = 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:

$$\begin{aligned} \mathbf{y}_n &= \rho_0 \mathbf{M}_n \mathbf{y}_n + \mathbf{X}_n \boldsymbol{\beta}_0 + \mathbf{u}_n = \mathbf{D}_n \boldsymbol{\theta}_0 + \mathbf{u}_n, \\ \mathbf{u}_n &= \lambda_0 \mathbf{M}_n \mathbf{u}_n + \boldsymbol{\varepsilon}_n. \end{aligned} \tag{5.41}$$

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}_n ,

note that:

$$\begin{aligned}
 \hat{\mathbf{u}}_n &= \mathbf{y}_n - \mathbf{D}_n \hat{\boldsymbol{\theta}}_n, \\
 &= \mathbf{D}_n \boldsymbol{\theta}_0 + \mathbf{u}_n - \mathbf{D}_n \hat{\boldsymbol{\theta}}_n, \quad \text{using (5.41)} \\
 &= \mathbf{u}_n - \mathbf{D}_n (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0).
 \end{aligned} \tag{5.42}$$

Using previous Equation (5.42), $Q_n^* = \hat{\mathbf{u}}_n^\top \mathbf{W}_n \hat{\mathbf{u}}_n$ can be expressed as:

$$\begin{aligned}
 \hat{\mathbf{u}}_n^\top \mathbf{W}_n \hat{\mathbf{u}}_n &= \left[\mathbf{u}_n - \mathbf{D}_n (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \right]^\top \mathbf{W}_n \left[\mathbf{u}_n - \mathbf{D}_n (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \right], \\
 &= \mathbf{u}_n^\top \mathbf{W}_n \mathbf{u}_n - \mathbf{u}_n^\top \mathbf{W}_n \mathbf{D}_n (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) - (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top \mathbf{D}_n^\top \mathbf{W}_n \mathbf{u}_n \\
 &\quad + (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top \mathbf{D}_n^\top \mathbf{W}_n \mathbf{D}_n (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0), \\
 &= \mathbf{u}_n^\top \mathbf{W}_n \mathbf{u}_n - \mathbf{u}_n^\top (\mathbf{W}_n + \mathbf{W}_n^\top) \mathbf{D}_n (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top \mathbf{D}_n^\top \mathbf{W}_n \mathbf{D}_n (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)
 \end{aligned}$$

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:

$$\begin{aligned}
 \frac{1}{\sqrt{n}} \hat{\mathbf{u}}_n^\top \mathbf{W}_n \hat{\mathbf{u}}_n &= \frac{1}{\sqrt{n}} \mathbf{u}_n^\top \mathbf{W}_n \mathbf{u}_n - \frac{1}{n} \mathbf{u}_n^\top (\mathbf{W}_n + \mathbf{W}_n^\top) \mathbf{D}_n \sqrt{n} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \\
 &\quad + (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top \left[\frac{1}{n} \mathbf{D}_n^\top \mathbf{W}_n \mathbf{D}_n \right] \sqrt{n} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)
 \end{aligned} \tag{5.43}$$

First, we will show that the last element of (5.43) converges to 0 as $n \rightarrow \infty$. Since $\mathbf{D}_n = [\mathbf{W}_n \mathbf{y}_n, \mathbf{X}_n]$ 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 6.6.4, 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:

$$\mathbf{v}^\top \mathbf{D} \mathbf{v} = \mathbf{v}^\top \left[\frac{\mathbf{D} + \mathbf{D}^\top}{2} \right] \mathbf{v}$$

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 (5.43) as:

$$\begin{aligned}
 \mathbf{u}_n^\top \mathbf{W}_n \mathbf{u}_n &= \boldsymbol{\varepsilon}_n^\top (\mathbf{I}_n - \lambda_0 \mathbf{M}_n^\top)^{-1} \left[\frac{\mathbf{M}_n + \mathbf{M}_n^\top}{2} \right] (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^{-1} \boldsymbol{\varepsilon}_n, \\
 &= \boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n
 \end{aligned}$$

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:

$$\begin{aligned}\mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) &= \sigma_0^2 \text{tr}(\mathbf{A}_n), \\ \mathbb{V}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) &= \sigma_0^4 [\text{tr}(\mathbf{A}_n \mathbf{A}_n^\top) + \text{tr}(\mathbf{A}_n^2)].\end{aligned}$$

For the second part of (5.43), let $\mathbf{R}_0 = (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)$ and note that

$$\frac{1}{n} \mathbf{u}_n^\top (\mathbf{W}_n + \mathbf{W}_n^\top) \mathbf{D}_n = \boldsymbol{\varepsilon}_n^\top \mathbf{R}_0^{-1\top} (\mathbf{W}_n + \mathbf{W}_n^\top) \mathbf{D}_n,$$

where $\mathbf{D}_n = [\mathbf{M}_n \mathbf{y}_n, \mathbf{X} \boldsymbol{\beta}_0]$, and

$$\mathbf{y}_n = (\mathbf{I}_n - \rho_0 \mathbf{M}_n)^{-1} \mathbf{X}_n \boldsymbol{\beta}_0 + (\mathbf{I}_n - \rho_0 \mathbf{M}_n)^{-1} (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^{-1} \boldsymbol{\varepsilon}_n,$$

so that:

$$\begin{aligned}\boldsymbol{\varepsilon}_n^\top \mathbf{R}_0^{-1\top} (\mathbf{W}_n + \mathbf{W}_n^\top) \mathbf{M}_n \mathbf{y}_n &= \boldsymbol{\varepsilon}_n^\top \mathbf{R}_0^{-1\top} (\mathbf{W}_n + \mathbf{W}_n) \mathbf{M}_n [(\mathbf{I}_n - \rho_0 \mathbf{M}_n)^{-1} \mathbf{X}_n \boldsymbol{\beta}_0 \\ &\quad + (\mathbf{I}_n - \rho_0 \mathbf{M}_n)^{-1} \mathbf{R}_0^{-1} \boldsymbol{\varepsilon}_n] \\ &= \boldsymbol{\varepsilon}_n^\top \mathbf{R}_0^{-1\top} (\mathbf{W}_n + \mathbf{W}_n^\top) \mathbf{M}_n (\mathbf{I}_n - \rho_0 \mathbf{M}_n)^{-1} \mathbf{X}_n \boldsymbol{\beta}_0 \\ &\quad + \boldsymbol{\varepsilon}_n^\top \mathbf{R}_0^{-1\top} (\mathbf{W}_n + \mathbf{W}_n^\top) \mathbf{M}_n (\mathbf{I}_n - \rho_0 \mathbf{M}_n)^{-1} \mathbf{R}_0^{-1} \boldsymbol{\varepsilon}_n \\ &= \boldsymbol{\varepsilon}_n^\top \mathbf{B}_n^* + \boldsymbol{\varepsilon}_n^\top \mathbf{C}_n^* \boldsymbol{\varepsilon}_n\end{aligned}$$

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

$$\mathbf{d}_n^\top = \mathbb{E} \left[\frac{1}{n} \mathbf{u}_n^\top (\mathbf{W}_n + \mathbf{W}_n^\top) \mathbf{D}_n \right] = O(1) \quad (5.44)$$

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:

$$\mathbb{V} \left[\frac{1}{n} \mathbf{u}_n^\top (\mathbf{W}_n + \mathbf{W}_n^\top) \mathbf{D}_n \right] = o(1),$$

and the Claim in Equation (5.44) follows by Chebychev's inequality. Therefore:

$$\frac{1}{n} \mathbf{u}_n^\top (\mathbf{W}_n + \mathbf{W}_n^\top) \mathbf{D}_n - \mathbf{d}_n^\top = o_p(1),$$

and

$$\begin{aligned} \frac{1}{n} \mathbf{u}_n^\top (\mathbf{W}_n + \mathbf{W}_n^\top) \mathbf{D}_n \sqrt{n} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) &= \mathbf{d}_n^\top \sqrt{n} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + o_p(1), \\ &= \mathbf{d}_n^\top \left(\mathbf{P}_n \left[\frac{1}{\sqrt{n}} \mathbf{F}_n^\top \boldsymbol{\varepsilon}_n \right] + o_p(1) \right) + o_p(1). \end{aligned}$$

Since $\mathbf{P}_n = \mathbf{P} + o_p(1)$ and $n^{-1/2} \mathbf{F}_n^\top \boldsymbol{\varepsilon}_n = O_p(1)$:

$$\frac{1}{n} \mathbf{u}_n^\top (\mathbf{W}_n + \mathbf{W}_n^\top) \mathbf{D}_n \sqrt{n} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}} \mathbf{d}_n^\top \mathbf{P} \mathbf{F}_n^\top \boldsymbol{\varepsilon}_n + o_p(1) = \frac{1}{\sqrt{n}} \mathbf{b}_n^\top \boldsymbol{\varepsilon}_n + o_p(1),$$

where $\mathbf{b}_n^\top = -\mathbf{d}_n^\top \mathbf{P} \mathbf{F}_n^\top$. Note that \mathbf{P} is the probability limit of \mathbf{P}_n and thus \mathbf{b}_n is nonstochastic. Finally, we can write Equation (5.43) as

$$\frac{1}{\sqrt{n}} \hat{\mathbf{u}}_n^\top \mathbf{W}_n \hat{\mathbf{u}}_n = \frac{1}{\sqrt{n}} [\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n + \mathbf{b}_n^\top \boldsymbol{\varepsilon}_n] + o_p(1).$$

Thus, the asymptotic distribution of the Moran's I statistic is based on estimated disturbances involves the large sample distribution of a linear-quadratic form in innovations.

Instrumental Variables and GMM

Maximum Likelihood Estimation (MLE) is a widely used method in spatial econometrics; however, its application can become computationally demanding as the number of spatial units grows. This is primarily due to the necessity of manipulating $n \times n$ matrices, which involves operations such as matrix multiplication, inversion, and eigenvalue computation. These challenges are exacerbated when working with large datasets, limiting the practicality of MLE in such contexts.

In response to these computational difficulties, alternative estimation techniques such as Instrumental Variables (IV) and the Generalized Method of Moments (GMM) have been developed (Kelejian and Prucha, 1998, 1999; Lee, 2007). These methods offer a more computationally efficient approach, as they circumvent the need for calculating the Jacobian determinant—a key component of the MLE framework—and do not depend on the normality assumption.

This chapter focuses on the theoretical foundations and practical implementation of IV and GMM methods in spatial econometrics, with a particular emphasis on coding in R and their practical applications. Section 6.1 provides a comprehensive review of the GMM estimator for spatial models, drawing heavily on insights from Prucha (2014). The Spatial Two-Stage Least Squares (S2SLS) estimator for the Spatial Lag Model (SLM) is discussed in Section 6.2, followed by an in-depth analysis of its corresponding GMM estimator in Section 6.3. Sections 6.4 and 6.5 explore the Spatial Feasible Generalized Least Squares (SFGLS) estimator and the GMM estimator for the Spatial Error Model (SEM), respectively. Finally, Section 6.6 outlines the Spatial Feasible Generalized Two-Stage Least Squares (SFG2SLS) estimation procedure for the Spatial Autoregressive Combined (SAC) model.

6.1 A Review of GMM

Before explaining the estimation procedure for the SLM, SEM, and SAC models, we review some key aspects of the Generalized Method of Moments (GMM) in the spatial econometrics context.

6.1.1 Model Specification

Suppose the data are generated by the following model:

$$f(y_{in}, \mathbf{x}_{in}, \boldsymbol{\theta}_0) = \epsilon_{in}, \quad i = 1, \dots, n,$$

where $f(y_{in}, \mathbf{x}_{in}, \boldsymbol{\theta}_0)$ represents a system of spatial equations, y_{in} is the dependent variable for unit i , \mathbf{x}_{in} is a vector of explanatory variables, ϵ_{in} is a disturbance term, $\boldsymbol{\theta}_0$ is a $k \times 1$ unknown parameter vector, and $f(\cdot)$ is a known function.

Assume there exists a $1 \times s$ vector of instruments \mathbf{h}_{in} , and let w_{in} denote all observable variables, including instruments, for the i -th unit. For simplicity, assume ϵ_{in} is i.i.d. $(0, \sigma^2)$, and \mathbf{h}_{in} is non-stochastic (these assumptions can be relaxed). The explanatory variables may take the form $\mathbf{x}_{in} = [\mathbf{x}_i, \bar{\mathbf{x}}_{in}, \bar{y}_{in}]$, where \mathbf{x}_i is exogenous, $\bar{\mathbf{x}}_{in} = \sum_j w_{ij} \mathbf{x}_j$, and $\bar{y}_{in} = \sum_j w_{ij} y_{jn}$ are spatial lags, with w_{ij} denoting spatial weights and $w_{ii} = 0$ for all $i = 1, \dots, n$.

Suppose that there exists a vector $s \times 1$ of sample moments

$$\mathbf{g}_n(\boldsymbol{\theta}) = \mathbf{g}_n(w_1, \dots, w_n, \boldsymbol{\theta}) = \begin{pmatrix} g_{1n}(w_1, \dots, w_n, \boldsymbol{\theta}) \\ \vdots \\ g_{sn}(w_1, \dots, w_n, \boldsymbol{\theta}) \end{pmatrix}, \quad (6.1)$$

with $s \geq k$ for identification. Further assume:

$$\mathbb{E}[\mathbf{g}_n(w_1, \dots, w_n, \boldsymbol{\theta})] = \mathbf{0} \iff \boldsymbol{\theta} = \boldsymbol{\theta}_0,$$

that is, the model is identified. Let $\boldsymbol{\Upsilon}_n$ be some $s \times s$ symmetric positive semidefinite weighting matrix, then the corresponding GMM estimator is defined as:

$$\hat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\operatorname{argmin}} \underset{(1 \times s)}{\mathbf{g}_n(w_1, \dots, w_n, \boldsymbol{\theta})}^\top \underset{(s \times s)}{\boldsymbol{\Upsilon}_n} \underset{(s \times 1)}{\mathbf{g}_n(w_1, \dots, w_n, \boldsymbol{\theta})}. \quad (6.2)$$

where $\boldsymbol{\Theta}$ is the parameter space. If $s = k$ (i.e., the model is just identified), the weighting matrix $\boldsymbol{\Upsilon}_n$ is irrelevant and $\hat{\boldsymbol{\theta}}_n$ can be found as a solution to the moment condition:

$$\mathbf{g}_n(w_1, \dots, w_n, \hat{\boldsymbol{\theta}}) = \mathbf{0}. \quad (6.3)$$

In classical GMM literature, **linear moment conditions** are of the form:

$$\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n \mathbf{h}_i^\top \epsilon_i \right] = \mathbf{0},$$

which holds under the maintained assumptions since $\mathbb{E}[\mathbf{h}_i^\top \epsilon_i] = \mathbf{h}_i^\top \mathbb{E}[\epsilon_i] = \mathbf{0}$.

The spatial econometrics literature often considers **quadratic moment conditions**. Let \mathbf{A}_q be an $n \times n$ matrix with $\operatorname{tr}(\mathbf{A}_q) = 0$. As explained in the following sections, such matrices are of class \mathcal{P}_1 . Assuming \mathbf{A}_q is non-stochastic, the quadratic moment conditions are:

$$\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n a_{ijq} \epsilon_i \epsilon_j \right] = \mathbf{0}, \quad (6.4)$$

which clearly holds under the maintained assumptions. To see this, let $\boldsymbol{\varepsilon} = [\epsilon_1, \dots, \epsilon_n]^\top$, then the moment conditions in (6.4) can be rewritten as:

$$\mathbb{E} \left[\frac{\boldsymbol{\varepsilon}^\top \mathbf{A}_q \boldsymbol{\varepsilon}}{n} \right] = \text{tr} \left[\frac{\mathbf{A}_q \mathbb{E}(\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top)}{n} \right] = \sigma^2 \frac{\text{tr}(\mathbf{A}_q)}{n} = \mathbf{0},$$

since under the maintained assumptions $\mathbb{E}[\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^\top] = \sigma^2 \mathbf{I}_n$ and $\text{tr}(\mathbf{A}_q) = 0$.

Now let $\boldsymbol{\theta}_0 = [\lambda_0, \boldsymbol{\delta}_0]^\top$ and suppose the sample moment vector in (6.1) 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:

$$\begin{aligned} \mathbb{E}[\mathbf{g}_n^\lambda(w_1, \dots, w_n, \lambda, \boldsymbol{\delta})] &= \mathbf{0} \iff \lambda = \lambda_0, \\ \mathbb{E}[\mathbf{g}_n^\delta(w_1, \dots, w_n, \lambda, \boldsymbol{\delta})] &= \mathbf{0} \iff \boldsymbol{\delta} = \boldsymbol{\delta}_0, \end{aligned}$$

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 $\boldsymbol{\mathcal{R}}_n^{\lambda\lambda}$:

$$\widehat{\lambda}_n = \underset{\lambda}{\text{argmin}} \quad \mathbf{g}_n^\lambda(w_1, \dots, w_n, \lambda, \widehat{\boldsymbol{\delta}}_n)^\top \boldsymbol{\mathcal{R}}_n^{\lambda\lambda}(w_1, \dots, w_n, \lambda, \widehat{\boldsymbol{\delta}}_n). \quad (6.5)$$

Using $\widehat{\lambda}_n$ we may further consider the following estimator for $\boldsymbol{\delta}_0$ corresponding to some weight matrix $\boldsymbol{\mathcal{R}}_n^{\delta\delta}$:

$$\widehat{\boldsymbol{\delta}}_n = \underset{\boldsymbol{\delta}}{\text{argmin}} \quad \mathbf{g}_n^\delta(w_1, \dots, w_n, \widehat{\lambda}_n, \boldsymbol{\delta})^\top \boldsymbol{\mathcal{R}}_n^{\delta\delta}(w_1, \dots, w_n, \widehat{\lambda}_n, \boldsymbol{\delta}). \quad (6.6)$$

GMM estimator like $\widehat{\boldsymbol{\theta}}$ in Equation (6.2) are often referred to as **one-step estimators**. Estimators like $\widehat{\lambda}_n$ and $\widehat{\boldsymbol{\delta}}_n$ in Equations (6.5) and (6.6) above, where the sample moments depend on some initial estimator, are often referred to as **two-step estimators**.

When the model conditions hold, the most efficient one-step estimator is expected to outperform even the most efficient two-step estimators in terms of statistical efficiency. However, practical trade-offs often influence the choice of estimator. One key trade-off involves computational complexity. For small sample sizes, maximum likelihood (ML) estimation may serve as an alternative to GMM, offering robust performance. Conversely, in large samples, computational efficiency and feasibility often outweigh gains in statistical efficiency, making two-step GMM estimators an attractive option.

Moreover, Monte Carlo studies indicate that the efficiency loss associated with two-step estimators is often modest in many practical scenarios. Another critical consideration is the potential impact of misspecifying a moment condition. Such misspecification can lead to inconsistent estimates for all model parameters, underscoring the importance of careful model specification and moment selection.

6.1.2 Asymptotic Distribution of One-Step GMM Estimator

Assuming that $\hat{\theta}_n$ is an interior point, the first-order condition for maximization of the objective function is:

$$\underset{(k \times 1)}{\mathbf{0}} = \frac{\partial Q_n(\hat{\theta}_n)}{\partial \underset{(k \times 1)}{\theta}} = -\underset{(k \times s)}{\mathbf{G}_n(\hat{\theta}_n)}^\top \underset{(s \times s)}{\mathbf{r}_n} \underset{(s \times 1)}{\mathbf{g}_n(\hat{\theta}_n)}, \quad (6.7)$$

where $Q_n = \mathbf{g}_n(\theta)^\top \mathbf{r}_n \mathbf{g}_n(\theta)$, and $\mathbf{G}_n(\theta)$ is the Jacobian of $\mathbf{g}_n(\theta)$:

$$\mathbf{G}_n(\theta) \equiv \frac{\partial \mathbf{g}_n(\theta)}{\partial \theta^\top}.$$

Using a Taylor expansion of $\mathbf{g}_n(\theta)$, we obtain:

$$\mathbf{g}_n(\hat{\theta}_n) = \mathbf{g}_n(\theta_0) + \mathbf{G}_n(\bar{\theta}) (\hat{\theta}_n - \theta_0). \quad (6.8)$$

Substituting (6.8) into the first-order condition (6.7), we get:

$$\mathbf{0} = \frac{\partial Q_n(\hat{\theta}_n)}{\partial \theta} = -\mathbf{G}_n(\hat{\theta}_n)^\top \mathbf{r}_n \mathbf{g}_n(\hat{\theta}_n) - \mathbf{G}_n(\hat{\theta}_n)^\top \mathbf{r}_n \mathbf{G}_n(\bar{\theta}) (\hat{\theta}_n - \theta_0).$$

Solving this for $(\hat{\theta}_n - \theta_0)$, multiplying by \sqrt{n} , and under some regularity conditions yields:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -[\mathbf{G}_0^\top \mathbf{r}_0 \mathbf{G}_0]^{-1} \mathbf{G}_0^\top \mathbf{r}_0 [\sqrt{n} \mathbf{g}_n(\theta_0)] + o_p(1),$$

where

$$\begin{aligned} \mathbf{G}_0 &= \text{plim}_{n \rightarrow \infty} \frac{\partial \mathbf{g}_n(\theta_0)}{\partial \theta} \quad \text{by some LLN,} \\ \mathbf{r}_0 &= \text{plim}_{n \rightarrow \infty} \mathbf{r}_n \quad \text{by some LLN,} \\ \sqrt{n} \mathbf{g}_n(\theta_0) &\xrightarrow{d} \text{N}(\mathbf{0}, \boldsymbol{\Psi}_0) \quad \text{by some CLT} \end{aligned} \quad (6.9)$$

where $\boldsymbol{\Psi}_0$ is some positive definite matrix. Then applying traditional asymptotic rules:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \text{N}(\mathbf{0}, \boldsymbol{\Phi}_0),$$

where:

$$\boldsymbol{\Phi}_0 = [\mathbf{G}_0^\top \mathbf{r}_0 \mathbf{G}_0]^{-1} \mathbf{G}_0^\top \mathbf{r}_0 \boldsymbol{\Psi}_0 \mathbf{r}_0 \mathbf{G}_0 [\mathbf{G}_0^\top \mathbf{r}_0 \mathbf{G}_0]^{-1}.$$

It can be seen that if we choose $\mathbf{r}_n = \hat{\boldsymbol{\Psi}}_n^{-1}$ (with weights given by the inverse of the variance-covariance matrix of the moment conditions), where $\hat{\boldsymbol{\Psi}}_n \xrightarrow{p} \boldsymbol{\Psi}_0$, the variance-covariance simplifies to

$$\boldsymbol{\Phi}_0 = [\mathbf{G}_0^\top \boldsymbol{\Psi}_0^{-1} \mathbf{G}_0]^{-1}.$$

Since $[\mathbf{G}_0^\top \mathbf{r}_0 \mathbf{G}_0]^{-1} \mathbf{G}_0^\top \mathbf{r}_0 \boldsymbol{\Psi}_0 \mathbf{r}_0 \mathbf{G}_0 [\mathbf{G}_0^\top \mathbf{r}_0 \mathbf{G}_0]^{-1} - [\mathbf{G}_0^\top \boldsymbol{\Psi}_0^{-1} \mathbf{G}_0]^{-1}$ is positive semidefinite it follows that $\mathbf{r}_n = \hat{\boldsymbol{\Phi}}_n^{-1}$ gives the optimal GMM estimator with lower asymptotic variance.

The most relevant literature on one-step GMM estimators in spatial econometrics includes:

- GMM estimator of SLM assuming homoskedasticity (Lee, 2007).

- GMM estimator of SLM assuming homoskedasticity, but reducing the joint maximization to the maximization with respect to ρ only.
- GMM estimator of SLM assuming heteroskedasticity (Lin and Lee, 2010).
- GMM estimator of SLM with additional endogenous variables (Liu and Saraiva, 2015).
- GMM estimator of SAC model assuming homoskedasticity (Lee and Liu, 2010; Liu et al., 2010).

6.1.3 Asympmtotic Distribution of Two-Step GMM Estimator

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 (6.5). Applying this approach, and assuming typical regularity conditions, we get:

$$\sqrt{n}(\hat{\lambda}_n - \lambda_0) = -[(\mathbf{G}_0^{\lambda\lambda})^\top \mathbf{r}_0^{\lambda\lambda} \mathbf{G}_0^{\lambda\lambda}]^{-1} (\mathbf{G}_0^{\lambda\lambda})^\top \mathbf{r}_0^{\lambda\lambda} \left[\sqrt{n} \mathbf{g}_n^\lambda(\lambda_0, \boldsymbol{\delta}_0) + \mathbf{G}_0^{\lambda\delta} \sqrt{n}(\tilde{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0) \right] + o_p(1), \quad (6.10)$$

where

$$\begin{aligned} \frac{\partial \mathbf{g}_n^\lambda(\lambda_0, \boldsymbol{\delta}_0)}{\partial \boldsymbol{\lambda}} &\xrightarrow{p} \mathbf{G}_0^{\lambda\lambda}, \\ \frac{\partial \mathbf{g}_n^\lambda(\lambda_0, \boldsymbol{\delta}_0)}{\partial \boldsymbol{\delta}} &\xrightarrow{p} \mathbf{G}_0^{\lambda\delta}, \\ \mathbf{r}_n^{\lambda\lambda} &\xrightarrow{p} \mathbf{r}_0^{\lambda\lambda}. \end{aligned}$$

In many cases the estimator $\tilde{\boldsymbol{\delta}}_n$ will be asymptotically linear in the sense that

$$\sqrt{n}(\tilde{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0) = \frac{1}{\sqrt{n}} \mathbf{T}_n^\top \boldsymbol{\varepsilon}_n + o_p(1),$$

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 $\boldsymbol{\varepsilon}_n = (\epsilon_1, \dots, \epsilon_n)^\top$. Now define:

$$\mathbf{g}_{*n}^\lambda(\lambda_0, \boldsymbol{\delta}_0) = \mathbf{g}_n^\lambda(\lambda_0, \boldsymbol{\delta}_0) + \frac{1}{n} \mathbf{G}_0^{\lambda\delta} \mathbf{T}_n^\top \boldsymbol{\varepsilon}_n.$$

Then Equation (6.10) can be rewritten as:

$$\sqrt{n}(\hat{\lambda}_n - \lambda_0) = -[(\mathbf{G}_0^{\lambda\lambda})^\top \mathbf{r}_0^{\lambda\lambda} \mathbf{G}_0^{\lambda\lambda}]^{-1} (\mathbf{G}_0^{\lambda\lambda})^\top \mathbf{r}_0^{\lambda\lambda} [\sqrt{n} \mathbf{g}_{*n}^\lambda(\lambda_0, \boldsymbol{\delta}_0)] + o_p(1). \quad (6.11)$$

Now suppose that

$$\sqrt{n} \mathbf{g}_{*n}^\lambda(\lambda_0, \boldsymbol{\delta}_0) \xrightarrow{d} N(0, \boldsymbol{\Psi}_*^{\lambda\lambda}),$$

where $\boldsymbol{\Psi}_*^{\lambda\lambda}$ is some positive definite matrix. Then

$$\sqrt{n}(\hat{\lambda}_n - \lambda_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Phi}_*^{\lambda\lambda}),$$

¹For more on two-step estimation see Newey and McFadden (1994, section 6)

with:

$$\Phi_*^{\lambda\lambda} = [(G_0^{\lambda\lambda})^\top \mathbf{r}_0^{\lambda\lambda} G_0^{\lambda\lambda}]^{-1} (G_0^{\lambda\lambda})^\top \mathbf{r}_0^{\lambda\lambda} \Psi_*^{\lambda\lambda} \mathbf{r}_0^{\lambda\lambda} G_0^{\lambda\lambda} [(G_0^{\lambda\lambda})^\top \mathbf{r}_0^{\lambda\lambda} G_0^{\lambda\lambda}]^{-1}.$$

From this it is seen that if we choose $\mathbf{r}_n^{\lambda\lambda} = (\tilde{\Psi}_{*n}^{\lambda\lambda})^{-1}$ where $\tilde{\Psi}_{*n}^{\lambda\lambda} \xrightarrow{p} \Psi_*^{\lambda\lambda}$, then variance-covariance simplifies to

$$\Phi_*^{\lambda\lambda} = [(G_0^{\lambda\lambda})^\top (\Psi_*^{\lambda\lambda})^{-1} G_0^{\lambda\lambda}]^{-1}.$$

So, using the weighting matrix $\mathbf{r}_n^{\lambda\lambda}$, a consistent estimator for the inverse of the limiting variance-covariance matrix $\Psi_*^{\lambda\lambda}$ yields the efficient two-step GMM estimator.

Suppose that Equation (6.9) holds and:

$$\Psi = \begin{pmatrix} \Psi^{\lambda\lambda} & \Psi^{\lambda\delta} \\ \Psi^{\delta\lambda} & \Psi^{\delta\delta} \end{pmatrix},$$

then the limiting distribution of the sample moment vector \mathbf{g}_n^λ evaluated at the true parameter is given by

$$\sqrt{n} \mathbf{g}_n^\lambda(\lambda_0, \delta_0) \xrightarrow{d} N(\mathbf{0}, \Psi^{\lambda\lambda}).$$

Note that in general $\Psi_*^{\lambda\lambda} \neq \Psi^{\lambda\lambda}$, unless $\mathbf{G}^{\lambda\delta} = \mathbf{0}$, and that in general $\Psi_*^{\lambda\lambda}$ will depend on \mathbf{T}_n , which in turn will depend on the employed estimator $\hat{\delta}_n$. In other words, unless $\mathbf{G}_0^{\lambda\delta} = \mathbf{0}$, for a two-step GMM estimator, we cannot simply use the variance-covariance matrix $\Psi^{\lambda\lambda}$ of the sample moment vector $\mathbf{g}^\lambda(\lambda_0, \delta_0)$, rather we need to work with the variance-covariance matrix $\Psi_*^{\lambda\lambda}$.

Prucha (2014) illustrate the difference between $\Psi^{\lambda\lambda}$, with elements $\Psi_{rs}^{\lambda\lambda}$, and $\Psi_*^{\lambda\lambda}$, 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

$$\mathbf{g}_n^\lambda(\lambda_0, \delta_0) = \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n a_{ij1} \epsilon_i \epsilon_j \\ \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n a_{ij2} \epsilon_i \epsilon_j \end{pmatrix}.$$

Now, for $r = 1, 2$, let a_{ir} denote the (i, r) th element of $\mathbf{G}_0^{\lambda\delta} \mathbf{T}_n^\top$, then by Equation (6.9):

$$\mathbf{g}_{*n}^\lambda(\lambda_0, \delta_0) = \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n a_{ij1} \epsilon_i \epsilon_j + \frac{1}{n} \sum_{i=1}^n a_{i1} \epsilon_i \\ \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n a_{ij2} \epsilon_i \epsilon_j + \frac{1}{n} \sum_{i=1}^n a_{i2} \epsilon_i \end{pmatrix}$$

It then follows from Limiting Distribution for linear-quadratic forms 3.31 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 $\hat{\delta}_n$ is employed in the sample moment vector $\mathbf{g}_n^\lambda(\lambda_0, \hat{\delta})$ used to form the objective function for the two-step GMM estimator $\hat{\lambda}_n$ defined in Equation (6.5). 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.

The relevant theoretical literature on two-step GMM estimators for spatial models are:

- (a) Generalized Spatial Two Stage Least Squares Estimator (GS2SLS) of the SAC model assuming homoskedasticity (Kelejian and Prucha, 1998).
- (b) Feasible Generalized Least Squares (FGLS) of the SEM model under homoskedasticity. (Kelejian and Prucha, 1999).
- (c) Feasible Generalized Two Stage Least Squares Estimator (FG2SLS) Estimator of the SAC model under heteroskedasticity (Kelejian and Prucha, 2010).

6.2 Spatial Two Stage Estimation of SLM

In this section, we derive the Spatial Two Stage Least Square (S2SLS) procedure for estimating the SLM. To do so, we rely on Kelejian and Prucha (1998) who first derived the asymptotic properties of the S2SLS estimator.² This estimation approach has two notable advantages:

- (a) It does not require the computation of the Jacobian term, making it computationally more efficient than ML-based methods.
- (b) It avoids the strong assumption of normality of the error terms, thereby offering greater flexibility.

To understand the essence of this procedure, let us revisit the formulation of the SLM, expressed as:

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

Alternatively, we can write the model more concisely as:

$$\mathbf{y}_n = \mathbf{Z}_n \boldsymbol{\delta}_0 + \boldsymbol{\varepsilon}_n,$$

where $\mathbf{Z}_n = [\mathbf{X}_n, \mathbf{W}_n \mathbf{y}_n]$ is an $n \times (k+1)$ matrix that combines the **exogenous regressors** and the spatially lagged dependent variable, and the $(k+1) \times 1$ coefficient column vector is rearranged as $\boldsymbol{\delta}_0 = (\boldsymbol{\beta}_0^\top, \rho_0)^\top$. As discussed in Section 4.1, the inclusion of the spatially lagged dependent variable, $\mathbf{W}_n \mathbf{y}_n$, on the right-hand side of the equation introduces endogeneity or simultaneous equation bias. Therefore, the OLS estimates are inconsistent.

To address this issue, we may use an instrumental variables (IV) approach rather than resorting to QML or ML methods. The IV principle relies on the existence of a matrix of instruments, \mathbf{H}_n , that is strongly correlated with \mathbf{Z}_n but asymptotically uncorrelated with the error term $\boldsymbol{\varepsilon}_n$.

At this point is important to stress that the only endogenous variable in this model is the spatially lagged dependent variable. Thus, the instrument matrix \mathbf{H}_n should include all the predetermined variables, that is, \mathbf{X}_n and the instrument(s) for $\mathbf{W}_n \mathbf{y}_n$.

6.2.1 Instruments in the Spatial Context

What constitutes the best instruments for $\mathbf{W}_n \mathbf{y}_n$? To construct the instrument matrix \mathbf{H}_n , it is essential to refer to the literature on **optimal instrumental variables**. Broadly, this

²In particular, Kelejian and Prucha (1998) derived this model as the first step in their Generalized S2SLS.

literature suggest that the ‘best instruments’ for the r.h.s variables are their conditional means. Consequently, the ideal instruments are expressed as:

$$\begin{aligned}\mathbb{E}(\mathbf{Z}_n|\mathbf{X}_n) &= [\mathbb{E}(\mathbf{X}_n|\mathbf{X}_n), \mathbb{E}(\mathbf{W}_n\mathbf{y}_n|\mathbf{X}_n)], \\ &= [\mathbf{X}_n, \mathbf{W}_n\mathbb{E}(\mathbf{y}_n|\mathbf{X}_n)] \quad \text{since } \mathbf{W}_n \text{ is non-stochastic.}\end{aligned}$$

Given that \mathbf{X}_n is exogenous, it serves as its own best instrument, whereas the best instruments for $\mathbf{W}_n\mathbf{y}_n$ are given by $\mathbf{W}_n\mathbb{E}(\mathbf{y}_n|\mathbf{X}_n)$. Noting that the reduced-form equation is $\mathbf{y}_n = (\mathbf{I}_n - \rho\mathbf{W}_n)^{-1}(\mathbf{X}_n\boldsymbol{\beta} + \boldsymbol{\varepsilon}_n)$, and applying Leontief Expansion (Lemma 2.4), the expected value of the reduced form is:

$$\mathbb{E}(\mathbf{y}_n|\mathbf{X}_n) = (\mathbf{I}_n - \rho\mathbf{W}_n)^{-1}\mathbf{X}_n\boldsymbol{\beta} = [\mathbf{I}_n + \rho\mathbf{W}_n + \rho^2\mathbf{W}_n^2 + \dots]\mathbf{X}_n\boldsymbol{\beta} = \left[\sum_{l=1}^{\infty} \rho^l \mathbf{W}_n^l\right]\mathbf{X}_n\boldsymbol{\beta}. \quad (6.12)$$

The challenge lies in approximating $\mathbb{E}(\mathbf{y}_n|\mathbf{X}_n)$ without directly inverting the $n \times n$ matrix $(\mathbf{I}_n - \rho\mathbf{W}_n)$. Equation (6.12) reveals that $\mathbb{E}(\mathbf{y}_n|\mathbf{X}_n)$ can be expressed as a linear function of $\mathbf{X}_n, \mathbf{W}_n\mathbf{X}_n, \mathbf{W}_n^2\mathbf{X}_n, \dots$. As a result, and given that the roots of $\rho\mathbf{W}_n$ are less than one in absolute value (or ρ is in its parameter space), the expectation can also be written as:

$$\begin{aligned}\mathbb{E}(\mathbf{W}_n\mathbf{y}_n|\mathbf{X}_n) &= \mathbf{W}_n(\mathbf{I}_n - \rho\mathbf{W}_n)^{-1}\mathbf{X}_n\boldsymbol{\beta}, \\ &= \mathbf{W}_n[\mathbf{I}_n + \rho\mathbf{W}_n + \rho^2\mathbf{W}_n^2 + \rho^3\mathbf{W}_n^3 + \dots]\mathbf{X}_n\boldsymbol{\beta}, \\ &= \mathbf{W}_n\left[\sum_{l=1}^{\infty} \rho^l \mathbf{W}_n^l\right]\mathbf{X}_n\boldsymbol{\beta}, \\ &= \mathbf{W}_n\mathbf{X}_n\boldsymbol{\beta} + \mathbf{W}_n^2\mathbf{X}_n(\rho\boldsymbol{\beta}) + \mathbf{W}_n^3\mathbf{X}_n(\rho^2\boldsymbol{\beta}) + \mathbf{W}_n^4\mathbf{X}_n(\rho^3\boldsymbol{\beta}) + \dots\end{aligned}$$

To avoid issues associated with the computation of the inverse of $(\mathbf{I}_n - \rho\mathbf{W}_n)$, [Kelejian and Prucha \(1998, 1999\)](#) propose an approximation of the optimal instruments. Recognizing that $\mathbb{E}(\mathbf{y}_n|\mathbf{X}_n)$ is linear in $\mathbf{X}_n, \mathbf{W}_n\mathbf{X}_n, \mathbf{W}_n^2\mathbf{X}_n, \dots$, they suggest using a set of instruments \mathbf{H}_n which consists of the linearly independent (LI) columns of $\mathbf{X}_n, \mathbf{W}_n\mathbf{X}_n, \mathbf{W}_n^2\mathbf{X}_n, \dots, \mathbf{W}_n^l\mathbf{X}_n$, where l is a pre-selected finite constant and is generally set to 2 in applied studies. Thus, if $l = 2$, the instrument matrix becomes:

$$\mathbf{H}_n = (\mathbf{X}_n, \mathbf{W}_n\mathbf{X}_n, \mathbf{W}_n^2\mathbf{X}_n).$$



The intuition behind the instruments is the following: Since \mathbf{X}_n determines \mathbf{y}_n , then it must be true that $\mathbf{W}_n\mathbf{X}_n, \mathbf{W}_n^2\mathbf{X}_n, \dots$ determines $\mathbf{W}_n\mathbf{y}_n$. Furthermore, since \mathbf{X}_n is uncorrelated with $\boldsymbol{\varepsilon}_n$, then $\mathbf{W}_n\mathbf{X}_n$ must be also uncorrelated with $\boldsymbol{\varepsilon}_n$.

6.2.2 Defining the S2SLS Estimator

With the instrument matrix \mathbf{H}_n defined, we can now apply the standard two-stage least squares (2SLS) procedure, modified to account for the asymptotic properties of the spatial weight matrices \mathbf{W}_n and $(\mathbf{I}_n - \rho\mathbf{W}_n)$. Due to the inclusion of these spatial components, this method is referred to as the spatial two stage least squares (S2SLS) ([Kelejian and Prucha, 1998](#)).

We begin by introducing the assumptions concerning the structure of the error term. Specifically, we assume that the errors form triangular arrays and exhibit heteroskedasticity. Note that [Kelejian and Prucha \(1998\)](#) derived the asymptotic properties under the assumption of homoskedastic errors, while [Kelejian and Prucha \(2010\)](#) extended this framework to accommodate heteroskedasticity.

Assumption 6.1 — Heteroskedastic Errors ([Kelejian and Prucha, 2010](#)). The errors terms $\{\epsilon_{i,n}, 1 \leq i \leq n, n \geq 1\}$ 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 $\sup_{1 \leq i \leq n, n \geq 1} \mathbb{E} |\epsilon_{i,n}|^{4+\eta}$ 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 characterizes the error terms' first two moments without imposing assumptions on their full distribution. It explicitly allows for heteroskedasticity, meaning the unobserved variables may have different variances across spatial units. Additionally, this assumption accommodates cases where the innovations depend on the sample size n by requiring the errors to form a **triangular arrays**. See our discussion in Section 3.7 about triangular arrays.

For reference, we also present the assumption of homoskedastic errors as initially proposed by [Kelejian and Prucha \(1998\)](#):

Assumption 6.2 — Homoskedastic Errors ([Kelejian and Prucha, 1998](#)). The errors $\{\epsilon_{i,n}, 1 \leq i \leq n, n \geq 1\}$ are distributed identically. Further, the errors $\{\epsilon_{i,n}, 1 \leq i \leq n\}$ are for each n distributed jointly independent with $\mathbb{E}(\epsilon_{i,n}) = 0$ and $\mathbb{E}(\epsilon_{i,n}^2) = \sigma_\epsilon^2$, with $0 < \sigma_\epsilon^2 < b < \infty$. Additionally the errors are assumed to possess fourth moments.

We now outline several key assumptions regarding the behavior of the spatial weight matrix, \mathbf{W}_n .

Assumption 6.3 — Diagonal elements of \mathbf{W}_n ([Kelejian and Prucha, 1998](#)). All diagonal elements of the spatial weighting matrix \mathbf{W}_n are zero.

Assumption 6.3 (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 6.4 — Nonsingularity ([Kelejian and Prucha, 1998](#)). The matrix $(\mathbf{I}_n - \rho_0 \mathbf{W}_n)$ is nonsingular with $|\rho_0| < 1$.

The nonsingularity condition in Assumption 6.4 allows us to express the reduced form of the true model as:

$$\mathbf{y}_n = (\mathbf{I}_n - \rho_0 \mathbf{W}_n)^{-1} \mathbf{X}_n \beta_0 + (\mathbf{I}_n - \rho_0 \mathbf{W}_n)^{-1} \boldsymbol{\epsilon}_n.$$

This assumption ensures that the model is well-defined, enabling us to solve for \mathbf{y}_n . Additionally, [Kelejian and Prucha \(1998\)](#) note that the elements of $(\mathbf{I}_n - \rho_0 \mathbf{W}_n)^{-1}$ depend on the sample size n , even if the elements of \mathbf{W}_n do not vary with n . Therefore, the elements of \mathbf{y}_n also depend on n , meaning that \mathbf{y}_n forms a triangular array, even in cases where the errors $\epsilon_{i,n}$ are independent of n .

Furthermore, Assumption 6.1 (Heteroskedastic Errors) implies that the population variance-

covariance matrix of \mathbf{y}_n is given by:

$$\mathbb{E}(\mathbf{y}_n \mathbf{y}_n^\top) = \boldsymbol{\Omega}_{y_n} = (\mathbf{I}_n - \rho_0 \mathbf{W}_n)^{-1} \boldsymbol{\Sigma}_n (\mathbf{I}_n - \rho_0 \mathbf{W}_n^\top)^{-1}, \quad (6.13)$$

where $\boldsymbol{\Sigma}_n = \text{Diag}(\sigma_{i,n}^2)$, and $\text{Diag}(\cdot)$ is the operator that generates a diagonal matrix.

Under the homoskedasticity assumption (Assumption 6.2), this variance-covariance matrix simplifies to:

$$\mathbb{E}(\mathbf{y}_n \mathbf{y}_n^\top) = \boldsymbol{\Omega}_{y_n} = \sigma_\epsilon^2 (\mathbf{I}_n - \rho_0 \mathbf{W}_n)^{-1} (\mathbf{I}_n - \rho_0 \mathbf{W}_n^\top)^{-1}. \quad (6.14)$$

Assumption 6.5 — 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.

Assumption 6.5 ensures that the variance of \mathbf{y}_n in Equation (6.13), which depends on \mathbf{W}_n and $(\mathbf{I}_n - \rho_0 \mathbf{W}_n)$, are uniformly bounded in absolute value as n goes to infinity. This limits the degree of correlation among the elements of both $\boldsymbol{\epsilon}_n$ and \mathbf{y}_n . This technical assumption is crucial for deriving the large-sample properties of the regression parameter estimators.

R Applied to \mathbf{W}_n Assumption 6.5 (Bounded matrices) means that each cross-sectional unit can only have a limited number of neighbors. When applied to $(\mathbf{I}_n - \rho_0 \mathbf{W}_n)$, it limits the degree of spatial correlation among units.

Assumption 6.6 — 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.

This assumption ensures that the regressors are not perfectly collinear, which is essential for identifying the parameters of the model. Uniform boundedness further ensures the stability of the regression coefficients in large samples

We now introduce the assumptions concerning the instrument matrix, \mathbf{H}_n .

Assumption 6.7 — 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 6.8 — 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.

The Rank Condition of Assumption 6.7 establishes that there exists a least p columns that are linearly independent such that $p \leq k + 1$. That is, the model can be just- or over-identified. In addition, since \mathbf{H}_n is composed of a subset of linearly independent columns of $(\mathbf{X}_n, \mathbf{W}_n \mathbf{X}_n, \mathbf{W}_n^2 \mathbf{X}_n, \dots)$, it does not depend on the parameters of the model, and hence, it is non-stochastic. This simplifies the asymptotic properties of the 2SLS estimator. This

contrast with the Best 2SLS estimator, which requires consistent estimate of the parameters. See Section 6.2.7.

Assumption 6.8 establishes conditions for **identification** of the model. The first condition in Assumption 6.8 ensures that the instrument matrix \mathbf{H}_n remains well-conditioned as the sample size increases, preventing issues such as multicollinearity among the instruments. Importantly, since \mathbf{H}_n contains spatially lagged explanatory variables, this condition also implies that $\mathbf{W}_n \mathbf{X}_n$ and \mathbf{X}_n cannot be linearly dependent. This condition would be violated if for example $\mathbf{W}_n \mathbf{X}_n$ included a spatial lag of a constant term or if the model is the pure SLM.

The second condition in Assumption 6.8 (Limits of Instruments) ensures a nonzero correlation between the instruments and the explanatory variables. Specifically, it guarantees that the instruments are valid in the sense of being correlated with the endogenous components of the model.

Note that $n^{-1} \mathbf{H}_n^\top \mathbf{Z}_n = [n^{-1} \mathbf{H}_n^\top \mathbf{X}_n, n^{-1} \mathbf{W}_n \mathbf{y}_n]$. Then part (b) of Assumption 6.8 implies:

$$\text{plim}_{n \rightarrow \infty} \frac{1}{n} \mathbf{H}_n^\top \mathbf{W}_n \mathbf{y}_n = \lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{H}_n^\top \mathbb{E}(\mathbf{W}_n \mathbf{y}_n). \quad (6.15)$$

Proof. We need to show that the expectation of $n^{-1} \mathbf{H}_n^\top \mathbf{W}_n \mathbf{y}_n$ exists and its variance converges to zero. The result then follows from Theorem 3.5 (Consistency of Unbiased Estimator).

Let $\boldsymbol{\psi}_n = \frac{1}{n} \mathbf{H}_n^\top \mathbf{W}_n \mathbf{y}_n$. Since $\mathbb{E}(\mathbf{y}_n) = \mathbf{A}_0^{-1} \mathbf{X}_n \boldsymbol{\beta}_0$, where $\mathbf{A}_0 = (\mathbf{I}_n - \rho_0 \mathbf{W}_n)$:

$$\mathbb{E}(\boldsymbol{\psi}_n) = \mathbb{E}\left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{W}_n \mathbf{y}_n\right) = \frac{1}{n} \mathbf{H}_n^\top \mathbf{W}_n \mathbb{E}(\mathbf{y}_n) = \frac{1}{n} \mathbf{H}_n^\top \mathbf{W}_n \mathbf{A}_0^{-1} \mathbf{X}_n \boldsymbol{\beta}_0.$$

This expectation exists as long as all its elements are uniformly bounded and \mathbf{A}_0 is invertible as $n \rightarrow \infty$.

The variance of $\boldsymbol{\psi}_n$ is

$$\begin{aligned} \mathbb{V}\left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{W}_n \mathbf{y}_n\right) &= \frac{1}{n^2} \mathbf{H}_n^\top \mathbf{W}_n \mathbb{E}(\mathbf{y}_n \mathbf{y}_n^\top) \mathbf{W}_n^\top \mathbf{H}_n, \\ &= \frac{1}{n^2} \mathbf{H}_n^\top \mathbf{W}_n \boldsymbol{\Omega}_{\mathbf{y}_n} \mathbf{W}_n^\top \mathbf{H}_n, \quad \text{using Equation (6.14)} \\ &= \frac{1}{n^2} \mathbf{H}_n^\top \mathbf{D}_n \mathbf{H}_n, \end{aligned}$$

where $\mathbf{D}_n = \mathbf{W}_n \boldsymbol{\Omega}_{\mathbf{y}_n} \mathbf{W}_n^\top$. Assumption 6.5 implies that the row and column sums of \mathbf{D}_n are uniformly bounded in absolute value. Using Definition 3.8.1, then there exists a constant c_d such that $\max_{1 \leq j \leq n} \sum_{i=1}^n |d_{n,ij}| \leq c_d$ and $\max_{1 \leq i \leq n} \sum_{j=1}^n |d_{n,ij}| \leq c_d$. By Assumption 6.6 and 6.5, the elements of $\mathbf{H}_n = [\mathbf{X}_n, \mathbf{W}_n \mathbf{X}_n, \mathbf{W}_n^2 \mathbf{X}_n, \dots]$ are uniformly bounded in absolute value

by some finite constant, say c_h . Let $\delta_{ij,n}$ be the (i, j) element of $\mathbb{V}(\frac{1}{n}\mathbf{H}_n^\top \mathbf{W}_n \mathbf{y}_n)$. Then

$$\begin{aligned}\delta_{ij,n} &= n^{-2} \sum_{r=1}^n \sum_{s=1}^n h_{si,n} d_{sr,n} h_{rj,n}, \\ |\delta_{ij,n}| &= n^{-2} \left| \sum_{r=1}^n \sum_{s=1}^n h_{si,n} d_{sr,n} h_{rj,n} \right|, \\ |\delta_{ij,n}| &\leq n^{-2} \sum_{r=1}^n \sum_{s=1}^n |h_{si,n} d_{sr,n} h_{rj,n}| \text{ by triangle inequality 3.B.2,} \\ &\leq n^{-2} c_h^2 \sum_{r=1}^n \sum_{s=1}^n |d_{sr}|, \\ &\leq n^{-2} c_h^2 \sum_{r=1}^n c_d, \\ &\leq n^{-2} c_h^2 c_d n, \\ &= n^{-1} c_h^2 c_d = o(1).\end{aligned}$$

Since $\mathbb{V}(\frac{1}{n}\mathbf{H}_n^\top \mathbf{W}_n \mathbf{y}_n) \rightarrow \mathbf{0}$ as $n \rightarrow \infty$, then by Theorem 3.5:

$$\frac{1}{n} \mathbf{H}_n^\top \mathbf{W}_n \mathbf{y}_n \xrightarrow{p} \lim_{n \rightarrow \infty} n^{-1} \mathbf{H}_n^\top \mathbb{E}(\mathbf{W}_n \mathbf{y}_n).$$

■

Given all these 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:

$$\hat{\delta}_{S2SLS} = \left(\hat{\mathbf{Z}}_n^\top \mathbf{Z}_n \right)^{-1} \hat{\mathbf{Z}}_n^\top \mathbf{y}_n, \quad (6.16)$$

where:

$$\hat{\mathbf{Z}}_n = \mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{Z}_n = \mathbf{P}_{H,n} \mathbf{Z}_n, \quad (6.17)$$

where the projection matrix $\mathbf{P}_{H,n}$ is defined as

$$\mathbf{P}_{H,n} = \mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top. \quad (6.18)$$

Alternatively, it can also be expressed as

$$\begin{aligned}\hat{\delta}_{S2SLS} &= \left[\mathbf{Z}_n^\top \mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{Z}_n \right]^{-1} \mathbf{Z}_n^\top \mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{y}_n, \\ &= \left[\mathbf{Z}_n^\top \mathbf{P}_{H,n} \mathbf{Z}_n \right]^{-1} \mathbf{Z}_n^\top \mathbf{P}_{H,n} \mathbf{y}_n.\end{aligned} \quad (6.19)$$

The S2SLS estimator in (6.16) is conceptually similar to the standard 2SLS, with adjustment to account for the spatial structure of the model. The estimation proceeds as follows:

- (a) **First stage:** The first stage involves regressing the endogenous variables \mathbf{Z}_n on the instruments \mathbf{H}_n via OLS. This regression can be expressed as $\mathbf{Z}_n = \mathbf{H}_n \boldsymbol{\theta} + \boldsymbol{\xi}_n$, where

$\hat{\theta}_n = (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{Z}_n$. The predicted values $\hat{\mathbf{Z}}_n$ are then obtained using Equation (6.17) where $\mathbf{P}_{H,n}$ is the projection matrix, which symmetric and idempotent, and hence nonsingular. It is also important to note that the projection matrix does not affect \mathbf{X}_n , but it does affect the endogenous variable $\mathbf{W}_n \mathbf{y}_n$:

$$\mathbf{P}_{H,n} \mathbf{Z}_n = [\mathbf{X}_n, \mathbf{P}_{H,n} \mathbf{W}_n \mathbf{y}_n] = [\mathbf{X}_n, \widehat{\mathbf{W}_n \mathbf{y}_n}]. \quad (6.20)$$

- (b) **Second stage:** In the second stage, the regression of \mathbf{y}_n on $\hat{\mathbf{Z}}_n$ is used to estimate the parameters, yielding the S2SLS estimator as defined in Equation (6.16) or (6.19).

6.2.3 Additional Endogenous Variables

In the specification considered so far, the only endogenous variable has been the spatially lagged dependent variable $\mathbf{W}\mathbf{y}$. However, in practice, other explanatory variables may also exhibit endogeneity, necessitating additional instruments beyond the spatially lagged exogenous variables required for $\mathbf{W}\mathbf{y}$.

For example, [Anselin and Lozano-Gracia \(2008\)](#) analyzed the effect of improved air quality on house prices. Since air quality variables were derived using interpolated air pollution measures, they argued that these measures could suffer from an “error in variable” problem, introducing an additional source of endogeneity alongside the spatial lag. Specifically, they considered 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:

$$\mathbf{H} = (\mathbf{X}_1, \mathbf{W}\mathbf{X}_1, \dots, \mathbf{W}^l\mathbf{X}_1, \mathbf{X}_2, \mathbf{W}\mathbf{X}_2, \dots, \mathbf{W}^l\mathbf{X}_2)_{LI} \quad (6.21)$$

6.2.4 Consistency of S2SLS Estimator

In this section, we provide a sketch of the proof for the consistency of the S2SLS estimator. For a formal proof see Kelejian and Prucha (1998) or Kelejian and Prucha (2010). To begin, recall that the $n \times n$ matrix $\mathbf{H}_n(\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top$ is symmetric and idempotent, implying that $\widehat{\mathbf{Z}}_n^\top \mathbf{Z}_n = \widehat{\mathbf{Z}}_n^\top \widehat{\mathbf{Z}}_n$. Using this property, we express the S2SLS estimator in terms of the population error term as follows:

$$\begin{aligned} \widehat{\boldsymbol{\delta}}_n &= \boldsymbol{\delta}_0 + \left(\widehat{\mathbf{Z}}_n^\top \widehat{\mathbf{Z}}_n \right)^{-1} \widehat{\mathbf{Z}}_n^\top \boldsymbol{\varepsilon}_n, \\ &= \boldsymbol{\delta}_0 + \left[\left(\mathbf{H}_n(\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{Z}_n \right)^\top \left(\mathbf{H}_n(\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{Z}_n \right) \right]^{-1} \left(\mathbf{H}_n(\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{Z}_n \right)^\top \boldsymbol{\varepsilon}_n, \\ &= \boldsymbol{\delta}_0 + \left[\mathbf{Z}_n^\top \mathbf{H}_n(\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{Z}_n \right]^{-1} \mathbf{Z}_n^\top \mathbf{H}_n(\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \boldsymbol{\varepsilon}_n. \end{aligned} \quad (6.22)$$

Solving for $\widehat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0$, we obtain:

$$\begin{aligned} (\widehat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0) &= \left[\left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right)^\top \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right) \right]^{-1} \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right)^\top \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \left(\frac{1}{n} \mathbf{H}_n^\top \boldsymbol{\varepsilon}_n \right), \\ &= \widetilde{\mathbf{P}}_n^\top \left(\frac{1}{n} \mathbf{H}_n^\top \boldsymbol{\varepsilon}_n \right), \end{aligned} \quad (6.23)$$

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 6.8 (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, $\widetilde{\mathbf{P}}_n \xrightarrow{p} \mathbf{P}_0$, where $\mathbf{P}_0 = \mathbf{Q}_{HH}^{-1} \mathbf{Q}_{HZ} (\mathbf{Q}_{HZ}^\top \mathbf{Q}_{HH}^{-1} \mathbf{Q}_{HZ})^{-1}$ is a finite matrix and exists because \mathbf{Q}_{HH} is invertible by Assumption 6.8 and $(\mathbf{Q}_{HZ}^\top \mathbf{Q}_{HH}^{-1} \mathbf{Q}_{HZ})^{-1}$ exists because \mathbf{Q}_{HZ} has full column rank by Assumption 6.8. Thus,

$$\widetilde{\mathbf{P}}_n - \mathbf{P}_0 = o_p(1) \implies \widetilde{\mathbf{P}}_n = \mathbf{P}_0 + o_p(1), \quad (6.24)$$

where $\mathbf{P}_0 = O(1)$ (why?).

By Assumption 6.6 and 6.5, the elements of \mathbf{H}_n are uniformly bounded in absolute value by some finite constant. Assumption 6.1 (Heteroskedastic Errors) implies that $\epsilon_{i,n}$ forms a

triangular array of identically distributed random variables. Furthermore, Assumption 6.1 states 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 $n \rightarrow \infty$, and the elements of $\frac{1}{n}\mathbf{H}_n^\top \boldsymbol{\Sigma}_n \mathbf{H}_n$ are uniformly bounded in absolute value (see proof 6.2.2) by Chebyshev's Theorem 3.5, $n^{-1}\mathbf{H}_n^\top \boldsymbol{\varepsilon}_n \xrightarrow{p} \mathbf{0}$ and $\widehat{\boldsymbol{\delta}}_n \xrightarrow{p} \boldsymbol{\delta}_0$.

6.2.5 Asymptotic Distribution of S2SLS Estimator

Multiplying Equation (6.23) by \sqrt{n} we obtain:

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0) &= \left[\left(\frac{1}{n}\mathbf{H}_n^\top \mathbf{Z}_n \right)^\top \left(\frac{1}{n}\mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \left(\frac{1}{n}\mathbf{H}_n^\top \mathbf{Z}_n \right) \right]^{-1} \left(\frac{1}{n}\mathbf{H}_n^\top \mathbf{Z}_n \right)^\top \left(\frac{1}{n}\mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \frac{1}{\sqrt{n}}\mathbf{H}_n^\top \boldsymbol{\varepsilon}_n, \\ &= \widetilde{\mathbf{P}}_n^\top \left(\frac{1}{\sqrt{n}}\mathbf{H}_n^\top \boldsymbol{\varepsilon}_n \right).\end{aligned}\tag{6.25}$$

From previous section we now that $\widetilde{\mathbf{P}}_n \xrightarrow{p} \mathbf{P}_0$, where $\mathbf{P}_0 = \mathbf{Q}_{HH}^{-1} \mathbf{Q}_{HZ} (\mathbf{Q}_{HZ}^\top \mathbf{Q}_{HH}^{-1} \mathbf{Q}_{HZ})^{-1}$ is a finite matrix and exists. By Assumption 6.5 and 6.6, the elements of \mathbf{H}_n are uniformly bounded in absolute value by some finite constant. Then, by CLT for triangular arrays with heteroskedastic innovations 3.31:

$$\frac{1}{\sqrt{n}}\mathbf{H}_n^\top \boldsymbol{\varepsilon}_n \xrightarrow{d} \text{N}(\mathbf{0}, \boldsymbol{\Delta}_0),$$

where $\boldsymbol{\Delta}_0 = \lim_{n \rightarrow \infty} \frac{1}{n}\mathbf{H}_n^\top \boldsymbol{\Sigma}_n \mathbf{H}_n$.

Finally :

$$\sqrt{n}(\widehat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0) \xrightarrow{d} \text{N}(\mathbf{0}, \boldsymbol{\Omega}_0),$$

where

$$\boldsymbol{\Omega}_0 = (\mathbf{Q}_{HZ}^\top \mathbf{Q}_{HH}^{-1} \mathbf{Q}_{HZ})^{-1} \mathbf{Q}_{HZ}^\top \mathbf{Q}_{HH}^{-1} \boldsymbol{\Delta}_0 \mathbf{Q}_{HH}^{-1} \mathbf{Q}_{HZ} (\mathbf{Q}_{HZ}^\top \mathbf{Q}_{HH}^{-1} \mathbf{Q}_{HZ})^{-1}.$$

If the error terms are homoskedatic, then by CLT for triangular arrays with homoskedastic innovations 3.30:

$$\frac{1}{\sqrt{n}}\mathbf{H}_n^\top \boldsymbol{\varepsilon}_n \xrightarrow{d} \text{N}(\mathbf{0}, \sigma_0^2 \mathbf{Q}_{HH}),$$

where $\mathbf{Q}_{HH} = \lim_{n \rightarrow \infty} n^{-1}\mathbf{H}_n^\top \mathbf{H}_n$. Thus:

$$\sqrt{n}(\widehat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0) \xrightarrow{d} \text{N}(\mathbf{0}, \boldsymbol{\Omega}_0^o),$$

where

$$\boldsymbol{\Omega}_0^o = \sigma_0^2 (\mathbf{Q}_{HZ}^\top \mathbf{Q}_{HH}^{-1} \mathbf{Q}_{HZ})^{-1}.$$

Now, we present a formal Theorem for the asymptotic properties of the 2SLS Estimator for SLM.

Theorem 6.9 — Spatial 2SLS Estimator for SLM. Suppose that Assumptions 6.1, 6.3, 6.4, 6.5, 6.6, 6.7, 6.8 hold. Then the 2SLS estimator defined as

$$\hat{\boldsymbol{\delta}}_n = \left(\hat{\mathbf{Z}}_n^\top \hat{\mathbf{Z}}_n \right)^{-1} \hat{\mathbf{Z}}_n^\top \mathbf{y}_n,$$

is consistent, and its asymptotic distribution is:

$$\sqrt{n}(\hat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Omega}_0)$$

where

$$\boldsymbol{\Omega}_0 = (\mathbf{Q}_{HZ}^\top \mathbf{Q}_{HH}^{-1} \mathbf{Q}_{HZ})^{-1} \mathbf{Q}_{HZ}^\top \mathbf{Q}_{HH}^{-1} \boldsymbol{\Delta}_0 \mathbf{Q}_{HH}^{-1} \mathbf{Q}_{HZ} (\mathbf{Q}_{HZ}^\top \mathbf{Q}_{HH}^{-1} \mathbf{Q}_{HZ})^{-1}, \quad (6.26)$$

with

$$\begin{aligned} \mathbf{Q}_{HH} &= \lim_{n \rightarrow \infty} n^{-1} \mathbf{H}_n^\top \mathbf{H}_n, \\ \mathbf{Q}_{HZ} &= \text{plim}_{n \rightarrow \infty} n^{-1} \mathbf{H}_n^\top \mathbf{Z}_n, \\ \boldsymbol{\Delta}_0 &= \lim_{n \rightarrow \infty} n^{-1} \mathbf{H}_n^\top \boldsymbol{\Sigma}_n \mathbf{H}_n. \end{aligned}$$

If Assumption 6.1 is replaced by Assumption 6.2, then

$$\sqrt{n}(\hat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Omega}_0^o),$$

where

$$\boldsymbol{\Omega}_0^o = \sigma_0^2 (\mathbf{Q}_{HZ}^\top \mathbf{Q}_{HH}^{-1} \mathbf{Q}_{HZ})^{-1}. \quad (6.27)$$

The variance-covariance under homoskedasticity in Equation (6.26) can be estimated as:

$$\begin{aligned} \hat{\boldsymbol{\Omega}}_n &= \left[\left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right)^\top \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right) \right]^{-1} \\ &\times \left[\left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right)^\top \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \hat{\boldsymbol{\Delta}}_n \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right) \right] \\ &\times \left[\left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right)^\top \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right) \right]^{-1}, \end{aligned} \quad (6.28)$$

where

$$\hat{\boldsymbol{\Delta}}_n = \frac{1}{n} \sum_{i=1}^n \hat{\epsilon}_i^2 \mathbf{h}_i \mathbf{h}_i^\top, \quad (6.29)$$

where \mathbf{h}_i is the $p \times 1$ vector of instruments for i observation. This is unknown as the White's, Robust or Sandwich estimator.

Under homoskedasticity, the variance-covariance matrix is estimated as

$$\hat{\boldsymbol{\Omega}}^o = \hat{\sigma}_n^2 \left[\left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right)^\top \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right) \right]^{-1}, \quad (6.30)$$

where:

$$\hat{\sigma}_n^2 = \frac{\hat{\boldsymbol{\epsilon}}_n^\top \hat{\boldsymbol{\epsilon}}_n}{n}, \quad \hat{\boldsymbol{\epsilon}}_n = \mathbf{y}_n - \hat{\mathbf{y}}_n. \quad (6.31)$$

6.2.6 Coding S2SLS in R

In this section, we demonstrate how to implement a custom function to estimate the S2SLS estimator. This implementation includes creating the instrument matrix, \mathbf{H} , estimating the model parameters, and providing variance-covariance matrix options for homoskedastic and robust errors. Additionally, we compare our implementation with the `stspls` function from the `spatialreg` package (Pebesma and Bivand, 2023) to ensure consistency.

We start by defining a function to create the instrument matrix \mathbf{H} , which includes spatial lags of the exogenous variables up to a user-specified order.

```
# Function that creates WXs
make.H <- function(W, X, l = 3){
  # This function creates the instruments (WX, ..., W^lX)
  # Drop constant (if any)
  names.x <- colnames(X)
  if (names.x[1] == "(Intercept)") X <- matrix(X[, -1],
                                                dim(X)[1],
                                                dim(X)[2] - 1) #Drop first column

  names.x <- names.x[which(names.x != "(Intercept)")]
  # Create lagged X variables
  sq1 <- seq(1, ncol(X) * l, ncol(X))
  sq2 <- seq(ncol(X), ncol(X) * l, ncol(X))
  Hmat <- matrix(NA, nrow = nrow(X), ncol = ncol(X) * l)
  names.ins <- c()
  for (i in 1:l) {
    Hmat[, sq1[i]:sq2[i]] <- as.matrix(W %*% X)
    X <- Hmat[, sq1[i]:sq2[i]]
    names.ins <- c(names.ins,
                   paste(paste(replicate(i, "W"), collapse = ""),
                         names.x, sep = "*"))
  }
  colnames(Hmat) <- names.ins
  return(Hmat)
}
```

This function takes as arguments the spatial weight matrix \mathbf{W} , which is assumed to be of class `matrix`, the $n \times K$ matrix of exogenous variables, \mathbf{X} , and order of spatial lags l . Then, it provides the matrix of instruments, $\mathbf{W}\mathbf{X}, \mathbf{W}^2\mathbf{X}^2, \dots, \mathbf{W}^l\mathbf{X}^l$, recursively. Note that the spatial lag of the constant is not included (why?).

We create an artificial DGP similar to Lee (2007) for testing the S2SLS implementation:

```
# Generate DGP
library("spatialreg")
library("spdep")
set.seed(1986)
n <- 529
```

```

rho      <- 0.6
W.nb2    <- cell2nb(sqrt(n), sqrt(n))
W        <- nb2mat(W.nb2)

# Exogenous variables
x1       <- rnorm(n)
x2       <- rnorm(n)
x3       <- rnorm(n)

# DGP parameters
b0 <- 0 ; b1 <- -1; b2 <- 0; b3 <- 1
sigma2 <- 2
epsilon <- rnorm(n, mean = 0, sd = sqrt(sigma2))

# Simulate the dependent variable
y <- solve(diag(n) - rho * W) %*% (b0 + b1*x1 + b2*x2 + b3*x3 + epsilon)

# Data as data.frame
data <- as.data.frame(cbind(y, x1, x2, x3))
names(data) <- c("y", "x1", "x2", "x3")

```

In the following lines, we show how the function `make.H` works using $l = 3$.

```

X <- cbind(1, x1, x2, x3)
colnames(X) <- c("(Intercept)", "x1", "x2", "x3")
H <- make.H(W = W, X = X, l = 3)
head(H)

```

	W*x1	W*x2	W*x3	WW*x1	WW*x2	WW*x3
[1,]	-0.1838084	1.4871684	1.06237375	-0.19093708	-0.66721937	0.11419352
[2,]	-0.1734224	-0.4801317	0.00889001	-0.18346010	1.16724914	0.84907535
[3,]	-0.1376933	0.7858546	0.65742663	-0.07600407	-0.06559888	-0.33095264
[4,]	0.2516524	0.5410386	-0.66094372	-0.17572619	0.04624316	0.12822625
[5,]	-0.2524966	-0.7357899	-0.27611007	0.53392070	0.46568956	-0.40883400
[6,]	0.7127103	0.4896574	-0.25597660	-0.08935028	-0.65274040	-0.05759884

	WWW*x1	WWW*x2	WWW*x3
[1,]	-0.2364339	1.1506372	0.80406622
[2,]	-0.1724420	-0.4360736	-0.07942842
[3,]	-0.1916439	0.6032002	0.46356720
[4,]	0.2365006	0.1309099	-0.37543431
[5,]	-0.1367966	-0.4243089	-0.02981086
[6,]	0.4918535	0.2858139	-0.35198454

Thus, the matrix `H` contains the spatial lag of the exogenous variables up to the third order.

The main function for estimating the S2SLS model is defined as follows:

```

# Main function to estimate S2SLS estimator
slm.2sls <- function(formula, data, W, instruments = 2){
  # Model Frame
  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())

  # Get variables and globals
  y <- model.response(mf)
  X <- model.matrix(formula, mf)
  n <- nrow(X)
  Wy <- W %**% y
  sn <- nrow(W)
  if (n != sn) stop("number of spatial units in W is different to the number of data")

  # Generate matrix of instruments H = [X, WX, ... ]
  # and select LI vars
  H <- cbind(X, make.H(W = W, X = X, l = instruments))
  H <- H[, qr(H)$pivot[seq_len(qr(H)$rank)]]

  # Get S2SLS estimates
  Z      <- cbind(X, Wy)
  colnames(Z) <- c(colnames(X), "Wy")

  # Create projection matrix
  HH     <- crossprod(H)
  PH     <- H %**% solve(HH) %**% t(H)

  # Compute S2SLS coefficients
  Z_hat  <- PH %**% Z
  b_2sls <- solve(crossprod(Z_hat)) %**% crossprod(Z_hat, y)
  y_hat  <- Z %**% b_2sls
  e_hat  <- y - y_hat

  # Save results
  results <- structure(
    list(
      coefficients = b_2sls,
      call        = callT,
      X           = X,
      H           = H,
      Z           = Z,
      y           = y,

```

```

    PH          = PH,
    e_hat       = e_hat
  ),
  class = 'mys2sls'
)
}

```

The function `slm.2sls` provides the S2SLS estimates using the matrix of instruments \mathbf{H} suggested by Kelejian and Prucha (1998). The 2SLS estimates are obtained using Equation (6.16). It then returns an object of class `mys2sls` along with some elements as a list such as the estimated coefficients, the call of the model, and different matrix that can be used in other functions.

In the next lines, we create the S3 method `vcov` for an object with class `mys2sls`:

```

# S3 Method vcov
vcov.mys2sls <- function(object, tse = c("homo", "rob"), ...){
  tse      <- match.arg(tse)
  n        <- nrow(object$Z)
  df       <- n - ncol(object$Z)
  Q.HZ     <- (t(object$H) %*% object$Z) / n
  Q.HH.i   <- solve(crossprod(object$H) / n)
  if (tse == "homo"){
    s2      <- crossprod(object$e_hat) / df
    var     <- drop(s2) * solve(t(Q.HZ) %*% Q.HH.i %*% Q.HZ) / n
  } else {
    Delta.hat <- 0
    for (i in 1:nrow(object$Z)){
      Delta.hat <- Delta.hat + drop(object$e_hat[i] ^2) * tcrossprod(object$H[i, ])
    }
    bread    <- solve(t(Q.HZ) %*% Q.HH.i %*% Q.HZ)
    cheese   <- t(Q.HZ) %*% Q.HH.i %*% (Delta.hat / n) %*% Q.HH.i %*% Q.HZ
    var      <- (bread %*% cheese %*% bread) / n
  }
  return(var)
}

```

This function computes the estimated variance-covariance (VC) matrix for the S2SLS estimators. If `tse = "homo"`, it returns the VC matrix under homoskedastic error, following Equation (6.30). In this case, $\hat{\sigma}_n^2$ is estimated using Equation (6.31), applying a small-sample by dividing by $n-k$ instead of n . If `tse = "rob"`, the function returns the robust VC matrix, estimated using Equation (6.28), where $\hat{\Delta}$ is computed according to Equation (6.29). In both cases, the VC matrices are divided by n for finite sample approximation.

In the following lines, we code the S3 functions `summary` and `print.summary` for the class of our model.

```

# S3 method summary
summary.mys2sls <- function(object,
                             tse = c("homo", "rob"),
                             table = TRUE,
                             digits = max(3, .Options$digits - 3),
                             ...){
  tse      <- match.arg(tse)
  n        <- nrow(object$Z)
  df       <- n - ncol(object$Z)
  b        <- object$coefficients
  std.err  <- sqrt(diag(vcov(object, tse = tse)))
  z        <- b / std.err
  p        <- 2 * pt(-abs(z), df = df)
  CoefTable <- cbind(b, std.err, z, p)
  colnames(CoefTable) <- c("Estimate", "Std.Error", "t-value", "Pr(>|t|)")
  result <- structure(
    list(
      CoefTable = CoefTable,
      digits    = digits,
      call      = object$call),
    class = 'summary.mys2sls'
  )
  return(result)
}

# S3 method print.summary
print.summary.mys2sls <- function(x,
                                  digits = x$digits,
                                  na.print = "",
                                  symbolic.cor = p > 4,
                                  signif.stars = getOption("show.signif.stars"),
                                  ...){
  {
    cat("\nCall:\n")
    cat(paste(deparse(x$call), sep = "\n", collapse = "\n"), "\n\n", sep = "")

    cat("\nCoefficients:\n")
    printCoefmat(x$CoefTable, digit = digits, P.value = TRUE, has.Pvalue = TRUE)
    invisible(NULL)
  }
}

```

Now, we can apply our function to estimate the S2SLS model using homoskedastic and robust standard errors. By default, the `summary` method reports results with homoskedastic standard errors. To obtain robust standard errors, we specify `tse = "rob"` in the `summary` function.

```
# Estimate S2SLS model
s2spls.e <- slm.2spls(y ~ x1 + x2 + x3, data = data, W = W, instruments = 2)
summary(s2spls.e)

##
## Call:
## slm.2spls(formula = y ~ x1 + x2 + x3, data = data, W = W, instruments = 2)
##
##
## Coefficients:
##              Estimate Std.Error t-value Pr(>|t|)
## (Intercept) -0.06983    0.06656  -1.049    0.295
## x1           -1.00574    0.06347 -15.846   <2e-16 ***
## x2           -0.04353    0.06276  -0.694    0.488
## x3            1.01485    0.06720  15.102   <2e-16 ***
## Wy            0.64731    0.06222  10.404   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(s2spls.e, tse = "rob")

##
## Call:
## slm.2spls(formula = y ~ x1 + x2 + x3, data = data, W = W, instruments = 2)
##
##
## Coefficients:
##              Estimate Std.Error t-value Pr(>|t|)
## (Intercept) -0.06983    0.06682  -1.045    0.296
## x1           -1.00574    0.06587 -15.269   <2e-16 ***
## x2           -0.04353    0.06290  -0.692    0.489
## x3            1.01485    0.06669  15.217   <2e-16 ***
## Wy            0.64731    0.05892  10.986   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now we check our results using `stspls` function:

```
spreg1 <- stspls(y ~ x1 + x2 + x3, data = data, listw = mat2listw(W, style = "W"))
spreg2 <- stspls(y ~ x1 + x2 + x3, data = data, listw = mat2listw(W, style = "W"),
  robust = TRUE, HC = "HCO")
cbind(as.numeric(coef(spreg1)[c(2, 3, 4, 5, 1)]), as.numeric(s2spls.e$coefficients))

##              [,1]          [,2]
## [1,] -0.06983166 -0.06983166
## [2,] -1.00573883 -1.00573883
## [3,] -0.04352912 -0.04352912
```

```
## [4,] 1.01485440 1.01485440
## [5,] 0.64730746 0.64730746

cbind(sqrt(diag(spreg1$var)[c(2, 3, 4, 5, 1)]),
      sqrt(diag(vcov(s2s1s.e))),
      sqrt(diag(spreg2$var)[c(2, 3, 4, 5, 1)]),
      sqrt(diag(vcov(s2s1s.e, tse = "rob"))))

##           [,1]      [,2]      [,3]      [,4]
## (Intercept) 0.06655890 0.06655890 0.06681918 0.06681918
## x1          0.06346892 0.06346892 0.06586903 0.06586903
## x2          0.06276194 0.06276194 0.06290024 0.06290024
## x3          0.06720175 0.06720175 0.06669244 0.06669244
## Rho         0.06221908 0.06221908 0.05892332 0.05892332
```

6.2.7 Best S2SLS Estimator

Lee (2003) suggested the so-called optimal instruments matrix, which gives rise to the Best Spatial Two Stage Least Square (BS2SLS) Estimator. Instead of using the IV matrices that are composed of a subset of the linearly independent columns of $(\mathbf{X}_n, \mathbf{W}_n \mathbf{X}_n, \mathbf{W}_n^2 \mathbf{X}_n, \dots)$, Lee (2003) suggests the instrument matrix:

$$\widetilde{\mathbf{H}}_n^* = \left[\mathbf{X}_n, \mathbf{W}_n (\mathbf{I}_n - \widetilde{\rho}_n \mathbf{W}_n)^{-1} \mathbf{X}_n \widetilde{\beta}_n \right],$$

which requires the use of consistent first-stage estimates for ρ_0 and β_0 .³ Lee (2003) shows that these instruments are asymptotically optimal in the sense that they provide the smallest asymptotic variance among all the IV estimators of the SLM.⁴

The resulting S2SLS estimator is called the Best Spatial Two Stage Least Squares (BS2SLS) estimator. However, since $\widetilde{\mathbf{H}}_n^*$ is an $n \times (k+1)$ matrix, the model is just-identified and the BS2SLS estimator is an IV estimator given by

$$\widehat{\delta}_n = \left[\widetilde{\mathbf{H}}_n^{*\top} \mathbf{Z}_n \right]^{-1} \widetilde{\mathbf{H}}_n^{*\top} \mathbf{y}_n. \quad (6.32)$$

We summarize Lee (2003)'s assumptions as follows:

Assumption 6.10 — Assumptions for the BS2SLSE of the SLM (Lee, 2003). Assume the following:

- (a) The errors $\{\epsilon_{i,n}, 1 \leq i \leq n, n \geq 1\}$ are distributed identically. Further, the errors $\{\epsilon_{i,n}, 1 \leq i \leq n\}$ are for each n distributed jointly independent with $\mathbb{E}(\epsilon_{i,n}) = 0$ and $\mathbb{E}(\epsilon_{i,n}^2) = \sigma_\epsilon^2$, with $0 < \sigma_\epsilon^2 < b < \infty$. Additionally the errors are assumed to possess fourth moments.

³In 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 $\mathbf{H} = [\mathbf{X}, \mathbf{W} (\sum_{l=1}^{\infty} \rho_0^l \mathbf{W}^l) \mathbf{X} \beta]$. In practice, the power expansion must be truncated at a finite value to ensure feasibility.

⁴It is important to note that Lee (2003) provides the BS2SLS estimator in the context of the SAC model. Thus, his estimator is termed Best Generalized Spatial Two Stage Least Squares (BGS2SLS) estimator.

- (b) The matrices $(\mathbf{I}_n - \rho_0 \mathbf{W}_n)$ are nonsingular.
- (c) The row and column sums of the matrices \mathbf{W}_n and $(\mathbf{I}_n - \rho_0 \mathbf{W}_n)^{-1}$ are uniformly bounded in absolute value.
- (d) The elements of the matrices \mathbf{X}_n are uniformly bounded in absolute value.
- (e) Let $\mathbf{G}_0 = \mathbf{W}_n(\mathbf{I}_n - \rho_0 \mathbf{W}_n)^{-1}$. The limit $\mathbf{J}^* = \lim_{n \rightarrow \infty} (1/n) \mathbf{H}_n^{*\top} \mathbf{H}_n^*$ exists and is nonsingular, where

$$\mathbf{H}_n^* = [\mathbf{X}_n, \mathbf{W}_n(\mathbf{I}_n - \rho_0 \mathbf{W}_n)^{-1} \mathbf{X}_n \boldsymbol{\beta}_0] = [\mathbf{X}_n, \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0].$$

- (f) $\tilde{\rho}_n \xrightarrow{p} \rho_0$ and $\tilde{\boldsymbol{\beta}}_n \xrightarrow{p} \boldsymbol{\beta}_0$.^a

^aThis assumption is stronger than Assumption 6 in Lee (2003). Lee (2003) shows consistency under the assumption that $\tilde{\rho}_n$ is n^γ -consistent for some $\gamma > 0$.

Most of the assumptions in Assumption 6.10 are standard in spatial settings. However, unlike the S2SLSE from previous sections, the BS2SLSE must account for the stochastic nature of $\tilde{\mathbf{H}}_n^*$, which depends nonlinearly on the initial consistent estimate $\tilde{\rho}_n$. In particular, the uniform boundedness properties in (c) are imposed only at the true value of ρ_0 , i.e., a single point in the parameter space. To ensure validity, we require uniform boundedness to hold for all parameter values in a neighborhood of ρ_0 . Fortunately, Lee (pag. 313 2003)'s Lemma shows that condition (c) is sufficient to guarantee that the uniform boundedness property extends to a small neighborhood of the true parameter value.

Lemma 6.11 Suppose that $\{\|\mathbf{W}_n\|\}$ and $\{\|(\mathbf{I}_n - \rho_0 \mathbf{W}_n)^{-1}\|\}$, where $\|\cdot\|$ is a matrix norm, are bounded. Then, $\{\|(\mathbf{I}_n - \eta \mathbf{W}_n)^{-1}\|\}$ are uniformly bounded in η in a neighborhood of ρ_0 .

Proof Lemma 6.11. First, we rewrite the inverse $(\mathbf{I}_n - \eta \mathbf{W}_n)^{-1}$ in terms of $(\mathbf{I}_n - \rho_0 \mathbf{W}_n)^{-1}$. Let $\Delta = \eta - \rho_0$, then

$$\begin{aligned} \mathbf{I}_n - \eta \mathbf{W}_n &= \mathbf{I}_n - (\Delta + \rho_0) \mathbf{W}_n, \\ &= \mathbf{I}_n - \rho_0 \mathbf{W}_n - \Delta \mathbf{W}_n, \\ &= \mathbf{S}_0 - \Delta \mathbf{W}_n, \\ &= \mathbf{S}_0 (\mathbf{I}_n - \Delta \mathbf{W}_n \mathbf{S}_0^{-1}), \\ &= \mathbf{S}_0 (\mathbf{I}_n - \Delta \mathbf{G}_0), \end{aligned}$$

where $\mathbf{S}_0 = \mathbf{I}_n - \rho_0 \mathbf{W}_n$ and $\mathbf{G}_0 = \mathbf{W}_n \mathbf{S}_0^{-1}$ provided that \mathbf{G}_0 is invertible. Using Neumann series expansion for small Δ , we obtain

$$(\mathbf{I}_n - \Delta \mathbf{G}_0)^{-1} = \sum_{k=0}^{\infty} (\eta - \rho_0)^k \mathbf{G}_0^k.$$

From assumption, $\{\|\mathbf{W}_n\|\}$ is bounded, say by c_a , so that $\|\mathbf{W}_n\| \leq c_a$ for all n . Similarly, by assumption, $\{\|(\mathbf{I}_n - \rho_0 \mathbf{W}_n)^{-1}\|\}$ is bounded, say by c_b , so that $\|(\mathbf{I}_n - \rho_0 \mathbf{W}_n)^{-1}\| \leq c_b$. Thus, for sufficiently small $|\Delta|$, and since the geometric series $\sum_{k=0}^{\infty} (\eta - \rho_0)^k \mathbf{G}_0^k$ converges

to $(1 - |\Delta|c_ac_b)^{-1}$:

$$\begin{aligned} \|(\mathbf{I}_n - \Delta \mathbf{G}_0)^{-1}\| &\leq \sum_{k=0}^{\infty} |\eta - \rho_0|^k \|\mathbf{G}_0\|^k, \\ &\leq \sum_{k=0}^{\infty} |\eta - \rho_0|^k c_a^k c_b^k, \\ &= \frac{1}{1 - |\Delta|c_ac_b}, \\ &< \infty, \end{aligned}$$

for sufficiently small $|\Delta|$. Thus

$$\begin{aligned} \sup_{\eta \in \mathcal{B}} \|\mathbf{I}_n - \eta \mathbf{W}_n\| &\leq \|\mathbf{S}_0^{-1}\| \cdot \sup_{\eta \in \mathcal{B}} \|(\mathbf{I}_n - \Delta \mathbf{G}_0)^{-1}\| \\ &\leq \sup_{\eta \in \mathcal{B}} c_b \cdot \frac{1}{1 - |\Delta|c_ac_b}, \\ &< \infty, \end{aligned}$$

taking a close neighborhood \mathcal{B} contained in $B_1 = \{\eta : |\eta - \rho_0| < 1/(c_ac_b)\}$. ■

The following theorem provides the asymptotic distribution of the BS2SLSE for the SLM model.

Theorem 6.12 — Asymptotic Distribution of BS2SLSE for SLM. Under Assumption 6.10, the BS2SLS estimator defined in Equation (6.32) is consistent and

$$\sqrt{n} \left(\hat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0 \right) \xrightarrow{d} \mathbf{N}(\mathbf{0}, \boldsymbol{\Psi}_0^o),$$

where

$$\boldsymbol{\Psi}_0^o = \sigma_\epsilon^2 \left[\lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{H}_n^{*\top} \mathbf{H}_n^* \right]^{-1} = \sigma_\epsilon^2 \mathbf{J}^{*-1}. \quad (6.33)$$

The limiting variance $\boldsymbol{\Psi}_0^o$ in Equation (6.33) can be compared with the limiting distribution in Equation (6.27), which can be written as

$$\boldsymbol{\Omega}_0^o = \sigma_\epsilon^2 \left[\text{plim}_{n \rightarrow \infty} \frac{1}{n} \mathbf{Z}_n^\top \mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{Z}_n \right].$$

In proof 6.2.2, we showed that:

$$\text{plim}_{n \rightarrow \infty} \frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n = \text{plim}_{n \rightarrow \infty} \frac{1}{n} \mathbf{H}_n^\top \mathbb{E}(\mathbf{Z}_n) = \text{plim}_{n \rightarrow \infty} \mathbf{H}_n^\top \mathbf{H}_n^*.$$

In addition, the generalized Schwartz inequality implies that:⁵

$$\mathbf{H}_n^{*\top} \mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{H}_n^* \leq \mathbf{H}_n^{*\top} \mathbf{H}_n^*.$$

⁵For any matrices \mathbf{A} and \mathbf{B} , the generalized Schwarz inequality states that:

$$\mathbf{A}^\top \mathbf{B} (\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top \mathbf{A} \leq \mathbf{A}^\top \mathbf{A}.$$

This follows from the fact that the projection matrix $\mathbf{B} (\mathbf{B}^\top \mathbf{B})^{-1} \mathbf{B}^\top$ is positive semi-definite and represents an orthogonal projection onto the column space of \mathbf{B} , which cannot increase the norm of any vector.

Then,

$$\Omega_0^o = \sigma_\epsilon^2 \left[\text{plim}_{n \rightarrow \infty} \frac{1}{n} \mathbf{Z}_n^\top \mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{Z}_n \right] \geq \sigma_\epsilon^2 \left[\lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{H}_n^{*\top} \mathbf{H}_n^* \right]^{-1} = \Psi_0^o.$$

This inequality states that the variance matrix of the S2SLS estimator using \mathbf{H}_n as instruments is at least as large as the variance matrix of an estimator that directly uses the optimal instruments \mathbf{H}_n^* .

Sketch of Proof for Asymptotic Normality of BS2SLSE. This sketch of proof is based on Lee (2003). The sampling error of the BS2SLSE is

$$\sqrt{n}(\hat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0) = \left(\frac{1}{n} \widetilde{\mathbf{H}}_n^{*\top} \mathbf{Z}_n \right)^{-1} \left(\frac{1}{\sqrt{n}} \widetilde{\mathbf{H}}_n^{*\top} \boldsymbol{\varepsilon}_n \right), \quad (6.34)$$

where

$$\widetilde{\mathbf{H}}_n^* = \left[\mathbf{X}_n, \mathbf{W}_n (\mathbf{I}_n - \tilde{\rho}_n \mathbf{W}_n)^{-1} \mathbf{X}_n \tilde{\boldsymbol{\beta}}_n \right].$$

First step: First, we show that:

$$\text{plim} \left(\frac{1}{n} \widetilde{\mathbf{H}}_n^{*\top} \mathbf{Z}_n \right) = \text{plim} \frac{1}{n} \mathbf{H}_n^{*\top} \mathbf{H}_n^* = \mathbf{J}^*,$$

where $\mathbf{Z}_n = [\mathbf{X}_n, \mathbf{W}_n \mathbf{y}_n]$, and in the limit $\mathbf{H}_n^* = [\mathbf{X}_n, \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0] = \mathbb{E}(\mathbf{Z}_n)$ is the population expectation of the reduced form equation, where $\mathbf{G}_0 = \mathbf{W}_n \mathbf{S}_0^{-1}$ and $\mathbf{S}_0 = (\mathbf{I}_n - \rho_0 \mathbf{W}_n)$.

Let $\mathbf{W}_n (\mathbf{I}_n - \tilde{\rho} \mathbf{W}_n)^{-1} = \mathbf{W}_n \mathbf{S}_n^{-1}(\tilde{\rho}_n) = \mathbf{G}_n(\tilde{\rho}_n)$. By definition of $\widetilde{\mathbf{H}}_n^*$:

$$\frac{1}{n} \widetilde{\mathbf{H}}_n^{*\top} \mathbf{Z}_n = \frac{1}{n} \left[\mathbf{X}_n, \mathbf{G}_n(\tilde{\rho}_n) \mathbf{X}_n \tilde{\boldsymbol{\beta}}_n \right]^\top \mathbf{Z}_n.$$

Since $\mathbf{Z}_n = [\mathbf{X}_n, \mathbf{W}_n \mathbf{y}_n]$, and the reduced form equation at the true parameters is $\mathbf{y}_n = \mathbf{S}_0^{-1} \mathbf{X}_n \boldsymbol{\beta}_0 + \mathbf{S}_0^{-1} \boldsymbol{\varepsilon}_n$, the previous expression can be written as

$$\frac{1}{n} \widetilde{\mathbf{H}}_n^{*\top} \mathbf{Z}_n = \frac{1}{n} \left[\mathbf{X}_n, \mathbf{G}_n(\tilde{\rho}_n) \mathbf{X}_n \tilde{\boldsymbol{\beta}}_n \right]^\top [\mathbf{X}_n, \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0 + \mathbf{G}_0 \boldsymbol{\varepsilon}_n].$$

We now analyze the asymptotic behavior of $\frac{1}{n} \widetilde{\mathbf{H}}_n^{*\top} \mathbf{Z}_n$ as $n \rightarrow \infty$. Clearly,

$$\frac{1}{n} \mathbf{X}_n^\top \mathbf{X}_n \rightarrow \lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{X}_n^\top \mathbf{X}_n,$$

which is bounded by Assumption 6.10(d). Similarly, since $\mathbb{E}(\boldsymbol{\varepsilon}_n) = \mathbf{0}$,

$$\frac{1}{n} (\mathbf{X}_n^\top \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0 + \mathbf{X}_n^\top \mathbf{G}_0 \boldsymbol{\varepsilon}_n) \xrightarrow{p} \frac{1}{n} \mathbf{X}_n^\top \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0 + \frac{1}{n} \mathbf{X}_n^\top \mathbf{G}_0 \mathbb{E}(\boldsymbol{\varepsilon}_n) = \frac{1}{n} \mathbf{X}_n^\top \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0.$$

Unlike the S2SLSE proposed by Kelejian and Prucha (1998), the matrix of instruments is stochastic. Thus, $\frac{1}{n} \tilde{\boldsymbol{\beta}}_n^\top \mathbf{X}_n^\top \mathbf{G}_n(\tilde{\rho}_n)^\top \mathbf{X}_n$ needs to be expanded around ρ_0 to establish whether $\mathbf{G}_n(\tilde{\rho}_n) \xrightarrow{p} \mathbf{G}_0$. Note that $\mathbf{S}_n(\tilde{\rho}_n)^{-1} - \mathbf{S}_0^{-1}$ can be written as

$$\begin{aligned} \mathbf{S}_n(\tilde{\rho}_n)^{-1} - \mathbf{S}_0^{-1} &= \mathbf{S}_n(\tilde{\rho}_n)^{-1} \mathbf{S}_0 \mathbf{S}_0^{-1} - \mathbf{S}_n(\tilde{\rho}_n)^{-1} \mathbf{S}_n(\tilde{\rho}_n) \mathbf{S}_0^{-1}, \\ &= \mathbf{S}_n(\tilde{\rho}_n)^{-1} (\mathbf{S}_0 \mathbf{S}_0^{-1} - \mathbf{S}_n(\tilde{\rho}_n) \mathbf{S}_0^{-1}), \\ &= \mathbf{S}_n(\tilde{\rho}_n)^{-1} [\mathbf{S}_0 - \mathbf{S}_n(\tilde{\rho}_n)] \mathbf{S}_0^{-1}. \end{aligned} \quad (6.35)$$

Since $\partial \mathbf{S}(\eta)/\partial \eta = -\mathbf{W}_n$, a **first-order expansion** for $\mathbf{S}_n(\tilde{\rho}_n)$ around ρ_0 is

$$\mathbf{S}_n(\tilde{\rho}_n) = \mathbf{S}_0 - \mathbf{W}_n(\tilde{\rho}_n - \rho_0). \quad (6.36)$$

Inserting Equation (6.36) into (6.35), and pre-multiplying by \mathbf{W}_n yields:

$$\begin{aligned} \mathbf{W}_n(\mathbf{I}_n - \tilde{\rho}_n \mathbf{W}_n)^{-1} &= \mathbf{W}_n(\mathbf{I}_n - \rho_0 \mathbf{W}_n)^{-1} + \mathbf{W}_n \mathbf{S}_n(\tilde{\rho}_n)^{-1} \mathbf{W}_n \mathbf{S}_0^{-1} (\tilde{\rho}_n - \rho_0), \\ \mathbf{G}_n(\tilde{\rho}_n) &= \mathbf{G}_0 + (\tilde{\rho}_n - \rho_0) \mathbf{G}_n(\tilde{\rho}_n) \mathbf{G}_0. \end{aligned}$$

Because \mathbf{G}_0 is uniformly bounded in both row and column sum and $\mathbf{G}_n(\eta)$ is **uniformly bounded** in row sums, uniformly in η belonging to its parameter space, the matrices in the previous expansion are uniformly bounded.⁶ Furthermore, since $(\tilde{\rho}_n - \rho_0) = o_p(1)$, we have that:

$$\mathbf{G}_n(\tilde{\rho}_n) - \mathbf{G}_0 = o_p(1).$$

Therefore, since $\tilde{\beta}_n \xrightarrow{p} \beta_0$

$$\begin{aligned} \frac{1}{n} \tilde{\beta}_n^\top \mathbf{X}_n^\top \mathbf{G}_n(\tilde{\rho}_n)^\top \mathbf{X}_n &\xrightarrow{p} \frac{1}{n} \beta_0^\top \mathbf{X}_n^\top \mathbf{G}_0^\top \mathbf{X}_n, \\ \frac{1}{n} \tilde{\beta}_n^\top \mathbf{X}_n^\top \mathbf{G}_n(\tilde{\rho}_n)^\top \mathbf{G}_0 \boldsymbol{\varepsilon}_n &\xrightarrow{p} \frac{1}{n} \beta_0^\top \mathbf{X}_n^\top \mathbf{G}_0^\top \mathbf{G}_0 \mathbb{E}(\boldsymbol{\varepsilon}_n) = 0. \end{aligned}$$

Collecting the results, we have

$$\begin{aligned} \frac{1}{n} \widetilde{\mathbf{H}}_n^{*\top} \mathbf{Z}_n &\xrightarrow{p} \lim_{n \rightarrow \infty} \frac{1}{n} [\mathbf{X}_n^\top, \beta_0^\top \mathbf{X}_n^\top \mathbf{G}_0^\top] [\mathbf{X}_n, \mathbf{G}_0 \mathbf{X}_n \beta_0], \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} [\mathbf{X}_n, \mathbf{W}_n(\mathbf{I}_n - \rho_0 \mathbf{W}_n)^{-1} \mathbf{X}_n \beta_0]^\top \mathbb{E}(\mathbf{Z}_n), \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{H}_n^{*\top} \mathbf{H}_n^*, \end{aligned}$$

which can also be written as

$$\text{plim} \frac{1}{n} \widetilde{\mathbf{H}}_n^{*\top} \mathbf{Z}_n = \lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{H}_n^{*\top} \mathbf{H}_n^* = \mathbf{J}^*$$

Second step: We need to show that

$$\frac{1}{\sqrt{n}} \widetilde{\mathbf{H}}_n^{*\top} \boldsymbol{\varepsilon}_n \xrightarrow{d} \mathbf{N}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{J}^{*-1}).$$

Since $\hat{\beta}_n \xrightarrow{p} \beta_0$,

$$\frac{1}{\sqrt{n}} \widetilde{\mathbf{H}}_n^{*\top} \boldsymbol{\varepsilon}_n = \frac{1}{\sqrt{n}} [\mathbf{X}_n, \mathbf{G}_0 \mathbf{X}_n \beta_0]^\top \boldsymbol{\varepsilon}_n + o_p(1) = \frac{1}{\sqrt{n}} \mathbf{H}_n^{*\top} \boldsymbol{\varepsilon}_n + o_p(1).$$

Since $\lim_{n \rightarrow \infty} n^{-1} \mathbf{H}_n^{*\top} \mathbf{H}_n^*$ exists and it is invertible, by Theorem 3.30, we have limiting distribution. ■

6.2.8 Coding BS2SLS Estimator

To compute the BS2SLS estimator, we modify our previous function, `slm.2sls`, and create a new function called `slm.b2sls`.

⁶Since $\mathbf{G}_n(\tilde{\rho}_n)$ is a non-linear function of $\tilde{\rho}_n$ we need to assume uniformly boundedness in a neighborhood of ρ_0 . See Lemma 2 in Lee (2003).

```

# Main function to estimate S2SLS estimator
slm.b2spls <- function(formula, data, W, instruments = 2){
  # Model frame setup
  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())

  # Get variables and check dimensions
  y <- model.response(mf)
  X <- model.matrix(formula, mf)
  n <- nrow(X)
  Wy <- W %%% y
  sn <- nrow(W)
  if (n != sn) stop("number of spatial units in W is different to the number of data")

  # Obtain first (and consistent) estimates
  H <- cbind(X, make.H(W = W, X = X, l = instruments))

  # Compute S2SLSE
  Z      <- cbind(X, Wy)
  colnames(Z) <- c(colnames(X), "Wy")
  HH     <- crossprod(H, H)
  PH     <- H %%% solve(HH) %%% t(H)
  Z_hat  <- PH %%% Z
  b_2spls <- solve(crossprod(Z_hat)) %%% crossprod(Z_hat, y)

  # Second step: BS2SLS estimation
  beta.hat <- b_2spls[1:ncol(X)]
  rho.hat  <- drop(tail(b_2spls, n = 1L))
  H.lee    <- W %%% solve(diag(n) - rho.hat * W) %%% X %%% beta.hat
  H.star   <- cbind(X, H.lee)
  b_2spls  <- solve(crossprod(H.star, Z)) %%% crossprod(H.star, y)

  # Compute residuals
  y_hat    <- Z %%% b_2spls
  e_hat    <- y - y_hat

  # Save results
  results <- structure(
    list(
      coefficients = b_2spls,
      call         = callT,
      X            = X,

```

```

      H          = H.star,
      Z          = Z,
      y          = y,
      PH         = PH,
      e_hat      = e_hat,
      W          = W
    ),
    class = 'mybs2sls'
  )
}

```

The `slm.b2sls` function first obtains the S2SLS estimates using the instruments proposed by Kelejian and Prucha (1998) instruments. It then, uses these consistent estimates to construct the optimal instrument \widetilde{H}_n^* . The BS2SLSE is then obtained using Equation (6.32).

The S3 method `vcov` is constructed using Equation (6.33) and can be applied to either the initial or final round estimates.

```

# S3 Method vcov
vcov.mybs2sls <- function(object, estimate = c("initial", "final"), ...){
  estimate <- match.arg(estimate)
  X          <- object$X
  n          <- nrow(X)
  k          <- ncol(X)
  df         <- n - (k + 1)
  s2         <- crossprod(object$e_hat) / df
  if (estimate == "final"){
    b          <- object$coefficients
    b.hat      <- b[1:k]
    rho.hat    <- drop(tail(b, n = 1L))
    W          <- object$W
    H.lee      <- W %*% solve(diag(n) - rho.hat * W) %*% X %*% b.hat
    H.star     <- cbind(X, H.lee)
  } else {
    H.star     <- object$H
  }
  var        <- drop(s2) * solve(crossprod(H.star) / n) / n
  return(var)
}

```

The S3 methods for `summary` are the following:

```

# S3 methods for summary
summary.mybs2sls <- function(object,
                              estimate = c("initial", "final"),
                              table = TRUE,
                              digits = max(3, .Options$digits - 3),
                              ...){

```

```

estimate <- match.arg(estimate)
n        <- nrow(object$Z)
df        <- n - ncol(object$Z)
b         <- object$coefficients
std.err   <- sqrt(diag(vcov(object, estimate = estimate)))
z         <- b / std.err
p         <- 2 * pt(-abs(z), df = df)
CoefTable <- cbind(b, std.err, z, p)
colnames(CoefTable) <- c("Estimate", "Std.Error", "t-value", "Pr(>|t|)")
result <- structure(
  list(
    CoefTable = CoefTable,
    digits    = digits,
    call      = object$call),
  class = 'summary.mybs2sls'
)
return(result)
}

print.summary.mybs2sls <- function(x,
                                digits = x$digits,
                                na.print = "",
                                symbolic.cor = p > 4,
                                signif.stars = getOption("show.signif.stars"),
                                ...)
{
  cat("\nCall:\n")
  cat(paste(deparse(x$call), sep = "\n", collapse = "\n"), "\n\n", sep = "")

  cat("\nCoefficients:\n")
  printCoefmat(x$CoefTable, digit = digits, P.value = TRUE, has.Pvalue = TRUE)
  invisible(NULL)
}

```

Using the same DGP as for the S2SLS, we obtain the BS2SLS estimates:

```

# Using the BS2SLS estimator
b2sls <- slm.b2sls(y ~ x1 + x2 + x3, data = data, W = W, instruments = 2)
summary(b2sls, estimate = "initial")

##
## Call:
## slm.b2sls(formula = y ~ x1 + x2 + x3, data = data, W = W, instruments = 2)
##
##
## Coefficients:
##             Estimate Std.Error t-value Pr(>|t|)

```

```
## (Intercept) -0.08156    0.06604   -1.235    0.217
## x1          -1.01266    0.06417  -15.781   <2e-16 ***
## x2          -0.04195    0.06291   -0.667    0.505
## x3           1.02569    0.06655   15.413   <2e-16 ***
## Wy           0.61460    0.05803   10.591   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(b2sls, estimate = "final")

##
## Call:
## slm.b2sls(formula = y ~ x1 + x2 + x3, data = data, W = W, instruments = 2)
##
##
## Coefficients:
##              Estimate Std. Error t-value Pr(>|t|)
## (Intercept) -0.08156    0.06631   -1.230    0.219
## x1          -1.01266    0.06405  -15.811   <2e-16 ***
## x2          -0.04195    0.06290   -0.667    0.505
## x3           1.02569    0.06638   15.452   <2e-16 ***
## Wy           0.61460    0.06032   10.189   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6.3 GMM Estimator of SLM

6.3.1 GMM Estimator Under Homoskedasticity

In the previous sections, we analyzed the S2SLS and BS2SLS estimators for the Spatial Lag Model (SLM). Both estimators belong to the broader class of Generalized Method of Moments (GMM) estimators. This section reviews the GMM estimator for the SLM, as proposed by [Lee \(2007\)](#), and derives the GMM and Optimal GMM (OGMM) estimators for the model. This GMM estimator can be considered as an one-step GMM estimator (see Section 6.1.2).

To construct the moment conditions, [Lee \(2007\)](#) uses linear and quadratic moments. Let \mathbf{H}_n denote an $n \times k_x$ matrix of instrumental variables (IVs) constructed as a function of \mathbf{X}_n and \mathbf{W}_n . This matrix can be constructed using the instruments proposed by [Kelejian and Prucha \(1998\)](#), such that

$$\mathbf{H}_n = [\mathbf{X}_n, \mathbf{W}_n \mathbf{X}_n, \mathbf{W}_n^2 \mathbf{X}_n, \dots],$$

or using the asymptotically optimal instruments proposed by [Lee \(2003\)](#):

$$\mathbf{H}_n = [\mathbf{X}_n, \mathbf{W}_n (\mathbf{I}_n - \tilde{\rho}_n \mathbf{W}_n)^{-1} \mathbf{X}_n \tilde{\beta}_n],$$

where $\tilde{\rho}_n$ and $\tilde{\beta}_n$ are consistent estimates of ρ_0 and β_0 .

Recall the error term for the SLM is defined as

$$\boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) = \mathbf{y}_n - \mathbf{Z}_n \boldsymbol{\delta}_n = (\mathbf{I}_n - \rho \mathbf{W}_n) \mathbf{y}_n - \mathbf{X}_n \boldsymbol{\beta}.$$

Unlike the S2SLSE, the GMME uses additional quadratic moments. Because of this, the optimally weighted GMM estimator can be asymptotically efficient relative to the S2SLSE. For the quadratic population moments, [Lee \(2007\)](#) proposes:

$$\mathbb{E}(n^{-1} \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{jn} \boldsymbol{\varepsilon}_n) = 0, \quad (6.37)$$

where \mathbf{P}_{jn} are $n \times n$ matrix of constant having **zero trace**. [Lee \(2007\)](#) refers to this class of matrices as $\mathcal{P}_{1n} = \{\mathbf{P} : \mathbf{P} \text{ is } n \times n \text{ matrix, } \text{tr}(\mathbf{P}) = 0\}$. A subclass \mathcal{P}_{2n} arises for $n \times n$ matrices having zero diagonal: $\mathcal{P}_{2n} = \{\mathbf{P} : \mathbf{P} \text{ is } n \times n \text{ matrix, } \text{diag}(\mathbf{P}) = \mathbf{0}\}$.

To understand the role played by \mathbf{P}_{jn} , note for any nonstochastic matrix \mathbf{A}_n , the quadratic moments are given by:

$$\begin{aligned} \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) &= \text{tr}[\mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n)], \\ &= \mathbb{E}[\text{tr}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n)], \\ &= \mathbb{E}[\text{tr}(\mathbf{A}_n \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^\top)], \\ &= \text{tr}[\mathbb{E}(\mathbf{A}_n \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^\top)], \\ &= \sigma_0^2 \text{tr}(\mathbf{A}_n). \end{aligned}$$

Thus $\mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) \neq 0$, unless $\text{tr}(\mathbf{A}_n) = 0$. If $\text{tr}(\mathbf{A}_n) \neq 0$, we can always create for example $\mathbf{P}_n = \mathbf{A}_n - \text{Diag}(\mathbf{P}_n)$ such that $\text{tr}(\mathbf{P}_n) = 0$.

It is important to stress the intuitions of the instruments. The regressor $\mathbf{W}_n \mathbf{y}_n$ is endogenous, as

$$\mathbb{E}[(\mathbf{W}_n \mathbf{y}_n)^\top \boldsymbol{\varepsilon}_n] = \sigma_0^2 \text{tr}(\mathbf{G}_n(\rho_0)) \neq 0,$$

where $\mathbf{G}_n(\rho_0) = \mathbf{W}_n \mathbf{S}_0^{-1}$, i.e., the elements of $\mathbf{W}_n \mathbf{y}_n$ are correlated with the elements of $\boldsymbol{\varepsilon}_n$. This can be observed from the expression $\mathbf{W}_n \mathbf{y}_n = \mathbf{G}_n(\rho_0) \mathbf{X}_n \boldsymbol{\beta}_0 + \mathbf{G}_n(\rho_0) \boldsymbol{\varepsilon}_n$, which follows from the reduced form $\mathbf{y}_n = \mathbf{S}_0^{-1} \mathbf{X}_n \boldsymbol{\beta} + \mathbf{S}_0^{-1} \boldsymbol{\varepsilon}_n$. The linear instruments \mathbf{H}_n is correlated with $\mathbf{G}_n(\rho_0) \mathbf{X}_n \boldsymbol{\beta}_0$ (since \mathbf{H}_n are mean of $\mathbf{W}_n \mathbf{y}_n$), but uncorrelated with $\boldsymbol{\varepsilon}_n$, because $\mathbb{E}(\mathbf{H}_n^\top \boldsymbol{\varepsilon}_n) = \mathbf{H}_n^\top \mathbb{E}(\boldsymbol{\varepsilon}_n) = \mathbf{0}$. Note that we require \mathbf{P}_{jn} to be correlated with \mathbf{G}_0 in the sense that $\text{tr}(\mathbf{P}_{jn} \mathbf{G}_0) \neq 0$. As long as $\text{tr}(\mathbf{P}_{jn}) = 0$, $\mathbf{P}_{jn} \boldsymbol{\varepsilon}_n$ is uncorrelated with $\boldsymbol{\varepsilon}_n$, and thus it may be used as an instrument of $\mathbf{W}_n \mathbf{y}_n$.

The quadratic moments in Equation (6.37) can be derived under homoskedasticity, where $\mathbb{E}(\boldsymbol{\varepsilon}_n^2) = \sigma_0^2$ ([Lee, 2001](#)). For example, we can use the following moment conditions provided by [Kelejian and Prucha \(1999\)](#):

$$\begin{aligned} \mathbb{E}\left[\frac{1}{n} \boldsymbol{\varepsilon}_n^\top \boldsymbol{\varepsilon}_n\right] &= \sigma_0^2, \\ \mathbb{E}\left[\frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{W}_n \mathbf{W}_n \boldsymbol{\varepsilon}_n\right] &= \sigma_0^2 \frac{1}{n} \text{tr}(\mathbf{W}_n^\top \mathbf{W}_n), \\ \mathbb{E}\left[\frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{W}_n \boldsymbol{\varepsilon}_n\right] &= 0, \end{aligned} \quad (6.38)$$

These three moment conditions can be reduced to two. Substituting out σ_0^2 into the

second moment equation yields:

$$\begin{aligned}
 \mathbb{E} \left[\frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{W}_n \mathbf{W}_n \boldsymbol{\varepsilon}_n \right] - \mathbb{E} \left[\frac{1}{n} \boldsymbol{\varepsilon}_n^\top \boldsymbol{\varepsilon}_n \right] \frac{1}{n} \text{tr}(\mathbf{W}_n^\top \mathbf{W}_n) &= 0, \\
 \frac{1}{n} \mathbb{E} \left[\boldsymbol{\varepsilon}_n^\top \mathbf{W}_n \mathbf{W}_n \boldsymbol{\varepsilon}_n - \boldsymbol{\varepsilon}_n^\top \boldsymbol{\varepsilon}_n \frac{1}{n} \text{tr}(\mathbf{W}_n^\top \mathbf{W}_n) \right] &= 0, \\
 \frac{1}{n} \mathbb{E} \left[\boldsymbol{\varepsilon}_n^\top \mathbf{W}_n \mathbf{W}_n \boldsymbol{\varepsilon}_n - \boldsymbol{\varepsilon}_n^\top \frac{1}{n} \text{tr}(\mathbf{W}_n^\top \mathbf{W}_n) \boldsymbol{\varepsilon}_n \right] &= 0, \\
 \frac{1}{n} \mathbb{E} \left[\boldsymbol{\varepsilon}_n^\top \left(\mathbf{W}_n \mathbf{W}_n - \frac{1}{n} \text{tr}(\mathbf{W}_n^\top \mathbf{W}_n) \mathbf{I}_n \right) \boldsymbol{\varepsilon}_n \right] &= 0, \\
 \frac{1}{n} \mathbb{E} [\boldsymbol{\varepsilon}_n^\top \mathbf{P}_{2n} \boldsymbol{\varepsilon}_n] &= 0,
 \end{aligned}$$

where \mathbf{P}_{2n} is symmetric with $\text{tr}(\mathbf{P}_{2n}) = 0$, but its diagonal elements are non zero (In the heteroskedasticity case it is!). Based on these results, [Lee \(2007\)](#) proposes:

$$\begin{aligned}
 \mathbf{P}_{1n} &= \mathbf{W}_n, \\
 \mathbf{P}_{2n} &= \mathbf{W}_n^2 - (\text{tr}(\mathbf{W}_n^2)/n) \mathbf{I}_n.
 \end{aligned}$$

The vector of moments functions is then defined as:

$$\underset{(2+k_x) \times 1}{\mathbf{g}_n(\boldsymbol{\theta})} = \begin{pmatrix} \boldsymbol{\varepsilon}_n(\boldsymbol{\theta})^\top \mathbf{P}_{1n} \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) \\ \boldsymbol{\varepsilon}_n(\boldsymbol{\theta})^\top \mathbf{P}_{2n} \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) \\ \mathbf{H}_n^\top \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) \end{pmatrix}, \quad (6.39)$$

which has dimension $(2 + k_k) \times 1$. It is important to note that the moment conditions at the population hold. At $\boldsymbol{\theta}_0$:

$$\begin{aligned}
 \mathbb{E}(\mathbf{H}_n^\top \boldsymbol{\varepsilon}_n) &= \mathbf{H}_n^\top \mathbb{E}(\boldsymbol{\varepsilon}_n) = \mathbf{0}, \\
 \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{P}_{jn} \boldsymbol{\varepsilon}_n) &= \sigma_0^2 \text{tr}(\mathbf{P}_{jn}) = \mathbf{0} \text{ for } j = 1, 2
 \end{aligned}$$

Let $\boldsymbol{\Upsilon}_n$ be some $(2 + k_k) \times (2 + k_k)$ symmetric positive semidefinite weighting matrix, then the corresponding GMM estimator is defined as:

$$\hat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\text{argmin}} \quad Q_n(\boldsymbol{\theta}) = \mathbf{g}_n(\boldsymbol{\theta})^\top \boldsymbol{\Upsilon}_n \mathbf{g}_n(\boldsymbol{\theta}), \quad (6.40)$$

where $\boldsymbol{\theta} = (\rho, \boldsymbol{\beta}^\top)^\top$ is an $(1 + k)$ -dimensional vector, and where $\mathbf{g}_n(\boldsymbol{\theta})$ is the $(2 + k_x) \times 1$ vector of moments.

Since $\mathbb{E}(\mathbf{g}_n) = \mathbf{0}$, the variance-covariance matrix of the population moment functions are given by:

$$\begin{aligned}
 \boldsymbol{\Omega}_n &= \mathbb{V}(\mathbf{g}_n) \\
 &= \mathbb{E}(\mathbf{g}_n \mathbf{g}_n^\top) \\
 &= \mathbb{E} \begin{pmatrix} \boldsymbol{\varepsilon}_n(\boldsymbol{\theta})^\top \mathbf{P}_{1n} \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) \cdot (\boldsymbol{\varepsilon}_n(\boldsymbol{\theta})^\top \mathbf{P}_{1n} \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}))^\top & \boldsymbol{\varepsilon}_n(\boldsymbol{\theta})^\top \mathbf{P}_{1n} \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) \cdot (\boldsymbol{\varepsilon}_n(\boldsymbol{\theta})^\top \mathbf{P}_{2n} \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}))^\top & \boldsymbol{\varepsilon}_n(\boldsymbol{\theta})^\top \mathbf{P}_{1n} \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) \cdot \boldsymbol{\varepsilon}_n(\boldsymbol{\theta})^\top \mathbf{H}_n \\ \boldsymbol{\varepsilon}_n(\boldsymbol{\theta})^\top \mathbf{P}_{2n} \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) \cdot (\boldsymbol{\varepsilon}_n(\boldsymbol{\theta})^\top \mathbf{P}_{1n} \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}))^\top & \boldsymbol{\varepsilon}_n(\boldsymbol{\theta})^\top \mathbf{P}_{2n} \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) \cdot (\boldsymbol{\varepsilon}_n(\boldsymbol{\theta})^\top \mathbf{P}_{2n} \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}))^\top & \boldsymbol{\varepsilon}_n(\boldsymbol{\theta})^\top \mathbf{P}_{2n} \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) \cdot \boldsymbol{\varepsilon}_n(\boldsymbol{\theta})^\top \mathbf{H}_n \\ \mathbf{H}_n^\top \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) \cdot (\boldsymbol{\varepsilon}_n(\boldsymbol{\theta})^\top \mathbf{P}_{1n} \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}))^\top & \mathbf{H}_n^\top \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) \cdot (\boldsymbol{\varepsilon}_n(\boldsymbol{\theta})^\top \mathbf{P}_{2n} \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}))^\top & \mathbf{H}_n^\top \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) \cdot \boldsymbol{\varepsilon}_n(\boldsymbol{\theta})^\top \mathbf{H}_n \end{pmatrix}
 \end{aligned}$$

To find these moments, we use Lemma 3.26. Thus, using the fact that $\text{tr}(\mathbf{P}_{jn}) = 0$, for $j = 1, 2$:

$$\begin{aligned}\mathbb{E} [\mathbf{H}_n^\top \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) \cdot \boldsymbol{\varepsilon}_n(\boldsymbol{\theta})^\top \mathbf{P}_{jn} \boldsymbol{\varepsilon}_n(\boldsymbol{\theta})] &= \mathbf{H}_n^\top \text{diag}(\mathbf{P}_{jn}) \mu_3, \\ \mathbb{E} [\boldsymbol{\varepsilon}(\boldsymbol{\theta})^\top \mathbf{P}_{jn} \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) \cdot \boldsymbol{\varepsilon}(\boldsymbol{\theta})^\top \mathbf{P}_{ln} \boldsymbol{\varepsilon}_n(\boldsymbol{\theta})] &= (\mu_4 - 3\sigma_0^4) \text{diag}(\mathbf{P}_{jn})^\top \text{diag}(\mathbf{P}_{ln}) + \sigma_0^4 [\text{tr}(\mathbf{P}_{jn}) \text{tr}(\mathbf{P}_{ln}) + \text{tr}(\mathbf{P}_{jn} \mathbf{P}_{ln}^s)], \\ &= (\mu_4 - 3\sigma_0^4) \text{diag}(\mathbf{P}_{jn})^\top \text{diag}(\mathbf{P}_{ln}) + \sigma_0^4 \text{tr}(\mathbf{P}_{jn} \mathbf{P}_{ln}^s),\end{aligned}$$

where $\mu_3 = \mathbb{E}(\epsilon_{ni}^3)$, $\mu_4 = \mathbb{E}(\epsilon_{ni}^4)$, and $\mathbf{P}_{jn}^s = \mathbf{P}_{jn} + \mathbf{P}_{jn}^\top$. Thus, we can write:

$$\boldsymbol{\Omega}_n = \begin{pmatrix} (\mu_4 - 3\sigma_0^4) \boldsymbol{\omega}_n^\top \boldsymbol{\omega}_n & \mu_3 \boldsymbol{\omega}_n^\top \mathbf{H}_n \\ (2 \times 2) & (2 \times k_x) \\ \mu_3 \mathbf{H}_n^\top \boldsymbol{\omega}_n & \mathbf{O} \\ (k_x \times 2) & (k_x \times k_x) \end{pmatrix} + \mathbf{V}_n,$$

with $\boldsymbol{\omega}_n = [\text{diag}(\mathbf{P}_{1n}), \text{diag}(\mathbf{P}_{2n})]$ is $n \times 2$ and

$$\mathbf{V}_n = \sigma_0^4 \begin{pmatrix} \text{tr}(\mathbf{P}_{1n} \mathbf{P}_{1n}^s) & \text{tr}(\mathbf{P}_{1n} \mathbf{P}_{2n}^s) & \mathbf{0} \\ (1 \times 1) & (1 \times 1) & (1 \times k_x) \\ \text{tr}(\mathbf{P}_{2n} \mathbf{P}_{1n}^s) & \text{tr}(\mathbf{P}_{2n} \mathbf{P}_{2n}^s) & \mathbf{0} \\ (1 \times k_x) & (1 \times k_x) & (1 \times k_x) \\ \mathbf{0} & \mathbf{0} & \frac{1}{\sigma_0^2} \mathbf{H}_n^\top \mathbf{H}_n \\ (k_x \times 1) & (k_x \times 1) & (k_x \times k_x) \end{pmatrix}.$$

As Lee (2007) remarks, when $\boldsymbol{\varepsilon}_n$ is normally distributed, $\boldsymbol{\Omega}_n$ is simplified to \mathbf{V}_n because $\mu_3 = 0$, and $\mu_4 = 3\sigma_0^4$.

For further reference note that the first derivatives of the moment functions in Equation (6.39) are:

$$\begin{aligned}\frac{\partial \mathbf{g}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} &= \begin{pmatrix} \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{1n}^s \\ \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{2n}^s \\ \mathbf{H}_n^\top \end{pmatrix} \frac{\partial \boldsymbol{\varepsilon}_n}{\partial \boldsymbol{\theta}^\top}, \\ ((2+k_x) \times (k+1)) & \\ &= \begin{pmatrix} \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{1n}^s \\ \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{2n}^s \\ \mathbf{H}_n^\top \end{pmatrix} \begin{pmatrix} -\mathbf{W}_n \mathbf{y}_n & -\mathbf{X}_n \end{pmatrix}. \\ ((2+k_x) \times n) & \quad (n \times (1+k)) \end{aligned} \tag{6.41}$$

Assumption 6.13 — Assumptions for GMM (Lee, 2007). Assume the following:

- (a) The ϵ_{ni} are i.i.d with zero mean, variance σ_0^2 and that a moment of order higher than the fourth exists.
- (b) The elements of \mathbf{X}_n are uniformly bounded constants, \mathbf{X}_n has the full rank k , and $\lim_{n \rightarrow \infty} (1/n) \mathbf{X}_n^\top \mathbf{X}_n$ exists and is nonsingular.
- (c) The spatial weights matrices $\{\mathbf{W}_n\}$ and $\{(\mathbf{I}_n - \rho \mathbf{W}_n)^{-1}\}$ at $\rho = \rho_0$ are uniformly bounded in both row and column sums in absolute value.
- (d) The matrices \mathbf{P}_{1n} and \mathbf{P}_{2n} are uniformly bonded in both row and column sums in absolute value, and elements of \mathbf{H}_n are uniformly bounded.
- (e) Either

- (a) $\lim_{n \rightarrow \infty} (1/n) \mathbf{H}_n^\top [\mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0, \mathbf{X}_n]$ has the full rank $(k+1)$, or

(b) $\lim_{n \rightarrow \infty} (1/n) \mathbf{H}_n^\top \mathbf{X}_n$ has the full rank k , $\lim_{n \rightarrow \infty} (1/n) \text{tr}(\mathbf{P}_{jn}^s \mathbf{G}_0) \neq 0$ for some j , $\mathbf{P}_{jn}^s = \mathbf{P}_{jn} + \mathbf{P}_{jn}^\top$, and

$$\lim_{n \rightarrow \infty} (1/n) [\text{tr}(\mathbf{P}_{1n}^s \mathbf{G}_0), \text{tr}(\mathbf{P}_{2n}^s \mathbf{G}_0)]^\top,$$

is linearly independent of

$$\lim_{n \rightarrow \infty} (1/n) [\text{tr}(\mathbf{G}_0^\top \mathbf{P}_{1n} \mathbf{G}_0), \text{tr}(\mathbf{G}_0^\top \mathbf{P}_{2n} \mathbf{G}_0)]^\top$$

where $\mathbf{G}_0 = \mathbf{W}_n \mathbf{S}_0^{-1}$.

Theorem 6.14 — GMM estimator for SLM under homoskedasticity (Lee, 2007). Let $\mathbf{Y}_n \rightarrow \mathbf{Y}_0$. Under Assumptions 6.13, suppose that \mathbf{P}_{jn} for $j = 1, 2$, are from \mathcal{P}_{1n} and \mathbf{H}_n is a $n \times k_k$ matrix so that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \mathbb{E}[\mathbf{g}_n(\boldsymbol{\theta})] = \mathbf{0},$$

has a unique root at $\boldsymbol{\theta}_0$ in $\boldsymbol{\Theta}$. Then, the GMM estimator $\hat{\boldsymbol{\theta}}_n$ derived from $\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbf{g}_n(\boldsymbol{\theta})^\top \mathbf{Y}_n \mathbf{g}_n(\boldsymbol{\theta})$ is a consistent estimator of $\boldsymbol{\theta}_0$, and $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_0)$, where

$$\begin{aligned} \boldsymbol{\Sigma}_0 &= \lim_{n \rightarrow \infty} \left[\left(\frac{1}{n} \mathbf{D}_n^\top \right) \mathbf{Y}_0 \left(\frac{1}{n} \mathbf{D}_n \right) \right]^{-1} \left(\frac{1}{n} \mathbf{D}_n^\top \right) \mathbf{Y}_0 \left(\frac{1}{n} \boldsymbol{\Omega}_n \right) \mathbf{Y}_0 \left(\frac{1}{n} \mathbf{D}_n \right) \\ &\quad \times \left[\left(\frac{1}{n} \mathbf{D}_n^\top \right) \mathbf{Y}_0 \left(\frac{1}{n} \mathbf{D}_n \right) \right]^{-1}, \end{aligned}$$

and

$$\mathbf{D}_n = \frac{\partial \mathbb{E}[\mathbf{g}_n(\boldsymbol{\theta}_0)]}{\partial \boldsymbol{\theta}^\top} = \begin{pmatrix} \sigma_0^2 \text{tr}(\mathbf{P}_{1n}^s \mathbf{G}_0) & \mathbf{0} \\ \sigma_0^2 \text{tr}(\mathbf{P}_{2n}^s \mathbf{G}_0) & \mathbf{0} \\ \mathbf{H}_n^\top \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0 & \mathbf{H}_n^\top \mathbf{X}_n \end{pmatrix} \quad (6.42)$$

where $\mathbf{G}_0 = \mathbf{W}_n \mathbf{S}_0^{-1}$, and under the assumption that $\lim_{n \rightarrow \infty} \mathbf{D}_n$ exists and has the full rank $k + 1$.

The GMME uses as linear instruments

$$\mathbf{H}_n = [\mathbf{X}_n, \mathbf{W}_n \mathbf{X}_n, \mathbf{W}_n^2 \mathbf{X}_n, \dots],$$

and

$$\begin{aligned} \mathbf{P}_{1n} &= \mathbf{W}_n, \\ \mathbf{P}_{2n} &= \mathbf{W}_n^2 - (\text{tr}(\mathbf{W}_n^2)/n) \mathbf{I}_n. \end{aligned}$$

Assuming that $\mathbf{Y}_n = \mathbf{I}_{(2+k_k)}$, the asymptotic variance can be estimated as

$$\hat{\boldsymbol{\Sigma}} = \left[\left(\frac{1}{n} \hat{\mathbf{D}}_n^\top \right) \left(\frac{1}{n} \hat{\mathbf{D}}_n \right) \right]^{-1} \left(\frac{1}{n} \hat{\mathbf{D}}_n^\top \right) \left(\frac{1}{n} \hat{\boldsymbol{\Omega}}_n \right) \left(\frac{1}{n} \hat{\mathbf{D}}_n \right) \left[\left(\frac{1}{n} \hat{\mathbf{D}}_n^\top \right) \left(\frac{1}{n} \hat{\mathbf{D}}_n \right) \right]^{-1},$$

where $\hat{\boldsymbol{\Omega}}_n$ is estimated as

$$\hat{\boldsymbol{\Omega}}_n = \begin{pmatrix} (\hat{\mu}_4 - 3\hat{\sigma}_n^4)\boldsymbol{\omega}_n^\top \boldsymbol{\omega}_n & \hat{\mu}_3\boldsymbol{\omega}_n^\top \mathbf{H}_n \\ \hat{\mu}_3\mathbf{H}_n^\top \boldsymbol{\omega}_n & \mathbf{O} \end{pmatrix} + \hat{\mathbf{V}}_n, \quad (6.43)$$

with $\hat{\mu}_4 = (1/n) \sum_{i=1}^n \hat{\epsilon}_{in}^4$, $\hat{\mu}_3 = (1/n) \sum_{i=1}^n \hat{\epsilon}_{in}^3$, $\hat{\sigma}_n^2 = (1/n) \sum_{i=1}^n \hat{\epsilon}_{in}^2$, $\hat{\boldsymbol{\epsilon}}_n = (\mathbf{I}_n - \hat{\rho}_n \mathbf{W}_n) \mathbf{y}_n - \mathbf{X}_n \hat{\boldsymbol{\beta}}_n$, and

$$\hat{\mathbf{V}}_n = \hat{\sigma}_n^4 \begin{pmatrix} \text{tr}(\mathbf{P}_{1n} \mathbf{P}_{1n}^s) & \text{tr}(\mathbf{P}_{1n} \mathbf{P}_{2n}^s) & \mathbf{0} \\ \text{tr}(\mathbf{P}_{2n} \mathbf{P}_{1n}^s) & \text{tr}(\mathbf{P}_{2n} \mathbf{P}_{2n}^s) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{1}{\hat{\sigma}_n^2} \mathbf{H}_n^\top \mathbf{H}_n \end{pmatrix}$$

The matrix $\hat{\mathbf{D}}_n$ can be estimated by replacing $\hat{\boldsymbol{\theta}}_n$ in Equation (6.42) or using the Jacobian in Equation (6.41) evaluated at $\hat{\boldsymbol{\theta}}_n$.

Consistency of GMM estimator

In the following lines, we provide and sketch for the proof of the asymptotic properties of the GMM estimator for SLM. First, we need the following results. Note that for any ρ in the parameter space (see Exercise 4.6)

$$\begin{aligned} \mathbf{S}_n^{-1} &= \mathbf{I}_n + \rho \mathbf{W}_n + \rho^2 \mathbf{W}_n^2 + \rho^3 \mathbf{W}_n + \cdots, \\ &= \mathbf{I}_n + \rho \mathbf{W}_n [\mathbf{I}_n + \rho \mathbf{W}_n + \rho^2 \mathbf{W}_n^2 + \cdots], \\ &= \mathbf{I}_n + \rho \mathbf{W}_n \mathbf{S}_n^{-1}. \end{aligned} \quad (6.44)$$

In addition, and using the result in Equation (6.44), the following expansion would be useful:

$$\begin{aligned} \mathbf{S}_n \mathbf{S}_0^{-1} &= (\mathbf{I}_n - \rho \mathbf{W}_n) (\mathbf{I}_n + \rho \mathbf{W}_n \mathbf{S}_0^{-1}), \\ &= \mathbf{I}_n + \rho_0 \mathbf{W}_n \mathbf{S}_0^{-1} - \rho \mathbf{W}_n - \rho \mathbf{W}_n \rho_0 \mathbf{W}_n \mathbf{S}_0^{-1}, \\ &= \mathbf{I}_n + \rho_0 \mathbf{W}_n \mathbf{S}_0^{-1} - (\rho \mathbf{W}_n [\mathbf{I}_n + \rho_0 \mathbf{W}_n \mathbf{S}_0^{-1}]) \\ &= \mathbf{I}_n + \rho_0 \mathbf{W}_n \mathbf{S}_0^{-1} - \rho \mathbf{W}_n \mathbf{S}_0^{-1} \\ &= \mathbf{I}_n + (\rho_0 - \rho) \mathbf{W}_n \mathbf{S}_0^{-1} \\ &= \mathbf{I}_n + (\rho_0 - \rho) \mathbf{G}_0, \end{aligned} \quad (6.45)$$

where $\mathbf{G}_0 = \mathbf{W}_n \mathbf{S}_0^{-1}$.

Consistency of GMM estimator relies on the conditions of consistency for extremum estimators (Hansen, 1982; Newey and McFadden, 1994). Two important conditions are **identification** conditions and **uniform convergence** (Lee, 2003, 2007).

To understand the conditions for identification, consider the moment functions given in Equation (6.39). To analyze the asymptotic behavior of these moment functions, we need to write the error term as a function of the true parameters of the models. Using the definition of the error term for the SLM and the reduced form equation, the error term can be expressed as:

$$\begin{aligned} \boldsymbol{\epsilon}_n(\boldsymbol{\theta}) &= (\mathbf{I}_n - \rho \mathbf{W}_n) \mathbf{y}_n - \mathbf{X}_n \boldsymbol{\beta}, \\ &= (\mathbf{I}_n - \rho \mathbf{W}_n) (\mathbf{S}_0^{-1} \mathbf{X}_n \boldsymbol{\beta}_0 + \mathbf{S}_0^{-1} \boldsymbol{\epsilon}_n) - \mathbf{X}_n \boldsymbol{\beta}, \\ &= \mathbf{S}_n \mathbf{S}_0^{-1} \mathbf{X}_n \boldsymbol{\beta}_0 - \mathbf{X}_n \boldsymbol{\beta} + \mathbf{S}_n \mathbf{S}_0^{-1} \boldsymbol{\epsilon}_n, \\ &= \mathbf{d}_n(\boldsymbol{\theta}) + \mathbf{S}_n \mathbf{S}_0^{-1} \boldsymbol{\epsilon}_n, \end{aligned} \quad (6.46)$$

where $\mathbf{S}_n = (\mathbf{I}_n - \rho \mathbf{W}_n)$, $\mathbf{S}_0 = (\mathbf{I}_n - \rho_0 \mathbf{W}_n)$, and $\mathbf{d}_n(\boldsymbol{\theta}) = \mathbf{S}_n \mathbf{S}_0^{-1} \mathbf{X}_n \boldsymbol{\beta}_0 - \mathbf{X}_n \boldsymbol{\beta}$. Then, using the result in Equation (6.45), $\mathbf{d}_n(\boldsymbol{\theta})$ in Equation (6.46) can be written as

$$\begin{aligned} \mathbf{d}_n(\boldsymbol{\theta}) &= \mathbf{S}_n \mathbf{S}_0^{-1} \mathbf{X}_n \boldsymbol{\beta}_0 - \mathbf{X}_n \boldsymbol{\beta}, \\ &= [\mathbf{I}_n + (\rho_0 - \rho) \mathbf{G}_0] \mathbf{X}_n \boldsymbol{\beta}_0 - \mathbf{X}_n \boldsymbol{\beta}, \\ &= \mathbf{X}_n (\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + (\rho_0 - \rho) \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0. \end{aligned}$$

Using these expression, we can write the population moments using the linear instruments as

$$\mathbb{E}(\mathbf{H}_n^\top \boldsymbol{\varepsilon}_n) = \mathbb{E}[\mathbf{H}_n^\top (\mathbf{d}_n(\boldsymbol{\theta}) + \mathbf{S}_n \mathbf{S}_0^{-1} \boldsymbol{\varepsilon}_n)] = \mathbf{H}_n^\top \mathbf{d}_n(\boldsymbol{\theta}),$$

whereas the population moments using the quadratic instruments can be written as

$$\begin{aligned} \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{P}_{jn} \boldsymbol{\varepsilon}_n) &= \mathbb{E}\left[(\mathbf{d}_n(\boldsymbol{\theta}) + \mathbf{S}_n \mathbf{S}_0^{-1} \boldsymbol{\varepsilon}_n)^\top \mathbf{P}_{jn} (\mathbf{d}_n(\boldsymbol{\theta}) + \mathbf{S}_n \mathbf{S}_0^{-1} \boldsymbol{\varepsilon}_n)\right], \\ &= \mathbf{d}_n(\boldsymbol{\theta})^\top \mathbf{P}_{jn} \mathbf{d}_n(\boldsymbol{\theta}) + \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{S}_0^{-1\top} \mathbf{S}_n^\top \mathbf{P}_{jn} \mathbf{S}_n \mathbf{S}_0^{-1} \boldsymbol{\varepsilon}_n), \\ &= \mathbf{d}_n(\boldsymbol{\theta})^\top \mathbf{P}_{jn} \mathbf{d}_n(\boldsymbol{\theta}) + \sigma_0^2 \text{tr}(\mathbf{S}_0^{-1\top} \mathbf{S}_n^\top \mathbf{P}_{jn} \mathbf{S}_n \mathbf{S}_0^{-1}), \end{aligned}$$

for $j = 1, 2$. Thus, for any possible value $\boldsymbol{\theta}$, the population moment conditions can be written as

$$\mathbb{E}[\mathbf{g}_n(\boldsymbol{\theta})] = \begin{pmatrix} \mathbf{d}_n(\boldsymbol{\theta})^\top \mathbf{P}_{1n} \mathbf{d}_n(\boldsymbol{\theta}) + \sigma_0^2 \text{tr}(\mathbf{S}_0^{-1\top} \mathbf{S}_n^\top \mathbf{P}_{1n} \mathbf{S}_n \mathbf{S}_0^{-1}) \\ \mathbf{d}_n(\boldsymbol{\theta})^\top \mathbf{P}_{2n} \mathbf{d}_n(\boldsymbol{\theta}) + \sigma_0^2 \text{tr}(\mathbf{S}_0^{-1\top} \mathbf{S}_n^\top \mathbf{P}_{2n} \mathbf{S}_n \mathbf{S}_0^{-1}) \\ \mathbf{H}_n^\top \mathbf{d}_n(\boldsymbol{\theta}) \end{pmatrix}.$$

Identification condition requires $\lim_{n \rightarrow \infty} (1/n) \mathbb{E}[\mathbf{g}_n(\boldsymbol{\theta})] = \mathbf{0}$ at $\boldsymbol{\theta}_0$. Consider the moment equations using \mathbf{H}_n . In the limit, the moment equations are:

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{H}_n^\top \mathbf{d}_n(\boldsymbol{\theta}) &= \lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{H}_n^\top [\mathbf{X}_n (\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + (\rho_0 - \rho) \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0], \\ &= \lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{H}_n^\top \mathbf{X}_n (\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + \lim_{n \rightarrow \infty} \frac{1}{n} (\rho_0 - \rho) \mathbf{H}_n^\top \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0, \end{aligned}$$

which will be $\lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{H}_n^\top \mathbf{d}_n = \mathbf{0}$ at the true parameters $\boldsymbol{\theta}_0$, if $[\mathbf{H}_n^\top \mathbf{X}_n, \mathbf{H}_n^\top \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0]$ has a full column rank, that is, its ranks equals $k + 1$, for large enough n . Intuitively, the sufficient conditions requires that \mathbf{H}_n to be correlated with the endogenous variables $\mathbf{W}_n \mathbf{y}$. On the other hand, **sufficient rank condition** implies the **necessary rank condition** that $[\mathbf{X}_n, \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0]$ has a full column rank $(k + 1)$, and that \mathbf{H}_n has a rank **at least** $k + 1$. This will occur if $\mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0$ and \mathbf{X}_n are not asymptotically linearly dependent.

For the quadratic moments, note that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{d}_n(\boldsymbol{\theta})^\top \mathbf{P}_{jn} \mathbf{d}_n(\boldsymbol{\theta}) + \lim_{n \rightarrow \infty} \sigma_0^2 \frac{1}{n} \text{tr}(\mathbf{S}_0^{-1\top} \mathbf{S}_n^\top \mathbf{P}_{jn} \mathbf{S}_n \mathbf{S}_0^{-1}).$$

The identification of ρ_0 requires that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \text{tr}(\mathbf{S}_0^{-1\top} \mathbf{S}_n^\top \mathbf{P}_{jn} \mathbf{S}_n \mathbf{S}_0^{-1}) = 0,$$

for $j = 1, 2$. Lee (2007) states that the set of limiting quadratic moment equations has a unique solution at ρ_0 if

$$\lim_{n \rightarrow \infty} (1/n) [\text{tr}(\mathbf{P}_{1n}^s \mathbf{G}_0), \text{tr}(\mathbf{P}_{2n}^s \mathbf{G}_0)]^\top,$$

is linearly independent of

$$\lim_{n \rightarrow \infty} (1/n) [\text{tr}(\mathbf{G}_0^\top \mathbf{P}_{1n} \mathbf{G}_0), \text{tr}(\mathbf{G}_0^\top \mathbf{P}_{2n} \mathbf{G}_0)]^\top$$

Next, we need to show that:

$$\frac{1}{n} \mathbf{g}_n(\boldsymbol{\theta}) \xrightarrow{p} \mathbb{E}[\mathbf{g}_n(\boldsymbol{\theta})],$$

uniformly in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. In previous proofs, we have shown that for the linear instruments $\frac{1}{n} \mathbf{H}_n^\top \boldsymbol{\varepsilon}_n \xrightarrow{p} \frac{1}{n} \mathbf{H}_n^\top \mathbf{d}_n(\boldsymbol{\theta})$.

The analysis for the quadratic moments is a bit more complicated. Using the error term as function of the population parameters, we can decompose the quadratic moment as

$$\begin{aligned} \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{jn} \boldsymbol{\varepsilon}_n &= \mathbf{d}_n(\boldsymbol{\theta})^\top \mathbf{P}_{jn} \mathbf{d}_n(\boldsymbol{\theta}) + \mathbf{d}_n(\boldsymbol{\theta})^\top \mathbf{P}_{jn} \mathbf{S}_n \mathbf{S}_0^{-1} \boldsymbol{\varepsilon}_n + \boldsymbol{\varepsilon}_n^\top \mathbf{S}_0^{-1\top} \mathbf{S}_n^\top \mathbf{d}_n(\boldsymbol{\theta}) + \boldsymbol{\varepsilon}_n^\top \mathbf{S}_0^{-1\top} \mathbf{S}_n^\top \mathbf{P}_{jn} \mathbf{S}_n \mathbf{S}_0^{-1} \boldsymbol{\varepsilon}_n \\ &= \mathbf{d}_n(\boldsymbol{\theta})^\top \mathbf{P}_{jn} \mathbf{d}_n(\boldsymbol{\theta}) + \mathbf{d}_n(\boldsymbol{\theta})^\top \mathbf{P}_{jn} (\mathbf{I}_n + (\rho_0 - \rho) \mathbf{G}_0) \boldsymbol{\varepsilon}_n \\ &\quad + \boldsymbol{\varepsilon}_n^\top (\mathbf{I}_n + (\rho_0 - \rho) \mathbf{G}_0)^\top \mathbf{d}_n(\boldsymbol{\theta}) \\ &\quad + \boldsymbol{\varepsilon}_n^\top (\mathbf{I}_n + (\rho_0 - \rho) \mathbf{G}_0)^\top \mathbf{P}_{jn} (\mathbf{I}_n + (\rho_0 - \rho) \mathbf{G}_0) \boldsymbol{\varepsilon}_n \\ &= \mathbf{d}_n(\boldsymbol{\theta})^\top \mathbf{P}_{jn} \mathbf{d}_n(\boldsymbol{\theta}) + \underbrace{\mathbf{d}_n(\boldsymbol{\theta})^\top \mathbf{P}_{jn}^s (\boldsymbol{\varepsilon}_n + (\rho_0 - \rho) \mathbf{G}_0 \boldsymbol{\varepsilon}_n)}_{\mathbf{l}_n(\boldsymbol{\theta})} \\ &\quad + \underbrace{(\boldsymbol{\varepsilon}_n^\top + (\rho_0 - \rho) \boldsymbol{\varepsilon}_n^\top \mathbf{G}_0^\top) \mathbf{P}_{jn} (\boldsymbol{\varepsilon}_n + (\rho_0 - \rho) \mathbf{G}_0 \boldsymbol{\varepsilon}_n)}_{\mathbf{q}_n(\boldsymbol{\theta})} \end{aligned}$$

Focusing on $(1/n) \mathbf{l}_n(\boldsymbol{\theta})$, and using $\mathbf{d}_n(\boldsymbol{\theta})$ yields.

$$\begin{aligned} \frac{1}{n} \mathbf{l}_n(\boldsymbol{\theta}) &= (\rho_0 - \rho) \frac{1}{n} (\mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{G}_0^\top \mathbf{P}_{jn}^s \boldsymbol{\varepsilon}_n + (\boldsymbol{\beta}_0 - \boldsymbol{\beta})^\top \frac{1}{n} \mathbf{X}_n^\top \mathbf{P}_{jn}^s \boldsymbol{\varepsilon}_n \\ &\quad + (\rho_0 - \rho)^2 \frac{1}{n} (\mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{G}_0^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \boldsymbol{\varepsilon}_n \\ &\quad + (\rho_0 - \rho) (\boldsymbol{\beta}_0 - \boldsymbol{\beta})^\top \frac{1}{n} \mathbf{X}_n^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \boldsymbol{\varepsilon}_n. \end{aligned}$$

Thus, because the previous expression is linear in $\boldsymbol{\varepsilon}_n$ it can be established that

$$\begin{aligned} \frac{1}{n} \mathbf{l}_n(\boldsymbol{\theta}) &\xrightarrow{p} (\rho_0 - \rho) \frac{1}{n} (\mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{G}_0^\top \mathbf{P}_{jn}^s \mathbb{E}(\boldsymbol{\varepsilon}_n) + (\boldsymbol{\beta}_0 - \boldsymbol{\beta})^\top \frac{1}{n} \mathbf{X}_n^\top \mathbf{P}_{jn}^s \mathbb{E}(\boldsymbol{\varepsilon}_n) \\ &\quad + (\rho_0 - \rho)^2 \frac{1}{n} (\mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{G}_0^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \mathbb{E}(\boldsymbol{\varepsilon}_n) \\ &\quad + (\rho_0 - \rho) (\boldsymbol{\beta}_0 - \boldsymbol{\beta})^\top \frac{1}{n} \mathbf{X}_n^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \mathbb{E}(\boldsymbol{\varepsilon}_n) \\ &\xrightarrow{p} \mathbf{0}, \end{aligned}$$

uniformly in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. The uniform convergence in probability follows because $\frac{1}{n} \mathbf{l}_n(\boldsymbol{\theta})$ is simply a quadratic function of ρ and $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}$ is a bounded set.

For the second term, operating over the matrix multiplications, and taking plim yields,

$$\begin{aligned} \frac{1}{n} \mathbf{q}_n(\boldsymbol{\theta}) &\xrightarrow{p} \frac{1}{n} \mathbb{E}[\boldsymbol{\varepsilon}_n^\top \mathbf{P}_{jn} \boldsymbol{\varepsilon}_n] + (\rho_0 - \rho) \frac{1}{n} \mathbb{E}[\boldsymbol{\varepsilon}_n^\top \mathbf{G}_0^\top \mathbf{P}_{jn}^s \boldsymbol{\varepsilon}_n] + (\rho_0 - \rho)^2 \frac{1}{n} \mathbb{E}[\boldsymbol{\varepsilon}_n^\top \mathbf{G}_0^\top \mathbf{P}_{jn} \mathbf{G}_0 \boldsymbol{\varepsilon}_n], \\ &\xrightarrow{p} (\rho_0 - \rho) \frac{\sigma_0^2}{n} \text{tr}(\mathbf{G}_0^\top \mathbf{P}_{js}^s) + (\rho_0 - \rho)^2 \frac{\sigma_0^2}{n} \text{tr}(\mathbf{G}_0^\top \mathbf{P}_{js}^s \mathbf{G}_0), \end{aligned}$$

which can be also be written as

$$\frac{1}{n} \mathbf{q}_n(\boldsymbol{\theta}) = (\rho_0 - \rho) \frac{\sigma_0^2}{n} \text{tr}(\mathbf{G}_0^\top \mathbf{P}_{js}^s) + (\rho_0 - \rho)^2 \frac{\sigma_0^2}{n} \text{tr}(\mathbf{G}_0^\top \mathbf{P}_{js}^s \mathbf{G}_0) + o_p(1),$$

Collecting terms yields,

$$\frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{jn} \boldsymbol{\varepsilon}_n \xrightarrow{p} \frac{1}{n} \mathbf{d}_n(\boldsymbol{\theta})^\top \mathbf{P}_{jn} \mathbf{d}_n(\boldsymbol{\theta}) + (\rho_0 - \rho) \frac{\sigma_0^2}{n} \text{tr}(\mathbf{G}_0^\top \mathbf{P}_{js}^s) + (\rho_0 - \rho)^2 \frac{\sigma_0^2}{n} \text{tr}(\mathbf{G}_0^\top \mathbf{P}_{js}^s \mathbf{G}_0),$$

uniformly in $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}$.

Thus,

$$\begin{aligned} \begin{pmatrix} \frac{1}{n} \boldsymbol{\varepsilon}_n \mathbf{P}_{1n} \boldsymbol{\varepsilon}_n \\ \frac{1}{n} \boldsymbol{\varepsilon}_n \mathbf{P}_{2n} \boldsymbol{\varepsilon}_n \\ \frac{1}{n} \mathbf{H}_n^\top \boldsymbol{\varepsilon}_n \end{pmatrix} &\xrightarrow{p} \frac{1}{n} \mathbb{E}[\mathbf{g}_n(\boldsymbol{\theta})] \\ &= \begin{pmatrix} \frac{1}{n} \mathbf{d}_n(\boldsymbol{\theta})^\top \mathbf{P}_{1n} \mathbf{d}_n(\boldsymbol{\theta}) + (\rho_0 - \rho) \frac{\sigma_0^2}{n} \text{tr}(\mathbf{G}_0^\top \mathbf{P}_{1s}^s) + (\rho_0 - \rho)^2 \frac{\sigma_0^2}{n} \text{tr}(\mathbf{G}_0^\top \mathbf{P}_{1s}^s \mathbf{G}_0) \\ \frac{1}{n} \mathbf{d}_n(\boldsymbol{\theta})^\top \mathbf{P}_{2n} \mathbf{d}_n(\boldsymbol{\theta}) + (\rho_0 - \rho) \frac{\sigma_0^2}{n} \text{tr}(\mathbf{G}_0^\top \mathbf{P}_{2s}^s) + (\rho_0 - \rho)^2 \frac{\sigma_0^2}{n} \text{tr}(\mathbf{G}_0^\top \mathbf{P}_{2s}^s \mathbf{G}_0) \\ \mathbf{H}_n^\top \mathbf{d}_n(\boldsymbol{\theta}) \end{pmatrix}. \end{aligned}$$

As $\mathbf{g}_n(\boldsymbol{\theta})$ is a quadratic function of $\boldsymbol{\theta}$ and $\boldsymbol{\Theta}$ is bounded, $(1/n)\mathbb{E}[\mathbf{g}_n(\boldsymbol{\theta})]$ is uniformly equicontinuous on $\boldsymbol{\Theta}$. The identification condition and the uniform equicontinuity of $(1/n)\mathbb{E}(\mathbf{g}_n(\boldsymbol{\theta}))$ imply that the identification uniqueness condition must be satisfied.

Asymptotic distribution of GMM estimator

Assuming that $\hat{\boldsymbol{\theta}}_n$ is an interior point, the first-order conditions yields

$$\frac{\partial \mathbf{g}_n(\hat{\boldsymbol{\theta}}_n)^\top \boldsymbol{\Upsilon}_n \mathbf{g}_n(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}} = - \left(\frac{\partial \mathbf{g}_n(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}^\top} \right)^\top \boldsymbol{\Upsilon}_n \mathbf{g}_n(\hat{\boldsymbol{\theta}}_n) = \mathbf{0}. \quad (6.47)$$

By Taylor expansion around $\boldsymbol{\theta}_0$ of $\mathbf{g}_n(\hat{\boldsymbol{\theta}}_n)$:

$$\mathbf{g}_n(\hat{\boldsymbol{\theta}}_n) = \mathbf{g}_n(\boldsymbol{\theta}_0) + \frac{\partial \mathbf{g}_n(\bar{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}^\top} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0). \quad (6.48)$$

Substituting (6.48) into (6.47), solving for $(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ and multiplying by \sqrt{n} gives

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = - \left[\left(\frac{1}{n} \frac{\partial \mathbf{g}_n(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}^\top} \right)^\top \boldsymbol{\Upsilon}_n \left(\frac{1}{n} \frac{\partial \mathbf{g}_n(\bar{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}^\top} \right) \right]^{-1} \left(\frac{1}{n} \frac{\partial \mathbf{g}_n(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}^\top} \right)^\top \boldsymbol{\Upsilon}_n \sqrt{n} \frac{1}{n} \mathbf{g}_n(\boldsymbol{\theta}_0).$$

where $\bar{\boldsymbol{\theta}}_n$ is some between value.

Consider the gradient

$$\begin{aligned} \frac{\partial \mathbf{g}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} &= \begin{pmatrix} \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{1n}^s \\ \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{2n}^s \\ \mathbf{H}_n^\top \end{pmatrix} \begin{pmatrix} -\mathbf{W}_n \mathbf{y}_n & -\mathbf{X}_n \end{pmatrix} \\ &= -(\mathbf{P}_{1n}^{s\top} \boldsymbol{\varepsilon}_n, \mathbf{P}_{2n}^{s\top} \boldsymbol{\varepsilon}_n, \mathbf{H}_n^\top)^\top (\mathbf{W}_n \mathbf{y}_n, \mathbf{X}_n) \\ &= - \begin{pmatrix} \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{1n}^s \mathbf{W}_n \mathbf{y}_n & \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{1n}^s \mathbf{X}_n \\ \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{2n}^s \mathbf{W}_n \mathbf{y}_n & \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{2n}^s \mathbf{X}_n \\ \mathbf{H}_n^\top \mathbf{W}_n \mathbf{y}_n & \mathbf{H}_n^\top \mathbf{X}_n \end{pmatrix}. \end{aligned}$$

We will show that $\frac{\partial g_n(\theta)}{\partial \theta^\top} = \frac{1}{n} \mathbf{D}_n + o_p(1)$, where $\mathbf{D}_n = \frac{\partial \mathbb{E}[g_n(\theta_0)]}{\partial \theta^\top}$. Note that for $j = 1, 2$,

$$\begin{aligned} \frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{jn} \mathbf{W}_n \mathbf{y}_n &= \frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{jn} \mathbf{W}_n (\mathbf{S}_0^{-1} \mathbf{X}_n \boldsymbol{\beta}_0 + \mathbf{S}_0^{-1} \boldsymbol{\varepsilon}_n) \\ &= \frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0 + \frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \boldsymbol{\varepsilon}_n. \end{aligned}$$

For the first element, and using Equation (6.46):

$$\begin{aligned} \frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0 &= \frac{1}{n} (\mathbf{d}_n(\theta) + \mathbf{S}_n \mathbf{S}_0^{-1} \boldsymbol{\varepsilon}_n)^\top (\mathbf{P}_{jn}^s \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0), \\ &= \mathbf{d}_n(\theta)^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0 + \frac{1}{n} \boldsymbol{\varepsilon}_n^\top (\mathbf{S}_n \mathbf{S}_0^{-1})^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0, \\ &= \mathbf{d}_n(\theta)^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0 + \frac{1}{n} \boldsymbol{\varepsilon}_n^\top [\mathbf{I}_n + (\rho_0 - \rho) \mathbf{G}_0]^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0, \\ &= \mathbf{d}_n(\theta)^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0 + \frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0 + \frac{1}{n} (\rho_0 - \rho) \boldsymbol{\varepsilon}_n^\top \mathbf{G}_0^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0. \end{aligned}$$

Then,

$$\begin{aligned} \frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0 &\xrightarrow{p} \frac{1}{n} \mathbf{d}_n(\theta)^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0 + \frac{1}{n} \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0) + \frac{1}{n} (\rho_0 - \rho) \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{G}_0^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0), \\ &\xrightarrow{p} \frac{1}{n} \mathbf{d}_n(\theta)^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0, \end{aligned}$$

uniformly in $\theta \in \boldsymbol{\Theta}$, since $\mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0) = o_p(1)$ and $\mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{G}_0^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0) = o_p(1)$. For the second term, note that

$$\begin{aligned} \frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \boldsymbol{\varepsilon}_n &= \frac{1}{n} \mathbf{d}_n(\theta)^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \boldsymbol{\varepsilon}_n + \frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \boldsymbol{\varepsilon}_n + (\rho_0 - \rho) \frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{G}_0^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \boldsymbol{\varepsilon}_n, \\ &\xrightarrow{p} \frac{1}{n} \mathbb{E}(\mathbf{d}_n(\theta)^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \boldsymbol{\varepsilon}_n) + \frac{1}{n} \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \boldsymbol{\varepsilon}_n) + (\rho_0 - \rho) \frac{1}{n} \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{G}_0^\top \mathbf{P}_{jn}^s \mathbf{G}_0 \boldsymbol{\varepsilon}_n), \\ &= \frac{\sigma_0^2}{n} \text{tr}(\mathbf{P}_{jn}^s \mathbf{G}_0) + (\rho_0 - \rho) \frac{\sigma_0^2}{n} \text{tr}(\mathbf{G}_0^\top \mathbf{P}_{jn}^s \mathbf{G}_0). \end{aligned}$$

Continuing with the other elements of the gradient:

$$\begin{aligned} \frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{jn}^s \mathbf{X}_n &= \frac{1}{n} [\mathbf{d}_n(\theta) + \mathbf{S}_n \mathbf{S}_0^{-1} \boldsymbol{\varepsilon}_n]^\top \mathbf{P}_{jn}^s \mathbf{X}_n, \\ &= \frac{1}{n} \mathbf{d}_n(\theta)^\top \mathbf{P}_{jn}^s \mathbf{X}_n + \frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{jn}^s \mathbf{X}_n + \frac{1}{n} (\rho_0 - \rho) \boldsymbol{\varepsilon}_n^\top \mathbf{G}_0^\top \mathbf{P}_{jn}^s \mathbf{X}_n, \\ &\xrightarrow{p} \frac{1}{n} \mathbf{d}_n(\theta)^\top \mathbf{P}_{jn}^s \mathbf{X}_n. \end{aligned}$$

Similarly:

$$\begin{aligned} \frac{1}{n} \mathbf{H}_n^\top \mathbf{W}_n \mathbf{y}_n &= \frac{1}{n} \mathbf{H}_n^\top \mathbf{W}_n (\mathbf{S}_0^{-1} \mathbf{X}_n \boldsymbol{\beta}_0 + \mathbf{S}_0^{-1} \boldsymbol{\varepsilon}_n), \\ &= \frac{1}{n} \mathbf{H}_n^\top \mathbf{W}_n \mathbf{S}_0^{-1} \mathbf{X}_n \boldsymbol{\beta}_0 + \frac{1}{n} \mathbf{H}_n^\top \mathbf{W}_n \mathbf{S}_0^{-1} \boldsymbol{\varepsilon}_n, \\ &\xrightarrow{p} \frac{1}{n} \mathbf{H}_n^\top \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0. \end{aligned}$$

Thus,

$$\frac{1}{n} \frac{\partial \mathbf{g}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} \xrightarrow{p} \begin{pmatrix} \frac{1}{n} \mathbf{d}_n(\boldsymbol{\theta})^\top \mathbf{P}_{1s}^s \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0 + \frac{\sigma_0^2}{n} \text{tr}(\mathbf{P}_{1n}^s \mathbf{G}_0) + (\rho_0 - \rho) \frac{\sigma_0^2}{n} \text{tr}(\mathbf{G}_0^\top \mathbf{P}_{1n}^s \mathbf{G}_0) & \frac{1}{n} \mathbf{d}_n(\boldsymbol{\theta})^\top \mathbf{P}_{1s}^s \mathbf{X}_n \\ \frac{1}{n} \mathbf{d}_n(\boldsymbol{\theta})^\top \mathbf{P}_{2s}^s \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0 + \frac{\sigma_0^2}{n} \text{tr}(\mathbf{P}_{2n}^s \mathbf{G}_0) + (\rho_0 - \rho) \frac{\sigma_0^2}{n} \text{tr}(\mathbf{G}_0^\top \mathbf{P}_{2n}^s \mathbf{G}_0) & \frac{1}{n} \mathbf{d}_n(\boldsymbol{\theta})^\top \mathbf{P}_{2s}^s \mathbf{X}_n \\ \frac{1}{n} \mathbf{H}_n^\top \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0 & \frac{1}{n} \mathbf{H}_n^\top \mathbf{X}_n \end{pmatrix}$$

At $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, $\mathbf{d}_n(\boldsymbol{\theta}) = \mathbf{0}$, Then

$$\begin{aligned} \frac{1}{n} \frac{\partial \mathbf{g}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} &\xrightarrow{p} -\frac{1}{n} \mathbf{D}_n, \\ &\xrightarrow{p} \frac{1}{n} \begin{pmatrix} \sigma_0^2 \text{tr}(\mathbf{P}_{1n}^s \mathbf{G}_0) & \mathbf{0} \\ \sigma_0^2 \text{tr}(\mathbf{P}_{2n}^s \mathbf{G}_0) & \mathbf{0} \\ \mathbf{H}_n^\top \mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0 & \mathbf{H}_n^\top \mathbf{X}_n \end{pmatrix}. \end{aligned}$$

It follows that

$$\left(\frac{1}{n} \frac{\partial \mathbf{g}_n(\bar{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}^\top} \right) = -\frac{1}{n} \mathbf{D}_n + o_p(1),$$

where $\frac{\partial \mathbf{g}_n(\bar{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}^\top} = O_p(1)$ and $\frac{1}{n} \mathbf{D}_n = O(1)$.

In addition, $\boldsymbol{\Upsilon}_n - \boldsymbol{\Upsilon}_0 = o_p(1)$. In addition, by central limit Theorem for linear-quadratic function 3.31

$$\begin{aligned} \frac{1}{\sqrt{n}} \mathbf{g}_n(\boldsymbol{\theta}_0) &= \frac{1}{\sqrt{n}} \begin{pmatrix} \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}_0) \mathbf{P}_{1n} \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}_0) \\ \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}_0) \mathbf{P}_{2n} \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}_0) \\ \mathbf{H}_n^\top \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}_0) \end{pmatrix} \\ &= \mathbf{N} \left(\mathbf{0}, \lim_{n \rightarrow \infty} \frac{1}{n} \boldsymbol{\Omega}_n \right), \end{aligned}$$

and

$$\boldsymbol{\zeta}_n = -\boldsymbol{\Omega}_n^{-1/2} \frac{1}{\sqrt{n}} \mathbf{g}_n(\boldsymbol{\theta}_0) \xrightarrow{d} \mathbf{N}(\mathbf{0}, \mathbf{I}_{2+k_x})$$

Thus:

$$\sqrt{n} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \left[\left(\frac{1}{n} \mathbf{D}_n^\top \right) \boldsymbol{\Upsilon}_0 \left(\frac{1}{n} \mathbf{D}_n \right) \right]^{-1} \left(\frac{1}{n} \mathbf{D}_n^\top \right) \boldsymbol{\Upsilon}_0 \boldsymbol{\zeta}_n + o_p(1).$$

6.3.2 OGMM Estimator

Different weighting matrices $\boldsymbol{\Upsilon}_n$ lead to different consistent estimators, each with potentially different asymptotic covariance structures. As established in Theorem 6.14, the GMME remains consistent for any sequence $\boldsymbol{\Upsilon}_n$, satisfying $\boldsymbol{\Upsilon}_n \rightarrow \boldsymbol{\Upsilon}_0$. However, among the class of GMM estimators indexed by $\boldsymbol{\Upsilon}_n$, our goal is to select the one that minimizes the asymptotic variance given the moment conditions $\mathbf{g}_n(\boldsymbol{\theta})$.

A particularly important case arises when the weighting matrix is chosen as $\boldsymbol{\Upsilon}_n = \left(\frac{1}{n} \hat{\boldsymbol{\Omega}}_n \right)^{-1}$. Under this specification, the GMME attains its optimal asymptotic efficiency and is referred to as the Optimal GMM Estimator (OGMME). The following Theorem provides the asymptotic properties of the OGMME for the SLM under homoskedasticity (Lee, 2007).

Theorem 6.15 — **OGMM estimator for SLM under homoskedasticity (Lee, 2007)**. In addition to Assumptions 6.13, suppose that

- (a) the limit of $(1/n)\mathbf{\Omega}_n$ exists and is a nonsingular matrix,
- (b) $\left(\widehat{\mathbf{\Omega}}_n/n\right)^{-1} - (\mathbf{\Omega}_n/n)^{-1} = o_p(1)$,

then the OGMM, $\widehat{\boldsymbol{\theta}}_n$ derived from $\min_{\boldsymbol{\theta} \in \Theta} \mathbf{g}_n^\top(\boldsymbol{\theta}) \widehat{\mathbf{\Omega}}_n^{-1} \mathbf{g}_n(\boldsymbol{\theta})$ with \mathbf{P}'_{jn} s from \mathcal{P}_{1n} has the asymptotic distribution

$$\sqrt{n} \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \right) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}^o)$$

where

$$\boldsymbol{\Sigma}^o = \left[\lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{D}_n^\top \mathbf{\Omega}_n \mathbf{D}_n \right]^{-1}.$$

Furthermore:

$$\mathbf{g}_n^\top(\widehat{\boldsymbol{\theta}}) \widehat{\mathbf{\Omega}}_n^{-1} \mathbf{g}_n(\widehat{\boldsymbol{\theta}}) \xrightarrow{d} \chi_{(2+k_x)-(k+1)}^2$$

6.3.3 Coding GMME and OGMME for SLM

First, we create a function that returns the moment conditions presented in Equation (6.39), and Jacobian matrix in Equation (6.41) (both divided by n). This function is next used to construct the objective function to be minimized.

```
# Trace function
tr <- function(A) return(sum(diag(A)))

# Create moments as in Lee 2007
moments.lee2007 <- function(theta, y, X, H, W){
  # y: vector of dependent variables
  # X: matrix of exogenous variables
  # H: matrix of linear instrument
  # W: Spatial weight matrix
  k      <- ncol(X)
  n      <- nrow(X)
  rho    <- theta[1L]
  beta   <- theta[(2L):(k + 1)]
  I      <- diag(n)
  S      <- I - rho * W
  epsi   <- crossprod(t(S), y) - crossprod(t(X), beta)
  P1     <- W
  #W2    <- crossprod(t(W), W)
  W2     <- tcrossprod(W, W)
  P2     <- W2 - (tr(W2) / n) * I
  g.lin  <- crossprod(H, epsi)           # k_x * 1
  g.q1   <- crossprod(epsi, P1) %*% epsi # 1*1
  g.q2   <- crossprod(epsi, P2) %*% epsi # 1*1
```

```

g      <- rbind(g.q1, g.q2, g.lin)

# Gradient
P1s <- P1 + t(P1)
P2s <- P2 + t(P2)
# (2 * k_k)* (k + 1)
D    <- -1 * rbind(crossprod(eps1, P1s),
                  crossprod(eps1, P2s),
                  t(H)) %*% cbind(W %*% y, X)

# Return results (note that they are divided by n)
out <- list(g = g / n , D = D / n)
return(out)
}

```

Next, we define the function to be optimized: $Q_n(\theta) = \mathbf{g}_n^\top \mathbf{\Upsilon}_n \mathbf{g}_n(\theta)$. The function `Qmin` returns the negative of $Q_n(\theta)$ along with its gradient to speed up the optimization process. The function's negative is returned because the optimization algorithm used latter maximizes rather than minimizes.

```

# Objective function to minimize
Qmin <- function(start, y, X, H, W, Psi, gradient){
  # Thus function returns the negative of:
  # - objective function g'Psi g
  # - gradient of g'Upsi g
  g.hat <- moments.lee2007(theta = start, y = y, X = X, H = H, W = W)
  Q      <- -1 * crossprod(g.hat$g, Psi) %*% g.hat$g
  if (gradient){
    D <- g.hat$D # D.hat is (2 * k_k)* (k + 1)
    Gr <- -2 * crossprod(D, Psi) %*% g.hat$g
    attr(Q, 'gradient') <- as.vector(Gr)
  }
  return(Q)
}

```

The following function returns the variance-covariance matrix of the moments given in Equation (6.43) (divided by n):

```

# Create var-cov of moments: Omega
make.vmom <- function(b.hat, y, X, H, W){
  k      <- ncol(X)
  n      <- nrow(X)
  k_x    <- ncol(H)
  rho    <- b.hat[1L]
  beta   <- b.hat[(2L:(k + 1))]
  I      <- diag(n)
  S      <- I - rho * W

```

```

epsi      <- crossprod(t(S), y) - crossprod(t(X), beta)
sigma2    <- as.numeric(crossprod(epsi) / n)
P1        <- W
#W2       <- crossprod(t(W), W)
W2        <- tcrossprod(t(W), W)
P2        <- W2 - (tr(W2) / n) * I
P1s       <- P1 + t(P1)
P2s       <- P2 + t(P2)

# Construct V: (2 + k.x) * (2 + k.x)
V22 <- (1 / sigma2) * crossprod(H) # k_x + k_x
Delta <- matrix(0, nrow = 2, ncol = 2)
Delta[1, 1] <- tr(P1 %*% P1s)
Delta[1, 2] <- tr(P1 %*% P2s)
Delta[2, 1] <- tr(P2 %*% P1s)
Delta[2, 2] <- tr(P2 %*% P2s)
V <- matrix(0, nrow = (k_x + 2), ncol = (k_x + 2))
V[1:2, 1:2] <- Delta
V[3:(k_x + 2), 3:(k_x + 2)] <- V22
V <- sigma2^2 * V
# Construct first part of Omega
omega <- cbind(diag(P1), diag(P2)) # n * 2
mu4.hat <- sum(epsi^4) / n
mu3.hat <- sum(epsi^3) / n
Vp1 <- matrix(0, nrow = (k_x + 2), ncol = (k_x + 2))
Vp1[1:2, 1:2] <- (mu4.hat - 3 * sigma2^2) * crossprod(omega) # 2 * 2
Vp1[1:2, 3:(k_x + 2)] <- mu3.hat * crossprod(omega, H)
Vp1[3:(k_x + 2), 1:2] <- mu3.hat * crossprod(H, omega)
Omega <- Vp1 + V
Omega <- Omega / n
}

```

With all the previous function, we now crate the main function to estimate either the GMM or OGMM estimator. If `estimator = "gmm"`, the function uses an identity matrix as weighting matrix, $\mathbf{Y}_n = \mathbf{I}_{2+k_x}$, for the optimization procedure of $Q_n(\boldsymbol{\theta})$. If `estimator = "ogmm"`, the GMM estimates are used in a second step to construct $\hat{\boldsymbol{\Omega}}_n$ and optimize $Q_n(\boldsymbol{\theta}) = \mathbf{g}_n(\boldsymbol{\theta})^\top \hat{\boldsymbol{\Omega}}_n^{-1} \mathbf{g}_n(\boldsymbol{\theta})$.

```

# Function to estimate the SLM using GMME or OGMME
slm.gmm <- function(formula, data, W, instruments = 2,
                    estimator = c("gmm", "ogmm"),
                    gradient = TRUE){
  # Model Frame
  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())

# Estimator
estimator <- match.arg(estimator)

# Get variables and globals
y <- model.response(mf)
X <- model.matrix(formula, mf)
n <- nrow(X)
Wy <- W %*% y
sn <- nrow(W)
if (n != sn) stop("number of spatial units in W is different to the number of data")

# Linear instruments
H <- cbind(X, make.H(W = W, X = X, l = instruments))

# Starting values for optimization
start <- c(cor(Wy, y), coef(lm(y ~ X - 1)))
names(start) <- c("Wy", colnames(X))

# GMM estimator with weighting matrix using a identity matrix
k_x <- ncol(H)
Psi <- diag(2 + k_x)
require("maxLik")
opt <- maxLik(logLik = Qmin,
              start = start,
              method = "bfgs",
              y = y,
              X = X,
              H = H,
              W = W,
              Psi = Psi,
              gradient = gradient,
              print.level = 3,
              finalHessian = FALSE)

# OGMM: GMM estimator with weighting matrix using the inverse of the var-cov of moment
if (estimator == "ogmm"){
  # Compute Omega.hat/n
  Omega.hat <- make.vmom(coef(opt), y = y, X = X, H = H, W = W)
  Psi <- solve(Omega.hat)
  opt <- maxLik(logLik = Qmin,
                start = coef(opt),
                method = "bfgs",
                y = y,

```

```

        X = X,
        H = H,
        W = W,
        Psi = Psi,
        gradient = gradient,
        print.level = 3,
        finalHessian = FALSE)
}

results <- structure(
  list(
    coefficients = coef(opt),
    call = callT,
    X = X,
    H = H,
    y = y,
    Psi = Psi,
    W = W,
    estimator = estimator
  ),
  class = "gmm.slm"
)
return(results)
}

```

The following function returns the matrix D_n evaluated at $\hat{\theta}_n$, given in Equation (6.42)

```

# Create D matrix for asymptotic distribution
make.D <- function(rho, beta, y, X, H, W){
  n      <- nrow(X)
  k      <- ncol(X)
  k_x    <- ncol(H)
  I      <- diag(n)
  S      <- I - rho * W
  epsi   <- crossprod(t(S), y) - crossprod(t(X), beta)
  sigma2 <- as.numeric(crossprod(epsi) / n)
  P1     <- W
  #W2     <- crossprod(t(W), W)
  W2     <- tcrossprod(W, W)
  P2     <- W2 - (tr(W2) / n) * diag(n)
  P1s    <- P1 + t(P1)
  P2s    <- P2 + t(P2)
  G      <- W %*% solve(S)

  # Gen D
  D <- matrix(0, nrow = (k_x + 2) , ncol = k + 1)

```

```

rownames(D) <- c("q1", "q2", colnames(H))
colnames(D) <- c("Wy", colnames(X))
D[1, 1] <- sigma2 * tr(P1s %*% G)
D[2, 1] <- sigma2 * tr(P2s %*% G)
D[3:(k_x + 2), 1] <- t(H) %*% (G %*% X %*% beta)
D[3:(k_x + 2), 2:(k + 1)] <- t(H) %*% X
return(D)
}

```

The following functions provides the S3 methods `vcov`, `summary`, and `print.summary`. The `vcov` function for `gmm.slm` class allows to compute the VC matrix using either the asymptotic matrix D_n evaluated at $\hat{\theta}$ or the gradient evaluated at $\hat{\theta}$.

```

# Variance-covariance matrix
vcov.gmm.slm <- function(object, D = c("population", "gradient"), ...){
  estimator <- object$estimator
  D.type <- match.arg(D)
  X <- object$X
  H <- object$H
  y <- object$y
  W <- object$W
  k <- ncol(X)
  n <- nrow(X)
  b.hat <- object$coefficients
  rho <- b.hat[1L]
  beta <- b.hat[(2L:(k + 1))]
  # Matrix D is (k_x + 2) * (k + 1)
  if (estimator == "gmm"){
    if (D.type == "population"){
      D <- make.D(rho = rho, beta = beta, y = y, X = X, H = H, W = W) / n
    }
    if (D.type == "gradient"){
      D <- moments.lee2007(b.hat, y = y, X = X, H = H, W = W)$D
    }
    Omega <- make.vmom(b.hat = b.hat, y = y, X = X, H = H, W = W)
    var <- solve(crossprod(D)) %*% t(D) %*% Omega %*% D %*% solve(crossprod(D)) / n
  }
  if (estimator == "ogmm"){
    if (D.type == "population"){
      D <- make.D(rho = rho, beta = beta, y = y, X = X, H = H, W = W) / n
    }
    if (D.type == "gradient"){
      D <- moments.lee2007(b.hat, y = y, X = X, H = H, W = W)$D
    }
    Psi <- object$Psi
    var <- solve(t(D) %*% Psi %*% D) / n
  }
}

```

```

    return(var)
}

summary.gmm.slm <- function(object,
                             D = c("population", "gradient"),
                             table = TRUE,
                             digits = max(3, .Options$digits - 3),
                             ...){
  D.type <- match.arg(D)
  n      <- nrow(object$X)
  df     <- n - length(object$coefficients)
  b      <- object$coefficients
  std.err <- sqrt(diag(vcov(object, D = D.type)))
  z      <- b / std.err
  p      <- 2 * pt(-abs(z), df = df)
  CoefTable <- cbind(b, std.err, z, p)
  colnames(CoefTable) <- c("Estimate", "Std.Error", "t-value", "Pr(>|t|)")
  result <- structure(
    list(
      CoefTable = CoefTable,
      digits    = digits,
      call      = object$call),
    class = 'summary.gmm.slm'
  )
  return(result)
}

print.summary.gmm.slm <- function(x,
                                  digits = x$digits,
                                  na.print = "",
                                  symbolic.cor = p > 4,
                                  signif.stars = getOption("show.signif.stars"),
                                  ...){
  {
    cat("\nCall:\n")
    cat(paste(deparse(x$call), sep = "\n", collapse = "\n"), "\n\n", sep = "")

    cat("\nCoefficients:\n")
    printCoefmat(x$CoefTable, digit = digits, P.value = TRUE, has.Pvalue = TRUE)
    invisible(NULL)
  }
}

```

Now, we test the function `slm.gmm` estimating the GMM and OGMM procedure, along with standard errors estimated using the matrix D_n evaluated at $\hat{\theta}$, $D = \text{"population"}$, and the gradient evaluated at $\hat{\theta}$, $D = \text{"gradient"}$.


```
# Test
gmm <- slm.gmm(y ~ x1 + x2 + x3, data = data, instruments = 2, W = W)

## Initial function value: -0.1282094
## Initial gradient value:
##           Wy (Intercept)           x1           x2           x3
## -0.51461366 -0.32569203  0.37973090  0.03900924 -0.48286479
## initial  value 0.128209
## iter    2 value 0.028741
## iter    3 value 0.019024
## iter    4 value 0.008846
## iter    5 value 0.008408
## iter    6 value 0.008161
## iter    7 value 0.008160
## iter    8 value 0.008160
## iter    9 value 0.008160
## iter   10 value 0.008160
## iter   10 value 0.008160
## iter   10 value 0.008160
## final   value 0.008160
## converged

summary(gmm, D = "population")

##
## Call:
## slm.gmm(formula = y ~ x1 + x2 + x3, data = data, W = W, instruments = 2)
##
##
## Coefficients:
##           Estimate Std.Error t-value Pr(>|t|)
## Wy           0.57527   0.04111  13.994   <2e-16 ***
## (Intercept)  0.06118   0.06287   0.973   0.3309
## x1          -1.01976   0.06190 -16.475   <2e-16 ***
## x2           0.11041   0.06586   1.676   0.0943 .
## x3           1.02969   0.06328  16.272   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(gmm, D = "gradient")

##
## Call:
## slm.gmm(formula = y ~ x1 + x2 + x3, data = data, W = W, instruments = 2)
##
##
## Coefficients:
```

```
##              Estimate Std.Error t-value Pr(>|t|)
## Wy          0.57527   0.03921  14.670   <2e-16 ***
## (Intercept) 0.06118   0.06271   0.976   0.330
## x1          -1.01976   0.06242 -16.337   <2e-16 ***
## x2           0.11041   0.06581   1.678   0.094 .
## x3           1.02969   0.06572  15.667   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

ogmm <- slm.gmm(y ~ x1 + x2 + x3, data = data, instruments = 2, W = W,
  estimator = "ogmm")

## Initial function value: -0.1282094
## Initial gradient value:
##              Wy (Intercept)          x1          x2          x3
## -0.51461366 -0.32569203  0.37973090  0.03900924 -0.48286479
## initial value 0.128209
## iter  2 value 0.028741
## iter  3 value 0.019024
## iter  4 value 0.008846
## iter  5 value 0.008408
## iter  6 value 0.008161
## iter  7 value 0.008160
## iter  8 value 0.008160
## iter  9 value 0.008160
## iter 10 value 0.008160
## iter 10 value 0.008160
## iter 10 value 0.008160
## final value 0.008160
## converged
## Initial function value: -0.01950219
## Initial gradient value:
##              Wy (Intercept)          x1          x2          x3
##  0.108923926 -0.003093326 -0.004256760 -0.004647850 -0.015398993
## initial value 0.019502
## iter  2 value 0.017867
## iter  3 value 0.017115
## iter  4 value 0.017106
## iter  5 value 0.017106
## iter  6 value 0.017106
## iter  6 value 0.017106
## iter  6 value 0.017106
## final value 0.017106
## converged

summary(ogmm, D = "population")
```

```
##
## Call:
## slm.gmm(formula = y ~ x1 + x2 + x3, data = data, W = W, instruments = 2,
##       estimator = "ogmm")
##
##
## Coefficients:
##           Estimate Std.Error t-value Pr(>|t|)
## Wy           0.61473   0.03510  17.513   <2e-16 ***
## (Intercept)   0.04586   0.06236   0.735    0.462
## x1          -1.01472   0.06157 -16.481   <2e-16 ***
## x2           0.10361   0.06567   1.578    0.115
## x3           0.99502   0.06276  15.855   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(ogmm, D = "gradient")

##
## Call:
## slm.gmm(formula = y ~ x1 + x2 + x3, data = data, W = W, instruments = 2,
##       estimator = "ogmm")
##
##
## Coefficients:
##           Estimate Std.Error t-value Pr(>|t|)
## Wy           0.61473   0.03499  17.569   <2e-16 ***
## (Intercept)   0.04586   0.06247   0.734    0.463
## x1          -1.01472   0.06166 -16.456   <2e-16 ***
## x2           0.10361   0.06561   1.579    0.115
## x3           0.99502   0.06405  15.535   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 6.1 compares different estimates for the SLM using all previous estimators: MLE, S2SLSE, BS2SLSE, GMME and OGMME. Similarly, Table 6.2 compares the SEs.

Table 6.1: Comparing coefficients for SLM.

	mle	s2sls	bs2sls	gmm	ogmm
b0	-0.1086	-0.0698	-0.0816	0.0612	0.0459
b1	-1.0286	-1.0057	-1.0127	-1.0198	-1.0147
b2	-0.0383	-0.0435	-0.0420	0.1104	0.1036
b3	1.0507	1.0149	1.0257	1.0297	0.9950
rho	0.5391	0.6473	0.6146	0.5753	0.6147

Table 6.2: Comparing SE for SLM.

	mle	s2sls-ho	s2sls-he	bs2sls-i	bs2sls-f	gmm-p	gmm-g	ogmm-p	ogmm-g
b0	0.0647	0.0666	0.0668	0.0660	0.0663	0.0629	0.0627	0.0624	0.0625
b1	0.0632	0.0635	0.0659	0.0642	0.0640	0.0619	0.0624	0.0616	0.0617
b2	0.0633	0.0628	0.0629	0.0629	0.0629	0.0659	0.0658	0.0657	0.0656
b3	0.0652	0.0672	0.0667	0.0665	0.0664	0.0633	0.0657	0.0628	0.0641
rho	0.0379	0.0622	0.0589	0.0580	0.0603	0.0411	0.0392	0.0351	0.0349

6.3.4 Best GMM estimator

Lee (2007) proposes the optimal instrument for the OGMM estimator. The best linear instruments are:

$$\mathbf{H}_n^* = \mathbb{E}(\mathbf{Z}_n | \mathbf{X}_n) = (\mathbf{G}_0 \mathbf{X}_n \beta_0, \mathbf{X}_n).$$

There are two cases which can make the selection of best quadratic matrices easier.

Selection from \mathcal{P}_{2n}

For the best selection of \mathbf{P}_{jn} , consider $\mathbf{P}_{jn} \in \mathcal{P}_{2n}$. In this case $\boldsymbol{\omega}_n$ is a $n \times 1$ matrix of zeros and $\boldsymbol{\Omega}_n$ becomes:

$$\boldsymbol{\Omega}_n = \mathbf{V}_n = \sigma_0^4 \begin{pmatrix} \text{tr}(\mathbf{P}_{1n} \mathbf{P}_{1n}^s) & \text{tr}(\mathbf{P}_{1n} \mathbf{P}_{2n}^s) & \mathbf{0} \\ \text{tr}(\mathbf{P}_{2n} \mathbf{P}_{1n}^s) & \text{tr}(\mathbf{P}_{2n} \mathbf{P}_{2n}^s) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{1}{\sigma_0^2} \mathbf{H}_n^\top \mathbf{H}_n \end{pmatrix} = \sigma_0^4 \begin{pmatrix} \boldsymbol{\Delta}_{2n} & \mathbf{0} \\ \mathbf{0} & \frac{1}{\sigma_0^2} \mathbf{H}_n^\top \mathbf{H}_n \end{pmatrix},$$

with $\boldsymbol{\Delta}_{2n} = [\text{vec}(\mathbf{P}_{1n})^\top, \text{vec}(\mathbf{P}_{2n})^\top]^\top [\text{vec}(\mathbf{P}_{1n}^s), \text{vec}(\mathbf{P}_{2n}^s)]$, because, for any conformable matrices \mathbf{A} and \mathbf{B} , $\text{tr}(\mathbf{AB}) = \text{vec}(\mathbf{A}^\top)^\top \text{vec}(\mathbf{B})$ and

$$\begin{aligned} \mathbf{D}_n^\top \boldsymbol{\Omega}_n^{-1} \mathbf{D}_n &= \frac{1}{\sigma_0^4} \begin{pmatrix} \sigma_0^2 \text{tr}(\mathbf{P}_{1n}^s \mathbf{G}_0) & \sigma_0^2 \text{tr}(\mathbf{P}_{2n}^s \mathbf{G}_0) & (\mathbf{G}_0 \mathbf{X}_n \beta_0)^\top \mathbf{H}_n \\ \mathbf{0} & \mathbf{0} & \mathbf{X}_n^\top \mathbf{H}_n \end{pmatrix} \\ &\quad \begin{pmatrix} \text{tr}(\mathbf{P}_{1n} \mathbf{P}_{1n}^s)^{-1} & \text{tr}(\mathbf{P}_{1n} \mathbf{P}_{2n}^s)^{-1} & \mathbf{0} \\ \text{tr}(\mathbf{P}_{2n} \mathbf{P}_{1n}^s)^{-1} & \text{tr}(\mathbf{P}_{2n} \mathbf{P}_{2n}^s)^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_0^2 (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \end{pmatrix} \begin{pmatrix} \sigma_0^2 \text{tr}(\mathbf{P}_{1n}^s \mathbf{G}_0) & \mathbf{0} \\ \sigma_0^2 \text{tr}(\mathbf{P}_{2n}^s \mathbf{G}_0) & \mathbf{0} \\ \mathbf{H}_n^\top \mathbf{G}_0 \mathbf{X}_n \beta_0 & \mathbf{H}_n^\top \mathbf{X}_n \end{pmatrix}, \\ &= \frac{1}{\sigma_0^4} \begin{pmatrix} \sigma_0^2 \mathbf{C}_{2n} & (\mathbf{G}_0 \mathbf{X}_n \beta_0)^\top \mathbf{H}_n \\ \mathbf{0} & \mathbf{X}_n^\top \mathbf{H}_n \end{pmatrix} \begin{pmatrix} \boldsymbol{\Delta}_{2n}^{-1} & \mathbf{0} \\ \mathbf{0} & \sigma_0^2 (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \end{pmatrix} \begin{pmatrix} \sigma_0^2 \mathbf{C}_{2n}^\top & \mathbf{0} \\ \mathbf{H}_n^\top \mathbf{G}_0 \mathbf{X}_n \beta_0 & \mathbf{H}_n^\top \mathbf{X}_n \end{pmatrix}, \\ &= \frac{1}{\sigma_0^4} \begin{pmatrix} \sigma_0^2 \mathbf{C}_{2n} \boldsymbol{\Delta}_{2n}^{-1} & \sigma_0^2 (\mathbf{G}_0 \mathbf{X}_n \beta_0)^\top \mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \\ \mathbf{0} & \sigma_0^2 \mathbf{X}_n^\top \mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \end{pmatrix} \begin{pmatrix} \sigma_0^2 \mathbf{C}_{2n}^\top & \mathbf{0} \\ \mathbf{H}_n^\top \mathbf{G}_0 \mathbf{X}_n \beta_0 & \mathbf{H}_n^\top \mathbf{X}_n \end{pmatrix}, \\ &= \begin{pmatrix} \mathbf{C}_{2n} \boldsymbol{\Delta}_{2n}^{-1} \mathbf{C}_{2n}^\top + \frac{1}{\sigma_0^2} (\mathbf{G}_0 \mathbf{X}_n \beta_0)^\top \mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{G}_0 \mathbf{X}_n \beta_0 & \frac{1}{\sigma_0^2} (\mathbf{G}_0 \mathbf{X}_n \beta_0)^\top \mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{X}_n \\ \frac{1}{\sigma_0^2} \mathbf{X}_n^\top \mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{G}_0 \mathbf{X}_n \beta_0 & \frac{1}{\sigma_0^2} \mathbf{X}_n^\top \mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{X}_n \end{pmatrix}, \\ &= \begin{pmatrix} \mathbf{C}_{2n} \boldsymbol{\Delta}_{2n}^{-1} \mathbf{C}_{2n}^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} + \frac{1}{\sigma_0^2} \begin{pmatrix} (\mathbf{G}_0 \mathbf{X}_n \beta_0)^\top \\ \mathbf{X}_n^\top \end{pmatrix} \mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top (\mathbf{G}_0 \mathbf{X}_n \beta_0 \quad \mathbf{X}_n), \end{aligned}$$

where $\mathbf{C}_{2n} = [\text{tr}(\mathbf{P}_{1n}^s \mathbf{G}_0), \text{tr}(\mathbf{P}_{2n}^s \mathbf{G}_0)]$. With the best instruments \mathbf{H}_n^* for \mathbf{H}_n , the optimal variance-covariance matrix is

$$\mathbf{D}_n^\top \boldsymbol{\Omega}_n^{-1} \mathbf{D}_n = \begin{pmatrix} \mathbf{C}_{2n} \boldsymbol{\Delta}_{2n}^{-1} \mathbf{C}_{2n}^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} + \frac{1}{\sigma_0^2} ((\mathbf{G}_0 \mathbf{X}_n \beta_0) \quad \mathbf{X}_n)^\top (\mathbf{G}_0 \mathbf{X}_n \beta_0 \quad \mathbf{X}_n).$$

Since $\text{tr}(\mathbf{P}_{jn}\mathbf{P}_{ln}^s) = \frac{1}{2} \text{tr}(\mathbf{P}_{jn}^s\mathbf{P}_{ln}^s)$, we can also write

$$\begin{aligned}\Delta_{2n} &= \frac{1}{2} \begin{pmatrix} \text{tr}(\mathbf{P}_{1n}^s\mathbf{P}_{1n}^s) & \text{tr}(\mathbf{P}_{1n}^s\mathbf{P}_{2n}^s) \\ \text{tr}(\mathbf{P}_{2n}^s\mathbf{P}_{1n}^s) & \text{tr}(\mathbf{P}_{2n}^s\mathbf{P}_{2n}^s) \end{pmatrix}, \\ &= \frac{1}{2} [\text{vec}(\mathbf{P}_{1n}^s), \text{vec}(\mathbf{P}_{2n}^s)]^\top [\text{vec}(\mathbf{P}_{1n}^s), \text{vec}(\mathbf{P}_{2n}^s)].\end{aligned}$$

If \mathbf{P}_{jn} s belong to \mathcal{P}_{2n} such that $\text{diag}(\mathbf{P}_{jn}) = \mathbf{0}$, then

$$\begin{aligned}\text{tr}(\mathbf{P}_{jn}^s\mathbf{G}_0) &= \frac{1}{2} \text{tr} [\mathbf{P}_{jn}^s (\mathbf{G}_0 - \text{Diag}(\mathbf{G}_0))^s], \\ &= \frac{1}{2} \text{vec} ([\mathbf{G}_0 - \text{Diag}(\mathbf{G}_0)]^s)^\top \text{vec}(\mathbf{P}_{jn}^s).\end{aligned}$$

Schwartz inequality implies that

$$\begin{aligned}\mathbf{C}_{2n}\Delta_{2n}^{-1}\mathbf{C}_{2n}^\top &\leq \frac{1}{2} \text{vec} ([\mathbf{G}_0 - \text{Diag}(\mathbf{G}_0)]^s)^\top \frac{1}{2} \text{vec} ([\mathbf{G}_0 - \text{Diag}(\mathbf{G}_0)]^s) \\ &= \text{tr} ([\mathbf{G}_0 - \text{Diag}(\mathbf{G}_0)]^s \mathbf{G}_0).\end{aligned}$$

Therefore, in the subclass \mathcal{P}_{2n} , $[\mathbf{G}_0 - \text{Diag}(\mathbf{G}_0)]$ and $[\mathbf{G}_0\mathbf{X}_n\boldsymbol{\beta}_0, \mathbf{X}_n]$ provide the set of best IV functions, and the vector of moment functions is

$$\mathbf{g}_{b,n} = \begin{pmatrix} \hat{\boldsymbol{\varepsilon}}_n(\boldsymbol{\theta})^\top (\hat{\mathbf{G}}_n - \text{Diag}(\hat{\mathbf{G}}_n)) \hat{\boldsymbol{\varepsilon}}_n(\boldsymbol{\theta}) \\ \hat{\boldsymbol{\beta}}_n^\top \mathbf{X}_n^\top \hat{\mathbf{G}}_n^\top \hat{\boldsymbol{\varepsilon}}_n(\boldsymbol{\theta}) \\ \mathbf{X}_n^\top \hat{\boldsymbol{\varepsilon}}_n(\boldsymbol{\theta}) \end{pmatrix},$$

and the corresponding estimated \mathbf{V}_n is

$$\hat{\mathbf{V}}_n = \hat{\sigma}_n^4 \begin{pmatrix} \text{tr} \left([\hat{\mathbf{G}}_n - \text{Diag}(\hat{\mathbf{G}}_n)]^s \hat{\mathbf{G}}_n \right) & \mathbf{0} \\ \mathbf{0} & \frac{1}{\hat{\sigma}_n^2} \left(\hat{\mathbf{G}}_n \mathbf{X}_n \hat{\boldsymbol{\beta}}, \mathbf{X}_n \right)^\top \left(\hat{\mathbf{G}}_n \mathbf{X}_n \hat{\boldsymbol{\beta}}, \mathbf{X}_n \right) \end{pmatrix}, \quad (6.49)$$

and the Feasible Best GMME (FBGMM) with \mathcal{P}_{2n} will be from minimizing:

$$Q_n(\boldsymbol{\theta}) = \mathbf{g}_{b,n}^\top(\boldsymbol{\theta}) \hat{\mathbf{V}}_n^{-1} \mathbf{g}_{b,n}(\boldsymbol{\theta}).$$

Assuming Normality

If $\boldsymbol{\varepsilon}_n \sim N(\mathbf{0}, \sigma_0^2 \mathbf{I}_n)$, then $\boldsymbol{\Omega}_n = \mathbf{V}_n$ even though \mathbf{P}_{jn} 's are selected from the class \mathcal{P}_{1n} .

Because, for any $\mathbf{P}_{jn} \in \mathcal{P}_{1n}$,

$$\text{tr}(\mathbf{P}_{jn}^s \mathbf{G}_0) = \frac{1}{2} \text{vec} \left(\left[\mathbf{G}_0 - \frac{\text{tr}(\mathbf{G}_0)}{n} \mathbf{I}_n \right]^s \mathbf{G}_0 \right)^\top \text{vec}(\mathbf{P}_{jn}^s),$$

the generalized Schwartz inequality implies that

$$\mathbf{C}_{2n}\Delta_{2n}^{-1}\mathbf{C}_{2n}^\top \leq \text{tr} \left(\left[\mathbf{G}_0 - \frac{\text{tr}(\mathbf{G}_0)}{n} \mathbf{I}_n \right]^s \mathbf{G}_0 \right).$$

Therefore, in the broader class \mathcal{P}_{1n} , $\left[\mathbf{G}_0 - \frac{\text{tr}(\mathbf{G}_0)}{n} \mathbf{I}_n\right]$ and $[\mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0, \mathbf{X}_n]$ provide the best set of IV functions. Thus, the vector of moment functions is

$$\mathbf{g}_{b,n} = \begin{pmatrix} \widehat{\boldsymbol{\varepsilon}}_n(\boldsymbol{\theta})^\top \left(\widehat{\mathbf{G}}_n - \frac{\text{tr}(\widehat{\mathbf{G}}_n)}{n} \mathbf{I}_n \right) \widehat{\boldsymbol{\varepsilon}}_n(\boldsymbol{\theta}) \\ \widehat{\boldsymbol{\beta}}_n^\top \mathbf{X}_n^\top \widehat{\mathbf{G}}_n^\top \widehat{\boldsymbol{\varepsilon}}_n(\boldsymbol{\theta}) \\ \mathbf{X}_n^\top \widehat{\boldsymbol{\varepsilon}}_n(\boldsymbol{\theta}) \end{pmatrix},$$

and the corresponding estimated \mathbf{V}_n is

$$\widehat{\mathbf{V}}_n = \widehat{\sigma}_n^4 \begin{pmatrix} \text{tr} \left(\left[\widehat{\mathbf{G}}_n - \frac{\text{tr}(\widehat{\mathbf{G}}_n)}{n} \mathbf{I}_n \right]^s \widehat{\mathbf{G}}_n \right) & \mathbf{0} \\ \mathbf{0} & \frac{1}{\widehat{\sigma}_n^2} \left(\widehat{\mathbf{G}}_n \mathbf{X}_n \widehat{\boldsymbol{\beta}}, \mathbf{X}_n \right)^\top \left(\widehat{\mathbf{G}}_n \mathbf{X}_n \widehat{\boldsymbol{\beta}}, \mathbf{X}_n \right) \end{pmatrix}. \quad (6.50)$$

Theorem 6.16 — BGMME for SLM (Lee, 2007). Under Assumptions (a)-(c) of 6.13, suppose that $\widehat{\rho}_n$ is a \sqrt{n} -consistent estimate of ρ_0 , $\widehat{\boldsymbol{\beta}}_n$ is a consistent estimate of $\boldsymbol{\beta}_0$, and $\widehat{\sigma}_n^2$ is a consistent estimate of σ_0^2 . Within the class of GMMEs derived with \mathcal{P}_{2n} , the Best GMM estimator (BGMME) $\widehat{\boldsymbol{\theta}}_n$ has the limiting distribution that

$$\sqrt{n} \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \right) \xrightarrow{d} \text{N} \left(\mathbf{0}, \boldsymbol{\Sigma}^{b2} \right),$$

where

$$\boldsymbol{\Sigma}^{b2} = \lim_{n \rightarrow \infty} \frac{1}{n} \begin{pmatrix} \text{tr} \left([\mathbf{G}_0 - \text{Diag}(\mathbf{G}_0)]^s \mathbf{G}_0 \right) + \frac{1}{\sigma_0^2} (\mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0)^\top (\mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0) & \frac{1}{\sigma_0^2} (\mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{X}_n \\ \frac{1}{\sigma_0^2} \mathbf{X}_n^\top (\mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0) & \frac{1}{\sigma_0^2} \mathbf{X}_n^\top \mathbf{X}_n \end{pmatrix},$$

with $\mathbf{G}_0 = \mathbf{W}_n (\mathbf{I}_n - \rho_0 \mathbf{W}_n)^{-1}$, which is assumed to exist.

When $\boldsymbol{\varepsilon}_n \sim \text{N}(\mathbf{0}, \sigma_0^2 \mathbf{I}_n)$, within the broader class of GMMEs derived with \mathcal{P}_{1n} , the BGMME $\widehat{\boldsymbol{\theta}}_n$ has the limiting distribution that

$$\sqrt{n} \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \right) \xrightarrow{d} \text{N} \left(\mathbf{0}, \boldsymbol{\Sigma}^{b1} \right),$$

where

$$\boldsymbol{\Sigma}^{b1} = \lim_{n \rightarrow \infty} \frac{1}{n} \begin{pmatrix} \text{tr} \left(\left[\mathbf{G}_0 - \frac{\text{tr}(\mathbf{G}_0)}{n} \mathbf{I}_n \right]^s \mathbf{G}_0 \right) + \frac{1}{\sigma_0^2} (\mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0)^\top (\mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0) & \frac{1}{\sigma_0^2} (\mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{X}_n \\ \frac{1}{\sigma_0^2} \mathbf{X}_n^\top (\mathbf{G}_0 \mathbf{X}_n \boldsymbol{\beta}_0) & \frac{1}{\sigma_0^2} \mathbf{X}_n^\top \mathbf{X}_n \end{pmatrix},$$

which is assumed to exist.

6.3.5 S2SLS Estimator as GMM Estimator

The moment conditions for the S2SLS estimators are

$$\mathbf{g}_n = \frac{1}{n} \mathbf{H}_n^\top \boldsymbol{\varepsilon}_n = \frac{1}{n} \mathbf{H}_n^\top (\mathbf{y}_n - \mathbf{Z}_n \boldsymbol{\delta}),$$

whose variance-covariance matrix is:

$$\frac{1}{n} \boldsymbol{\Omega}_n = \frac{1}{n} \sigma_0^2 \mathbf{H}_n^\top \mathbf{H}_n.$$

The matrix \mathbf{r}_n^{-1} is the optimal weight matrix, which correspond to the inverse of the covariance matrix of the sample moments:

$$\mathbf{r}_n^{-1} = \left(\frac{1}{n} \sigma_0^2 \mathbf{H}_n^\top \mathbf{H}_n \right)^{-1}.$$

Then, the function to minimize is:

$$Q = \frac{1}{n\hat{\sigma}^2} \left\{ [\mathbf{H}_n^\top \mathbf{y}_n - \mathbf{H}_n^\top \mathbf{Z}_n \boldsymbol{\delta}]^\top (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} [\mathbf{H}_n^\top \mathbf{y}_n - \mathbf{H}_n^\top \mathbf{Z}_n \boldsymbol{\delta}] \right\}$$

Obtaining the first order conditions and solving for $\boldsymbol{\delta}$, we obtain:

$$\hat{\boldsymbol{\delta}}_n = (\mathbf{Z}_n^\top \mathbf{P}_{H,n} \mathbf{Z}_n)^{-1} \mathbf{Z}_n^\top \mathbf{P}_{H,n} \mathbf{y}_n. \quad (6.51)$$

Its asymptotic distribution is

$$\sqrt{n} (\hat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0) \xrightarrow{d} N \left(\mathbf{0}, \sigma_0^2 \lim_{n \rightarrow \infty} \left\{ \frac{1}{n} (\mathbf{G}_n \mathbf{X}_n \boldsymbol{\beta}_0)^\top \mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top (\mathbf{G}_n \mathbf{X}_n \boldsymbol{\beta}_0) \right\}^{-1} \right).$$

6.4 Feasible Generalized Least Squares Estimator for SEM Model

[Kelejian and Prucha \(1999\)](#) derive a Method of Moments (MOM) estimator for λ , which is later used in the Feasible Generalized Least Squares (FGLS) estimator. The motivation behind this new estimator, according to [Kelejian and Prucha \(1999\)](#), is that the (quasi) maximum likelihood estimator may not be computationally feasible, especially in cases with moderate or large-sized samples. As they note, the MOM estimator is computationally simple, regardless of sample size, making it highly attractive when dealing with large spatial datasets. Additionally, the IV/GMM estimators avoid the Jacobian term, alleviating many problems associated with matrix inversion, the computation of characteristic roots, and/or Cholesky decomposition. Another motivation for proposing this estimator was that, at the time, there were no formal results 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 estimator in 2004 for the Spatial Lag Model (SLM).

Recall that the SEM model is given by:

$$\begin{aligned} \mathbf{y}_n &= \mathbf{X}_n \boldsymbol{\beta}_0 + \mathbf{u}_n, \\ \mathbf{u}_n &= \lambda_0 \mathbf{M}_n \mathbf{u}_n + \boldsymbol{\varepsilon}_n. \end{aligned} \quad (6.52)$$

In brief, [Kelejian and Prucha \(1999\)](#) propose the use of **nonlinear least square** to obtain a consistent generalized moment estimate for λ_0 , which can be used to obtain consistent estimates for $\boldsymbol{\beta}_0$ in a FGLS framework. The key distinction between the MOM estimation discussed here and the Generalized Method of Moment (GMM) estimator presented later is that in the MOM approach, there is no inference for the spatial autoregressive coefficient λ . In other words, λ is treated purely as a nuisance parameter, whose only purpose is to aid in obtaining consistent estimates for $\boldsymbol{\beta}_0$.

- 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 , which is why some software packages only provide the estimate $\hat{\lambda}$ without its standard error.

One advantage of the MOM estimator is that it does not rely on the assumption of normality of the disturbances ϵ_n . Nonetheless, it does assume that ϵ_i are independently and identically distributed for all i with zero mean and constant variance σ^2 . We adopt the following assumption about the error terms, as stated by [Kelejian and Prucha \(1999\)](#).

Assumption 6.17 — Homoskedastic 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.

Next, we present the following assumptions regarding the weight matrix:

Assumption 6.18 — 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 (6.52) and Assumption 6.18 on the weight matrix, we can express the error term as $\mathbf{u}_n = (I_n - \lambda_0 M_n)^{-1} \epsilon_n = \mathbf{R}_0^{-1} \epsilon_n$. Therefore, the expectation and variance of \mathbf{u}_n are $\mathbb{E}(\mathbf{u}_n) = 0$ and $\mathbb{E}(\mathbf{u}_n \mathbf{u}_n^\top) = \Omega_{n0}$, respectively, where:

$$\Omega_{n0} = \sigma_0^2 (I_n - \lambda_0 M_n)^{-1} (I_n - \lambda_0 M_n^\top)^{-1} = \sigma_0^2 \mathbf{R}_0^{-1} \mathbf{R}_0^{-1\top}.$$

Note that a row-standardized spatial weight matrix is typically not symmetric, such that $M_n \neq M_n^\top$ and thus $(I_n - \lambda_0 M_n)^{-1} \neq (I_n - \lambda_0 M_n^\top)^{-1}$.

6.4.1 Generalized Least Squares Estimator

The primary objective in [Kelejian and Prucha \(1999\)](#) is to derive a **consistent estimator** for λ_0 to ensure the consistency of the resulting spatially weighted estimator. In this context, [Kelejian and Prucha \(1999\)](#) were not primarily concerned with inference on λ , but rather viewed it as a tool to obtain consistent estimates for β . This implies that λ is considered a **nuisance parameter**.

The Spatially Weighted Least Squares (SWLS) boils down to:

$$\hat{\beta}_{SWLS} = (\mathbf{X}_{s,n}^\top \mathbf{X}_{s,n})^{-1} \mathbf{X}_{s,n}^\top \mathbf{y}_{s,n}, \quad (6.53)$$

where $\mathbf{X}_{s,n} = \mathbf{X}_n - \lambda_0 M_n \mathbf{X}_n$ and $\mathbf{y}_{s,n} = \mathbf{y}_n - \lambda_0 M_n \mathbf{y}_n$, using a consistent estimate $\hat{\lambda}_n$ for the autoregressive parameter. Note that this model is essentially an OLS applied to spatially filtered variables. Moreover, it is important to note that SWLS is a special case of the Feasible Generalized Least Squares (FGLS) estimator.

Under the assumptions made, the Generalized Least Squares (GLS) estimator—assuming we know λ_0 —for β is:

$$\begin{aligned}\hat{\beta}_{GLS} &= [\mathbf{X}_n^\top \boldsymbol{\Omega}_{n0}^{-1} \mathbf{X}_n]^{-1} \mathbf{X}_n^\top \boldsymbol{\Omega}_{n0}^{-1} \mathbf{y}_n, \\ &= \left[\mathbf{X}_n^\top \frac{1}{\sigma^2} (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^\top (\mathbf{I}_n - \lambda_0 \mathbf{M}_n) \mathbf{X}_n \right]^{-1} \mathbf{X}_n^\top \frac{1}{\sigma^2} (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^\top (\mathbf{I}_n - \lambda_0 \mathbf{M}_n) \mathbf{y}_n, \\ &= \left[\mathbf{X}_n^\top (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^\top (\mathbf{I}_n - \lambda_0 \mathbf{M}_n) \mathbf{X}_n \right]^{-1} \mathbf{X}_n^\top (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^\top (\mathbf{I}_n - \lambda_0 \mathbf{M}_n) \mathbf{y}_n, \\ &= (\mathbf{X}_n^\top \mathbf{R}_{n0}^\top \mathbf{R}_{n0} \mathbf{X}_n)^{-1} \mathbf{X}_n^\top \mathbf{R}_{n0}^\top \mathbf{R}_{n0} \mathbf{y}_n.\end{aligned}$$

The FGLS estimator substitutes a consistent estimate for λ into this expression, resulting in the following:

$$\begin{aligned}\hat{\beta}_{FGLS} &= \left[\mathbf{X}_n^\top (\mathbf{I}_n - \hat{\lambda}_n \mathbf{M}_n)^\top (\mathbf{I}_n - \hat{\lambda}_n \mathbf{M}_n) \mathbf{X}_n \right]^{-1} \mathbf{X}_n^\top (\mathbf{I}_n - \hat{\lambda}_n \mathbf{M}_n)^\top (\mathbf{I}_n - \hat{\lambda}_n \mathbf{M}_n) \mathbf{y}_n, \\ &= (\mathbf{X}_n^\top \mathbf{R}_n^\top \mathbf{R}_n \mathbf{X}_n)^{-1} \mathbf{X}_n^\top \mathbf{R}_n^\top \mathbf{R}_n \mathbf{y}_n,\end{aligned}$$

which is precisely the same as Equation (6.53).

6.4.2 Moment Conditions

The core idea behind a MOM estimator is to find a set of population moments equations that relate population moments to the parameters of interest. These population moments are then replaced by sample moments to derive consistent estimates for the parameters, such as λ , which can be used in Equation (6.53).

Given the DGP in Equation (6.52), we express the relationship as:

$$\boldsymbol{\varepsilon}_n = \mathbf{u}_n - \lambda_0 \mathbf{M}_n \mathbf{u}_n,$$

where $\boldsymbol{\varepsilon}_n$ denotes the idiosyncratic error and \mathbf{u}_n is the regression error. The Kelejian and Prucha (1999)'s MOM estimator of λ_0 is based on these three moment conditions (see Section 6.3.1):

Definition 6.4.1 — Moment Conditions. Under homoskedasticity (Kelejian and Prucha, 1999) the moment conditions are:

$$\begin{aligned}\mathbb{E} [n^{-1} \boldsymbol{\varepsilon}_n^\top \boldsymbol{\varepsilon}_n] &= \sigma_0^2, \\ \mathbb{E} [n^{-1} \boldsymbol{\varepsilon}_n^\top \mathbf{M}_n^\top \mathbf{M}_n \boldsymbol{\varepsilon}_n] &= \frac{\sigma_0^2}{n} \text{tr} (\mathbf{M}_n^\top \mathbf{M}_n), \\ \mathbb{E} [n^{-1} \boldsymbol{\varepsilon}_n^\top \mathbf{M}_n \boldsymbol{\varepsilon}_n] &= 0.\end{aligned}$$

To implement the moment conditions, we need to express the conditions on $\boldsymbol{\varepsilon}_n$ in terms of \mathbf{u}_n , since $\boldsymbol{\varepsilon}_n$ is not observed. Given that $\mathbf{u}_n = \lambda \mathbf{M}_n \mathbf{u}_n + \boldsymbol{\varepsilon}_n$, it follows that $\boldsymbol{\varepsilon}_n = \mathbf{u}_n - \lambda_0 \mathbf{M}_n \mathbf{u}_n$, i.e., the spatially filtered regression error terms. This leads to the following relationships:

$$\begin{aligned}\boldsymbol{\varepsilon}_n^\top \boldsymbol{\varepsilon}_n &= (\mathbf{u}_n - \lambda_0 \mathbf{M}_n \mathbf{u}_n)^\top (\mathbf{u}_n - \lambda_0 \mathbf{M}_n \mathbf{u}_n), \\ &= \mathbf{u}_n^\top \mathbf{u}_n - 2\lambda_0 \mathbf{u}_n^\top \mathbf{M}_n \mathbf{u}_n + \lambda_0^2 \mathbf{u}_n^\top \mathbf{M}_n^\top \mathbf{M}_n \mathbf{u}_n, \\ \boldsymbol{\varepsilon}_n^\top \mathbf{M}_n^\top \mathbf{M}_n \boldsymbol{\varepsilon}_n &= (\mathbf{u}_n - \lambda_0 \mathbf{M}_n \mathbf{u}_n)^\top \mathbf{M}_n^\top \mathbf{M}_n (\mathbf{u}_n - \lambda_0 \mathbf{M}_n \mathbf{u}_n),\end{aligned}\tag{6.54}$$

$$\begin{aligned}
 &= \mathbf{u}_n^\top \mathbf{M}_n^\top \mathbf{M}_n \mathbf{u}_n - 2\lambda_0 \mathbf{u}_n^\top \mathbf{M}_n^\top \mathbf{M}_n \mathbf{M}_n \mathbf{u}_n + \lambda_0^2 \mathbf{u}_n^\top \mathbf{M}_n^\top \mathbf{M}_n \mathbf{M}_n^\top \mathbf{M}_n \mathbf{u}_n, \\
 \varepsilon_n^\top \mathbf{M}_n \varepsilon_n &= (\mathbf{u}_n - \lambda_0 \mathbf{M}_n \mathbf{u}_n)^\top \mathbf{M}_n (\mathbf{u}_n - \lambda_0 \mathbf{M}_n \mathbf{u}_n), \\
 &= \mathbf{u}_n^\top \mathbf{M}_n \mathbf{u}_n - 2\lambda_0 \mathbf{u}_n^\top \mathbf{M}_n \mathbf{M}_n \mathbf{u}_n + \lambda_0^2 \mathbf{u}_n^\top \mathbf{M}_n^\top \mathbf{M}_n \mathbf{M}_n \mathbf{u}_n.
 \end{aligned} \tag{6.55}$$

Defining $\mathbf{u}_L = \mathbf{M}_n \mathbf{u}_n$, $\mathbf{u}_{LL} = \mathbf{M}_n \mathbf{M}_n \mathbf{u}_n$,⁷ we take expectations in (6.54) under the assumption of homoskedasticity:

$$\begin{aligned}
 \mathbb{E}[\varepsilon_n^\top \varepsilon_n] &= \mathbb{E}[\mathbf{u}_n^\top \mathbf{u}_n] - 2\lambda_0 \mathbb{E}[\mathbf{u}_n^\top \mathbf{M}_n \mathbf{u}_n] + \lambda_0^2 \mathbb{E}[\mathbf{u}_n^\top \mathbf{M}_n^\top \mathbf{M}_n \mathbf{u}_n], \\
 \sigma_0^2 &= \frac{1}{n} \mathbb{E}[\mathbf{u}_n^\top \mathbf{u}_n] - \lambda_0 \frac{2}{n} \mathbb{E}[\mathbf{u}_n^\top \mathbf{u}_L] + \lambda_0^2 \frac{1}{n} \mathbb{E}[\mathbf{u}_L^\top \mathbf{u}_L] \quad \text{since } \mathbb{E}[n^{-1} \varepsilon_n^\top \varepsilon_n] = \sigma_0^2, \\
 0 &= \sigma_0^2 - \frac{1}{n} \mathbb{E}[\mathbf{u}_n^\top \mathbf{u}_n] + \lambda_0 \frac{2}{n} \mathbb{E}[\mathbf{u}_n^\top \mathbf{u}_L] - \lambda_0^2 \frac{1}{n} \mathbb{E}[\mathbf{u}_L^\top \mathbf{u}_L], \\
 0 &= \lambda_0 \frac{2}{n} \mathbb{E}[\mathbf{u}_n^\top \mathbf{u}_L] - \lambda_0^2 \frac{1}{n} \mathbb{E}[\mathbf{u}_L^\top \mathbf{u}_L] + \frac{1}{n} \sigma_0^2 - \frac{1}{n} \mathbb{E}[\mathbf{u}_n^\top \mathbf{u}_n], \\
 0 &= \begin{pmatrix} \frac{2}{n} \mathbb{E}[\mathbf{u}_n^\top \mathbf{u}_L] & -\frac{1}{n} \mathbb{E}[\mathbf{u}_L^\top \mathbf{u}_L] & 1 \end{pmatrix} \begin{pmatrix} \lambda_0 \\ \lambda_0^2 \\ \sigma_0^2 \end{pmatrix} - \frac{1}{n} \mathbb{E}[\mathbf{u}_n^\top \mathbf{u}_n].
 \end{aligned} \tag{6.57}$$

Similarly, we derive two additional moment conditions:

$$0 = \begin{pmatrix} \frac{2}{n} \mathbb{E}[\mathbf{u}_{LL}^\top \mathbf{u}_L] & -\frac{1}{n} \mathbb{E}[\mathbf{u}_{LL}^\top \mathbf{u}_{LL}] & \frac{1}{n} \text{tr}(\mathbf{M}_n^\top \mathbf{M}_n) \end{pmatrix} \begin{pmatrix} \lambda_0 \\ \lambda_0^2 \\ \sigma_0^2 \end{pmatrix} - \frac{1}{n} \mathbb{E}[\mathbf{u}_L^\top \mathbf{u}_L], \tag{6.58}$$

$$0 = \begin{pmatrix} \frac{1}{n} \mathbb{E}[\mathbf{u}^\top \mathbf{u}_{LL} + \mathbf{u}_L^\top \mathbf{u}_L] & -\frac{1}{n} \mathbb{E}[\mathbf{u}_L^\top \mathbf{u}_{LL}] & 0 \end{pmatrix} \begin{pmatrix} \lambda_0 \\ \lambda_0^2 \\ \sigma_0^2 \end{pmatrix} - \frac{1}{n} \mathbb{E}[\mathbf{u}_n^\top \mathbf{u}_L]. \tag{6.59}$$

At this point it is important to realized that we have three equations an three unknowns: λ_0 , λ_0^2 and σ_0^2 . Consider the following three-equations system implied by Equations (6.57), (6.58) and (6.59):

$$\mathbf{\Gamma}_n \boldsymbol{\alpha}_0 = \boldsymbol{\gamma}_n, \tag{6.60}$$

where $\mathbf{\Gamma}_n$ is given in Equation (6.61), and $\boldsymbol{\alpha}_0 = (\lambda_0, \lambda_0^2, \sigma_0^2)$.⁸ If $\mathbf{\Gamma}_n$ where known, Assumption 6.21 (Identification) implies that Equation (6.60) determines $\boldsymbol{\alpha}_0$ as:

$$\boldsymbol{\alpha}_0 = \mathbf{\Gamma}_n^{-1} \boldsymbol{\gamma}_n,$$

where:

$$\mathbf{\Gamma}_n = \begin{pmatrix} \frac{2}{n} \mathbb{E}[\mathbf{u}_n^\top \mathbf{u}_L] & -\frac{1}{n} \mathbb{E}[\mathbf{u}_L^\top \mathbf{u}_L] & 1 \\ \frac{2}{n} \mathbb{E}[\mathbf{u}_{LL}^\top \mathbf{u}_L] & -\frac{1}{n} \mathbb{E}[\mathbf{u}_{LL}^\top \mathbf{u}_{LL}] & \frac{1}{n} \text{tr}(\mathbf{M}_n^\top \mathbf{M}_n) \\ \frac{1}{n} \mathbb{E}[\mathbf{u}_n^\top \mathbf{u}_{LL} + \mathbf{u}_L^\top \mathbf{u}_L] & -\frac{1}{n} \mathbb{E}[\mathbf{u}_L^\top \mathbf{u}_{LL}] & 0 \end{pmatrix}, \tag{6.61}$$

and

$$\boldsymbol{\gamma}_n = \begin{pmatrix} \frac{1}{n} \mathbb{E}[\mathbf{u}_n^\top \mathbf{u}_n] \\ \frac{1}{n} \mathbb{E}[\mathbf{u}_L^\top \mathbf{u}_L] \\ \frac{1}{n} \mathbb{E}[\mathbf{u}_n^\top \mathbf{u}_L] \end{pmatrix} .. \tag{6.62}$$

⁷Spatially lagged variables are denoted by bar superscripts in the articles. For clarity, we use the L subscript throughout. That is, a first order spatial lag of \mathbf{y}_n , $\mathbf{W}_n \mathbf{y}_n$, is denoted by \mathbf{y}_L . Higher order spatial lags are symbolized by adding additional L subscripts.

⁸Note that we are assuming that λ^2 is a new parameter.

Now, we express the moment conditions $\gamma_n = \mathbf{I}_n \alpha_n$ as sample averages in observables spatial lags of OLS residuals:

$$\mathbf{g}_n = \mathbf{G}_n \alpha_n + \mathbf{v}_n(\lambda_n, \sigma_n^2), \quad (6.63)$$

where $\mathbf{v}_n(\lambda_n, \sigma_n^2)$ can be viewed as a vector of residuals and \mathbf{G}_n is an 3×3 matrix given by

$$\mathbf{G}_n = \begin{pmatrix} \frac{2}{n} \hat{\mathbf{u}}_n^\top \hat{\mathbf{u}}_L & -\frac{1}{n} \hat{\mathbf{u}}_L^\top \hat{\mathbf{u}}_L & 1 \\ \frac{2}{n} \hat{\mathbf{u}}_{LL}^\top \hat{\mathbf{u}}_L & -\frac{1}{n} \hat{\mathbf{u}}_{LL}^\top \hat{\mathbf{u}}_{LL} & \frac{1}{n} \text{tr}(\mathbf{M}_n^\top \mathbf{M}_n) \\ \frac{1}{n} [\hat{\mathbf{u}}_n^\top \hat{\mathbf{u}}_{LL} + \hat{\mathbf{u}}_L^\top \hat{\mathbf{u}}_L] & -\frac{1}{n} \hat{\mathbf{u}}_L^\top \hat{\mathbf{u}}_{LL} & 0 \end{pmatrix}, \quad (6.64)$$

and

$$\mathbf{g}_n = \begin{pmatrix} \frac{1}{n} \hat{\mathbf{u}}_n^\top \hat{\mathbf{u}}_n \\ \frac{1}{n} \hat{\mathbf{u}}_L^\top \hat{\mathbf{u}}_L \\ \frac{1}{n} \hat{\mathbf{u}}_n^\top \hat{\mathbf{u}}_L \end{pmatrix}. \quad (6.65)$$

Equation (6.63) can be thought as an OLS regression where (Kelejian and Prucha, 1998):

$$\tilde{\alpha}_n = \mathbf{G}_n^{-1} \mathbf{g}_n. \quad (6.66)$$

However, the estimator in (6.66) is based on an overparameterization in the sense that it does not use the information that the second element of α_n , λ_n^2 , is the squared of the first element. To address this issue, Kelejian and Prucha (1998) and Kelejian and Prucha (1999) define the MOM estimator for λ and σ^2 as the nonlinear least square estimator corresponding to Equation (6.63):⁹

$$(\hat{\lambda}_{NLS,n}, \hat{\sigma}_{NLS,n}^2) = \text{argmin} \{ \mathbf{v}_n(\lambda_n, \sigma_n^2)^\top \mathbf{v}_n(\lambda_n, \sigma_n^2) : \lambda_n \in [-a, a], \sigma_n^2 \in [0, b] \} \quad (6.67)$$

Note that $(\hat{\lambda}_{NLS,n}, \hat{\sigma}_{NLS,n}^2)$ are defined as the minimizers of

$$Q_n(\lambda_n, \lambda_n^2, \sigma_n^2) = \left[\mathbf{g}_n - \mathbf{G}_n \begin{pmatrix} \lambda_n \\ \lambda_n^2 \\ \sigma_n^2 \end{pmatrix} \right]^\top \left[\mathbf{g}_n - \mathbf{G}_n \begin{pmatrix} \lambda_n \\ \lambda_n^2 \\ \sigma_n^2 \end{pmatrix} \right]. \quad (6.68)$$

We provide the following assumptions.

Assumption 6.19 — Bounded Matrices (Kelejian and Prucha, 1999). The row and column sums of the matrices \mathbf{M}_n and $(\mathbf{I}_n - \lambda_0 \mathbf{M}_n)$ are bounded uniformly in absolute value.

Assumption 6.20 — Residuals (Kelejian and Prucha, 2010). 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

$$\sqrt{n} \|\boldsymbol{\Delta}_n\| = O_p(1). \quad (6.69)$$

⁹They argue that this estimator is more efficient than the OLS estimator. However, both estimator are consistent. See Theorem 2 in (Kelejian and Prucha, 1998).

The vector $\mathbf{d}_{i,n}$ denotes the i -th row of the regressor matrix, and $\mathbf{\Delta}_n$ denotes the difference between the parameter estimator and the true parameter values. This assumption should be satisfied for most cases in which $\tilde{\mathbf{u}}_n$ is based on \sqrt{n} -consistent estimators of the regression coefficients (non-linear OLS, linear OLS, S2SLS). Assumption 6.20 comes from Kelejian and Prucha (2010) and is a bit stronger than the corresponding assumption in Kelejian and Prucha (1999).

Assumption 6.21 — Identification (Kelejian and Prucha, 1999). Let $\mathbf{\Gamma}_n$ be the matrix in Equation (6.61). The smallest eigenvalues of $\mathbf{\Gamma}_n^\top \mathbf{\Gamma}_n$ is bounded away from zero, that is, $\omega_{\min}(\mathbf{\Gamma}_n^\top \mathbf{\Gamma}_n) \geq \omega_* > 0$, where ω_* may depend on λ and σ^2

The following Theorem establishes the consistency of the NLS estimator.

Theorem 6.22 — Consistency. Let $(\hat{\lambda}_{NLS,n}, \hat{\sigma}_{NLS,n}^2)$ given by:

$$(\hat{\lambda}_{NLS,n}, \hat{\sigma}_{NLS,n}^2) = \operatorname{argmin} \{ \mathbf{v}_n(\lambda, \sigma^2)^\top \mathbf{v}_n(\lambda, \sigma^2) : \lambda \in [-a, a], \sigma^2 \in [0, b] \}$$

Then, given Assumptions 6.1 (Heteroskedastic errors), 6.18 (Weight Matrix \mathbf{M}_n), 6.19 (Bounded Matrices), 6.20 (Residuals), and 6.21 (Identification),

$$(\hat{\lambda}_{NLS,n}, \hat{\sigma}_{NLS,n}^2) \xrightarrow{p} (\lambda_0, \sigma_0^2) \quad \text{as } n \rightarrow \infty$$

An important remark is that Theorem 6.22 establishes the consistency of the NLS estimates but does not provide information about the asymptotic distribution of $\hat{\lambda}_{NLS,n}$.

Sketch or proof for NLS estimator of $\hat{\lambda}_n$. The proof is based on Kelejian and Piras (2017) and consist into two steps. First, we prove consistency of $\hat{\lambda}_n$ for the OLS estimate of λ_0 —which is more simple—and assuming that the vector \mathbf{u}_n is observed. We then show that \mathbf{u}_n can be replaced in the GM estimator for λ by $\hat{\mathbf{u}}_n$. For a more general proof see Kelejian and Prucha (1998, 1999).

- (a) *Assuming that \mathbf{u}_n is observed.* Recall that in Equation (6.63) the sample moments are based on the estimated $\hat{\mathbf{u}}_n$. But, if \mathbf{u}_n were observed, then we would use the following sample moments:

$$\mathbf{g}_n^* = \mathbf{G}_n^* \boldsymbol{\alpha}_0$$

where

$$\mathbf{G}_n^* = \begin{pmatrix} \frac{2}{n} \mathbf{u}_n^\top \mathbf{u}_L & -\frac{1}{n} \mathbf{u}_L^\top \mathbf{u}_L & 1 \\ \frac{2}{n} \mathbf{u}_{LL}^\top \mathbf{u}_L & -\frac{1}{n} \mathbf{u}_{LL}^\top \mathbf{u}_{LL} & \frac{1}{n} \operatorname{tr}(\mathbf{M}_n^\top \mathbf{M}_n) \\ \frac{1}{n} [\mathbf{u}^\top \mathbf{u}_{LL} + \mathbf{u}_L^\top \mathbf{u}_L] & -\frac{1}{n} \mathbf{u}_L^\top \mathbf{u}_{LL} & 0 \end{pmatrix}, \quad \mathbf{g}_n^* = \begin{pmatrix} \frac{1}{n} \mathbf{u}_n^\top \mathbf{u}_n \\ \frac{1}{n} \mathbf{u}_L^\top \mathbf{u}_L \\ \frac{1}{n} \mathbf{u}_n^\top \mathbf{u}_L \end{pmatrix}$$

Recall that:

$$\begin{aligned} \mathbf{u}_n &= (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^{-1} \boldsymbol{\varepsilon}_n \\ \mathbf{u}_L &= \mathbf{M}_n (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^{-1} \boldsymbol{\varepsilon}_n \\ \mathbf{u}_{LL} &= \mathbf{M}_n \mathbf{M}_n (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^{-1} \boldsymbol{\varepsilon}_n \end{aligned}$$

and first and second column of \mathbf{G}_n^* are quadratic forms of $\boldsymbol{\varepsilon}_n$. Since \mathbf{M}_n is uniformly bounded then, using Lemma of consistency of quadratic forms 3.27, we can state that:

$$\mathbf{G}_n^* \xrightarrow{p} \mathbf{\Gamma}_n.$$

Also:

$$\begin{aligned}\text{plim } \mathbf{g}_n^* &= \text{plim } \mathbf{G}_n^* \boldsymbol{\alpha}_0, \\ &= \boldsymbol{\Gamma}_n \boldsymbol{\alpha}_0.\end{aligned}$$

If \mathbf{u}_n would be observed, a linear GMM estimator for λ_0 , say $\tilde{\lambda}$, would be the first element of the least squared estimator $\tilde{\boldsymbol{\alpha}}_n$, namely:

$$\tilde{\boldsymbol{\alpha}}_n = \mathbf{G}_n^{-1*} \mathbf{g}_n^*,$$

since \mathbf{G}_n^* is a 3×3 matrix which is nonsingular. Thus, using our previous results:

$$\text{plim } \tilde{\boldsymbol{\alpha}}_n = \text{plim } \mathbf{G}_n^{-1*} \text{plim } \mathbf{g}_n^* = \boldsymbol{\Gamma}_n^{-1} \boldsymbol{\gamma}_n = \boldsymbol{\alpha}_0. \quad (6.70)$$

- (b) *Replacing \mathbf{u}_n by $\hat{\mathbf{u}}_n$.* Now consider the estimator of $\boldsymbol{\alpha}_0$ based on $\hat{\mathbf{u}}_n$. The OLS estimator is consistent and can be expressed as:

$$\tilde{\boldsymbol{\beta}}_n = \boldsymbol{\beta}_0 + \Delta_n, \quad \Delta_n \xrightarrow{p} \mathbf{0}.$$

Then, the OLS estimator $\hat{\mathbf{u}}_n$ is:¹⁰

$$\begin{aligned}\hat{\mathbf{u}}_n &= \mathbf{y}_n - \mathbf{X}_n \hat{\boldsymbol{\beta}}_n, \\ &= \mathbf{y}_n - \mathbf{X}_n (\boldsymbol{\beta}_0 + \Delta_n), \\ &= \mathbf{y}_n - \mathbf{X}_n \boldsymbol{\beta}_0 - \mathbf{X}_n \Delta_n, \\ &= \mathbf{u}_n - \mathbf{X}_n \Delta_n.\end{aligned}$$

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 terms of the form $\boldsymbol{\varepsilon}_n^\top \mathbf{S}_n \boldsymbol{\varepsilon}_n / n$, where \mathbf{S}_n is an $n \times n$ matrix whose row and columns are uniformly bounded in absolute value given our assumption 6.19. For example:

$$\frac{1}{n} \mathbf{u}_n^\top \mathbf{u}_n = \frac{1}{n} \boldsymbol{\varepsilon}_n^\top (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^{-1\top} \mathbf{M}_n (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^{-1} \boldsymbol{\varepsilon}_n = \frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{S}_n \boldsymbol{\varepsilon}_n.$$

Then:

$$\begin{aligned}\frac{\hat{\mathbf{u}}_n^\top \mathbf{S}_n \hat{\mathbf{u}}_n}{n} &= \frac{(\mathbf{u}_n - \mathbf{X}_n \Delta_n)^\top \mathbf{S}_n (\mathbf{u}_n - \mathbf{X}_n \Delta_n)}{n}, \\ &= \frac{\mathbf{u}_n^\top \mathbf{S}_n \mathbf{u}_n}{n} - \frac{2\Delta_n^\top \mathbf{X}_n^\top \mathbf{S}_n \mathbf{u}_n}{n} + \frac{\Delta_n^\top \mathbf{X}_n^\top \mathbf{S}_n \mathbf{X}_n \Delta_n}{n}.\end{aligned}$$

We need to show that:

$$\begin{aligned}\frac{2\Delta_n^\top \mathbf{X}_n^\top \mathbf{S}_n \mathbf{u}_n}{n} &\xrightarrow{p} \mathbf{0}, \\ \frac{\Delta_n^\top \mathbf{X}_n^\top \mathbf{S}_n \mathbf{X}_n \Delta_n}{n} &\xrightarrow{p} \mathbf{0},\end{aligned}$$

so that we can conclude that:

$$\frac{\hat{\mathbf{u}}_n^\top \mathbf{S}_n \hat{\mathbf{u}}_n}{n} \xrightarrow{p} \frac{1}{n} \mathbf{u}_n^\top \mathbf{S}_n \mathbf{u}_n,$$

¹⁰For a more formal proof with any consistent estimate see Lemma C.1 in Kelejian and Prucha (2010).

and finally say that:

$$\mathbf{g}_n \xrightarrow{p} \mathbf{g}_n^* \xrightarrow{p} \boldsymbol{\gamma}_n, \quad \mathbf{G}_n \xrightarrow{p} \mathbf{G}_n^* \xrightarrow{p} \boldsymbol{\Gamma}_n.$$

Given Equation (6.70), consistency is proved. ■

We can also derive the MOM estimation using a GMM approach without weighting the moments:

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} Q_n(\boldsymbol{\theta}) = \mathbf{g}_n(\boldsymbol{\theta})^\top \mathbf{g}_n(\boldsymbol{\theta}).$$

Using the moment conditions under homoskedasticity results in the following empirical moments:

$$\mathbf{g}_n(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{\varepsilon}_n^\top \boldsymbol{\varepsilon}_n - n\hat{\sigma}^2 \\ \boldsymbol{\varepsilon}_n^\top \mathbf{M}_n^\top \mathbf{M}_n \boldsymbol{\varepsilon}_n - \hat{\sigma}^2 \text{tr}(\mathbf{M}_n^\top \mathbf{M}_n) \\ \boldsymbol{\varepsilon}_n^\top \mathbf{M}_n \boldsymbol{\varepsilon}_n \end{pmatrix} = \begin{pmatrix} \mathbf{R}_n^\top \mathbf{u}_n^\top \mathbf{R}_n \mathbf{u}_n - n\hat{\sigma}^2 \\ \mathbf{R}_n^\top \mathbf{u}_n^\top \mathbf{M}_n^\top \mathbf{M}_n \mathbf{R}_n \mathbf{u}_n - \hat{\sigma}^2 \text{tr}(\mathbf{M}_n^\top \mathbf{M}_n) \\ \mathbf{R}_n^\top \mathbf{u}_n^\top \mathbf{M}_n \mathbf{R}_n \mathbf{u}_n \end{pmatrix}$$

where \mathbf{u}_n are replaced by least squares residuals.

6.4.3 Feasible Generalized Least Squares Estimator

In Section 6.4.1 we established that the GLS estimator is given by:

$$\boldsymbol{\beta}_{GLS}(\lambda_0) = [\mathbf{X}_n^\top \boldsymbol{\Omega}(\lambda_0)^{-1} \mathbf{X}_n]^{-1} \mathbf{X}_n^\top \boldsymbol{\Omega}(\lambda_0)^{-1} \mathbf{y}_n, \quad (6.71)$$

where $\boldsymbol{\Omega}(\lambda_0) = (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^{-1} (\mathbf{I}_n - \lambda_0 \mathbf{M}_n^\top)^{-1}$. The MOM procedure provides a consistent estimate for λ_0 , which can be used to obtain the FGLS estimator:

$$\boldsymbol{\beta}_{FGLS}(\hat{\lambda}_n) = [\mathbf{X}_n^\top \boldsymbol{\Omega}(\hat{\lambda}_n)^{-1} \mathbf{X}_n]^{-1} \mathbf{X}_n^\top \boldsymbol{\Omega}(\hat{\lambda}_n)^{-1} \mathbf{y}_n. \quad (6.72)$$

The following theorem provides the identification conditions for the Spatial FGLS estimator.

Assumption 6.23 — Limiting Behavior. The elements of \mathbf{X}_n are non-stochastic and bounded in absolute value by c_X , $0 < c_X < \infty$. Also, \mathbf{X}_n has full rank, and the matrix $\mathbf{Q}_X = \lim_{n \rightarrow \infty} n^{-1} \mathbf{X}_n^\top \mathbf{X}_n$ is finite and nonsingular. Furthermore, the matrices

$$\mathbf{Q}_X(\lambda_0) = \lim_{n \rightarrow \infty} n^{-1} \mathbf{X}_n^\top \boldsymbol{\Omega}(\lambda_0)^{-1} \mathbf{X}_n$$

are finite and nonsingular for all $|\lambda_0| < 1$

The following Theorem proposes the asymptotic distribution for the SFGLS Estimator:

Theorem 6.24 — Asymptotic Properties of FGLS Estimator. If assumptions 6.1 (Homoskedastic errors), 6.18 (Weight Matrix \mathbf{M}_n), 6.19 (Bounded Matrices), and 6.23 (Limiting Behavior) hold:

(a) The true GLS estimator $\hat{\boldsymbol{\beta}}_{GLS}$ is a consistent estimator for $\boldsymbol{\beta}_0$, and

$$\sqrt{n} \left(\hat{\boldsymbol{\beta}}_{GLS} - \boldsymbol{\beta}_0 \right) \xrightarrow{d} N(\mathbf{0}, \sigma^2 \mathbf{Q}_X(\lambda_0)^{-1})$$

- (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 $\hat{\sigma}_n^2 \left[n^{-1} \mathbf{X}_n^\top \boldsymbol{\Omega}(\hat{\lambda}_n)^{-1} \mathbf{X}_n \right]$ is a consistent estimator for $\sigma^2 \mathbf{Q}_X(\lambda_0)^{-1}$.

Note that Theorem 6.24 assumes the existence of a consistent estimator of λ_0 and σ_0^2 . It can be shown that the OLS estimator:

$$\hat{\beta}_n = (\mathbf{X}_n^\top \mathbf{X}_n)^{-1} \mathbf{X}_n^\top \mathbf{y}_n,$$

is \sqrt{n} -consistent. Thus, the OLS residuals $\tilde{u}_i = y_i - \mathbf{x}_i^\top \hat{\beta}_n$ satisfy Assumption 6.20 with $d_{i,n} = |\mathbf{x}_i|$ and $\Delta_n = \hat{\beta}_n - \beta_0$. Thus, OLS residuals can be used to obtain consistent estimators of λ_0 and σ_0^2 .

Then, the SFGLS estimator is given by

$$\hat{\beta}_{FGLS} = \left[\mathbf{X}^\top(\tilde{\lambda}) \mathbf{X}(\tilde{\lambda}) \right]^{-1} \mathbf{X}^\top(\tilde{\lambda}) \mathbf{y}(\tilde{\lambda}),$$

where:

$$\mathbf{X}(\tilde{\lambda}) = (\mathbf{I}_n - \tilde{\lambda}_n \mathbf{M}_n) \mathbf{X}_n, \quad \mathbf{y}(\tilde{\lambda}) = (\mathbf{I}_n - \tilde{\lambda}_n \mathbf{M}_n) \mathbf{y}_n.$$

The variance covariance matrix of $\hat{\beta}_{FGLS}$ is estimated as:

$$\hat{\mathbb{V}}(\hat{\beta}_{FGLS}) = \hat{\sigma}_n^2 \left[\mathbf{X}^\top(\tilde{\lambda}) \mathbf{X}(\tilde{\lambda}) \right]^{-1},$$

where:

$$\begin{aligned} \hat{\sigma}_n^2 &= \hat{\boldsymbol{\varepsilon}}^\top(\tilde{\lambda}) \hat{\boldsymbol{\varepsilon}}(\tilde{\lambda}), \\ \hat{\boldsymbol{\varepsilon}}(\tilde{\lambda}) &= \mathbf{y}(\tilde{\lambda}) - \mathbf{X}(\tilde{\lambda}) \hat{\beta}_{FGLS} = (\mathbf{I} - \tilde{\lambda} \mathbf{M}) \hat{\mathbf{u}}_n, \\ \hat{\mathbf{u}}_n &= \mathbf{y}_n - \mathbf{X}_n \hat{\beta}_{FGLS}. \end{aligned}$$

Sketch of Proof of Theorem 6.24. We first prove part (a). Recall that the GLS and FGSL estimator are given by:

$$\begin{aligned} \hat{\beta}_{GLS} &= [\mathbf{X}_n^\top \boldsymbol{\Omega}(\lambda_0)^{-1} \mathbf{X}_n]^{-1} \mathbf{X}_n^\top \boldsymbol{\Omega}(\lambda_0)^{-1} \mathbf{y}_n, \\ \hat{\beta}_{FGLS} &= [\mathbf{X}_n^\top \hat{\boldsymbol{\Omega}}(\lambda)^{-1} \mathbf{X}_n]^{-1} \mathbf{X}_n^\top \hat{\boldsymbol{\Omega}}(\lambda)^{-1} \mathbf{y}_n. \end{aligned}$$

Since $\mathbf{y}_n = \mathbf{X}_n \beta_0 + \mathbf{u}_n = \mathbf{X}_n \beta_0 + (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^{-1} \boldsymbol{\varepsilon}_n$, the sampling error of $\hat{\beta}_{GLS}$ is,

$$\begin{aligned} \hat{\beta}_n &= \beta_0 + [\mathbf{X}_n^\top \boldsymbol{\Omega}(\lambda_0)^{-1} \mathbf{X}_n]^{-1} \mathbf{X}_n^\top \boldsymbol{\Omega}(\lambda_0)^{-1} \mathbf{u}_n, \\ \hat{\beta}_n - \beta_0 &= [\mathbf{X}_n^\top \boldsymbol{\Omega}(\lambda_0)^{-1} \mathbf{X}_n]^{-1} \mathbf{X}_n^\top (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^\top (\mathbf{I}_n - \lambda_0 \mathbf{M}_0) (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^{-1} \boldsymbol{\varepsilon}_n, \\ \hat{\beta}_n - \beta_0 &= [\mathbf{X}_n^\top \boldsymbol{\Omega}(\lambda_0)^{-1} \mathbf{X}_n]^{-1} \mathbf{X}_n^\top (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^\top \boldsymbol{\varepsilon}_n, \\ \sqrt{n}(\hat{\beta}_n - \beta_0) &= \left[\frac{1}{n} \mathbf{X}_n^\top \boldsymbol{\Omega}(\lambda_0)^{-1} \mathbf{X}_n \right]^{-1} \frac{1}{\sqrt{n}} \mathbf{A}_0^\top \boldsymbol{\varepsilon}_n, \end{aligned}$$

where $\mathbf{A}_0 = (\mathbf{I}_n - \lambda_0 \mathbf{M}_n) \mathbf{X}_n$. By Assumption 6.23 (Limiting Behavior):

$$\frac{1}{n} \mathbf{X}_n^\top \boldsymbol{\Omega}(\lambda_0)^{-1} \mathbf{X}_n \rightarrow \mathbf{Q}_X(\lambda_0).$$

Since \mathbf{Q}_X is not singular, by continuous mapping theorem:

$$\left[\frac{1}{n} \mathbf{X}_n^\top \boldsymbol{\Omega}(\lambda_0)^{-1} \mathbf{X}_n \right]^{-1} \rightarrow \mathbf{Q}_X^{-1}(\lambda_0).$$

Because \mathbf{A}_0 is bounded in absolute value (why?), by Theorem 3.30 it follows that:

$$\frac{1}{\sqrt{n}} \mathbf{A}_0^\top \boldsymbol{\varepsilon}_n \xrightarrow{d} \mathcal{N} \left(\mathbf{0}, \lim_{n \rightarrow \infty} n^{-1} \sigma^2 \mathbf{A}_0^\top \mathbf{A}_0 \right),$$

where $\lim_{n \rightarrow \infty} n^{-1} \sigma^2 \mathbf{A}_0^\top \mathbf{A}_0 = \sigma^2 \lim_{n \rightarrow \infty} n^{-1} \mathbf{X}_n^\top (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^\top (\mathbf{I}_n - \lambda_0 \mathbf{M}_n) \mathbf{X}_n = \sigma^2 \mathbf{Q}_X(\lambda_0)$.
Consequently:

$$\begin{aligned} \sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) &= \underbrace{\left[\frac{1}{n} \mathbf{X}_n^\top \boldsymbol{\Omega}(\lambda_0)^{-1} \mathbf{X}_n \right]^{-1}}_{\rightarrow \mathbf{Q}_X^{-1}(\lambda_0)} \underbrace{\frac{1}{\sqrt{n}} \mathbf{A}_0^\top \boldsymbol{\varepsilon}}_{\xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{Q}_X(\lambda_0))} \\ &\xrightarrow{d} \mathcal{N} \left[\mathbf{0}, \mathbf{Q}_X^{-1}(\lambda_0) \sigma^2 \mathbf{Q}_X(\lambda_0) \mathbf{Q}_X^{-1}(\lambda_0)^\top \right], \\ &\xrightarrow{d} \mathcal{N} \left[\mathbf{0}, \sigma^2 \mathbf{Q}_X^{-1}(\lambda_0) \right]. \end{aligned}$$

This also implies that $\hat{\boldsymbol{\beta}}_{GLS}$ is consistent (why?). To show part (b), we can show that:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{GLS} - \boldsymbol{\beta}_{FGLS}) \xrightarrow{p} \mathbf{0}.$$

Recall that $\text{plim}(X_n - Y_n) = 0$ implies that the random variables X_n and Y_n have the same asymptotic distribution. Following Kelejian and Prucha (1999), it suffices to show that

$$\frac{1}{n} \mathbf{X}_n^\top \left[\boldsymbol{\Omega}(\hat{\lambda}_n)^{-1} - \boldsymbol{\Omega}(\lambda_0)^{-1} \right] \mathbf{X}_n \xrightarrow{p} \mathbf{0} \quad (6.73)$$

and

$$\frac{1}{\sqrt{n}} \mathbf{X}_n^\top \left[\boldsymbol{\Omega}(\hat{\lambda}_n)^{-1} - \boldsymbol{\Omega}(\lambda_0)^{-1} \right] \mathbf{u}_n \xrightarrow{p} \mathbf{0}.$$

To derive the expression for $\boldsymbol{\Omega}(\hat{\lambda}_n)^{-1} - \boldsymbol{\Omega}(\lambda_0)^{-1}$, we start by expanding both matrices in terms of λ_0 and $\hat{\lambda}_n$.

Given that:

$$\boldsymbol{\Omega}(\lambda_0)^{-1} = (\mathbf{I}_n - \lambda_0 \mathbf{M}_n^\top)(\mathbf{I}_n - \lambda_0 \mathbf{M}_n),$$

and

$$\boldsymbol{\Omega}(\hat{\lambda}_n)^{-1} = (\mathbf{I}_n - \hat{\lambda}_n \mathbf{M}_n^\top)(\mathbf{I}_n - \hat{\lambda}_n \mathbf{M}_n),$$

we subtract these two expressions to obtain:

$$\boldsymbol{\Omega}(\hat{\lambda}_n)^{-1} - \boldsymbol{\Omega}(\lambda_0)^{-1} = \left[(\mathbf{I}_n - \hat{\lambda}_n \mathbf{M}_n^\top)(\mathbf{I}_n - \hat{\lambda}_n \mathbf{M}_n) - (\mathbf{I}_n - \lambda_0 \mathbf{M}_n^\top)(\mathbf{I}_n - \lambda_0 \mathbf{M}_n) \right].$$

Expanding yields:

$$(\mathbf{I}_n - \lambda \mathbf{M}_n^\top)(\mathbf{I}_n - \lambda \mathbf{M}_n) = \mathbf{I}_n - \lambda \mathbf{M}_n^\top - \lambda \mathbf{M}_n + \lambda^2 \mathbf{M}_n^\top \mathbf{M}_n,$$

By subtracting these expansions, we obtain the first-order difference:

$$\begin{aligned} \boldsymbol{\Omega}(\hat{\lambda}_n)^{-1} - \boldsymbol{\Omega}(\lambda)^{-1} &= \mathbf{I}_n - \lambda \mathbf{M}_n^\top - \lambda \mathbf{M}_n + \lambda^2 \mathbf{M}_n^\top \mathbf{M}_n \\ &\quad - (\mathbf{I}_n - \lambda_0 \mathbf{M}_n^\top - \lambda_0 \mathbf{M}_n + \lambda_0^2 \mathbf{M}_n^\top \mathbf{M}_n), \\ &= (\lambda_0 - \hat{\lambda}_n)(\mathbf{M} + \mathbf{M}^\top) + (\lambda_0^2 - \hat{\lambda}_n^2) \mathbf{M}^\top \mathbf{M}. \end{aligned}$$

Then using the fact that we have summable matrices,

$$\frac{1}{n} \mathbf{X}_n^\top \left[\boldsymbol{\Omega}(\hat{\lambda}_n)^{-1} - \boldsymbol{\Omega}(\lambda_0)^{-1} \right] \mathbf{X}_n = \underbrace{(\lambda_0 - \hat{\lambda}_n)}_{\xrightarrow{p} 0} \underbrace{n^{-1} \mathbf{X}_n^\top (\mathbf{M}_n + \mathbf{M}_n^\top) \mathbf{X}_n}_{O(1)} + \underbrace{(\lambda_0^2 - \hat{\lambda}_n^2)}_{\xrightarrow{p} 0} \underbrace{n^{-1} \mathbf{X}_n^\top \mathbf{M}_n^\top \mathbf{M}_n \mathbf{X}_n}_{O(1)},$$

where $(\lambda - \hat{\lambda}_n) = o_p(1)$ since $\hat{\lambda}_n$ is a consistent estimate of λ , and:

$$\begin{aligned} \frac{1}{\sqrt{n}} \mathbf{X}_n^\top \left[\boldsymbol{\Omega}(\hat{\lambda}_n)^{-1} - \boldsymbol{\Omega}(\lambda_0)^{-1} \right] \mathbf{u}_n &= \underbrace{(\lambda_0 - \hat{\lambda}_n)}_{\xrightarrow{p} 0} \underbrace{n^{-1/2} \mathbf{X}_n^\top (\mathbf{M}_n + \mathbf{M}_n^\top) \mathbf{u}_n}_{O_p(1)} + \underbrace{(\lambda_0^2 - \hat{\lambda}_n^2)}_{\xrightarrow{p} 0} \underbrace{n^{-1/2} \mathbf{X}_n^\top \mathbf{M}_n^\top \mathbf{M}_n \mathbf{u}_n}_{O_p(1)}, \\ &= o_p(1) * O_p(1) + o_p(1) * O_p(1), \\ &= o_p(1) + o_p(1), \\ &= o_p(1), \\ &\xrightarrow{p} 0. \end{aligned} \tag{6.74}$$

To see that $n^{-1/2} \mathbf{X}_n^\top (\mathbf{M}_n + \mathbf{M}_n^\top) \mathbf{u}_n = O_p(1)$ note that

$$\begin{aligned} \mathbb{E} [n^{-1/2} \mathbf{X}_n^\top (\mathbf{M}_n + \mathbf{M}_n^\top) \mathbf{u}_n] &= 0, \\ \mathbb{V} [n^{-1/2} \mathbf{X}_n^\top (\mathbf{M}_n + \mathbf{M}_n^\top) \mathbf{u}_n] &= n^{-1} \mathbf{X}^\top \underbrace{(\mathbf{M}_n + \mathbf{M}_n^\top) \boldsymbol{\Omega}_0 (\mathbf{M}_n^\top + \mathbf{M}_n)}_{\substack{\text{absolutely summable} \\ O(n)}} \mathbf{X} = O(1). \end{aligned}$$

A similar result holds for $n^{-1/2} \mathbf{X}_n^\top \mathbf{M}_n^\top \mathbf{M}_n \mathbf{u}_n$.

Part 3 of the theorem follows from (6.73) 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 6.25 — GLS (FGLS) Algorithm of SEM. The steps are the following:

- (a) First of all obtain a consistent estimate of $\boldsymbol{\beta}$, say $\tilde{\boldsymbol{\beta}}$ 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 (6.67),
- (d) Estimate $\boldsymbol{\beta}$ using Equation (6.72)

6.4.4 Coding the FSGLS estimator in R

In this Section, we show how to create a function for the FSGLS estimator of the SEM outlined in Algorithm 6.25. First, we generate a function that create the matrix \mathbf{G}_n and the moments vector \mathbf{g}_n given in Equation (6.64) and (6.65), respectively.

```
# Function that generates g and G
mom.sem <- function(u, M){
  # This function generates the moment conditions
```

```

# inputs: Consistent residuals and W matrix
n      <- length(u)
u_l    <- W %*% u
u_ll   <- W %*% u_l
trMM   <- sum(diag(crossprod(M)))
uu     <- crossprod(u)
uul    <- crossprod(u, u_l)
uull   <- crossprod(u, u_ll)
ullul  <- crossprod(u_ll, u_l)
ullull <- crossprod(u_ll, u_ll)
ulul   <- crossprod(u_l, u_l)
ulull  <- crossprod(u_l, u_ll)
G      <- matrix(0, 3, 3)
G[1, 1] <- 2 * uul
G[2, 1] <- 2 * ullul
G[3, 1] <- uull + ulul
G[1, 2] <- -ulul
G[2, 2] <- -ullull
G[3, 2] <- -ulull
G[1, 3] <- 1 * n
G[2, 3] <- trMM
G <- G / n
g <- c(uu, ulul, uul) / n
list(G = G, g = g)
}

```

In the following lines, we provide a function that returns the $Q_n = \mathbf{v}_n^\top \mathbf{v}_n$ defined in Equation (6.68).

```

# Function to be optimized
Qn <- function(par, mom, verbose = verbose){
  # par has lambda and sigma
  upsilon <- mom$g - mom$G %*% c(par[1], par[1]^2, par[2])
  upup <- crossprod(upsilon)
  if (verbose)
    cat("function:", upup, "lambda:", par[1], "sig2:",
        par[2], "\n")
  return(upup)
}

```

Next, we create the main function that generates the FSGLS estimator of the SEM model:

```

# Main function for the FSGLSE
sem.sfgls <- function(formula, data, M, verbose = TRUE){
  # Model Frame
  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())

# Get variables and globals
y <- model.response(mf)
X <- model.matrix(formula, mf)
n <- nrow(X)
k <- ncol(X)
sn <- nrow(M)
if (n != sn) stop("number of spatial units in W is different to the number of data")

# First Step: Obtain consistent residuals from OLS
ols <- lm(y ~ X - 1)
u.hat <- residuals(ols)

# Generate Moments
mom.hat <- mom.sem(u = u.hat, M = M)

# Initial values for lambda and sigma
scorr <- crossprod(W %*% u.hat, u.hat) / crossprod(u.hat)
par <- c(scorr, var(u.hat))

# Optimization
opt <- nlminb(start = par, Qn, mom = mom.hat, verbose = verbose)

# B FGLS
lambda.hat <- opt$par[1L]
ys <- y - drop(lambda.hat) * W %*% y
Xs <- X - drop(lambda.hat) * W %*% X
b.hat <- solve(crossprod(Xs)) %*% crossprod(Xs, ys)

# Residuals
e.hat <- ys - Xs %*% b.hat

# Save results
results <- structure(
  list(
    coefficients = c(b.hat, lambda.hat),
    call        = callT,
    X           = X,
    y           = y,
    Xs          = Xs,
    e.hat       = e.hat
  )
)

```

```

    ),
    class = 'myfs2sls'
  )
}

```

We generate the following DGP to test our function:

```

# Generate DGP
library("spatialreg")
library("spdep")
set.seed(1)
n      <- 529
lambda <- 0.6
M.nb2  <- cell2nb(sqrt(n), sqrt(n))
M      <- nb2mat(M.nb2)

# Exogenous variables
x1      <- rnorm(n)
x2      <- rnorm(n)
x3      <- rnorm(n)

# DGP parameters
b0 <- 0 ; b1 <- -1; b2 <- 0; b3 <- 1
sigma2 <- 2
epsilon <- rnorm(n, mean = 0, sd = sqrt(sigma2))

# Simulate the dependent variable
y <- b0 + b1*x1 + b2*x2 + b3*x3 + solve(diag(n) - lambda * M) %*% epsilon

data <- as.data.frame(cbind(y, x1, x2, x3))
names(data) <- c("y", "x1", "x2", "x3")

```

Generate S3 methods:

```

vcov.myfs2sls <- function(object, ...){
  sigma2 <- crossprod(object$e.hat) / n
  var    <- drop(sigma2) * solve(crossprod(object$Xs))
  return(var)
}
summary.myfs2sls <- function(object,
                             table = TRUE,
                             digits = max(3, .Options$digits - 3),
                             ...){
  n      <- nrow(object$X)
  K      <- ncol(object$X)
  df     <- n - K
  b      <- object$coefficients[1:K]

```

```

std.err <- sqrt(diag(vcov(object)))
z       <- b / std.err
p       <- 2 * pt(-abs(z), df = df)
CoefTable <- cbind(b, std.err, z, p)
colnames(CoefTable) <- c("Estimate", "Std.Error", "t-value", "Pr(>|t|)")
result <- structure(
  list(
    CoefTable = CoefTable,
    digits    = digits,
    call      = object$call),
  class = 'summary.myfs2sls'
)
return(result)
}

print.summary.myfs2sls <- function(x,
                                digits = x$digits,
                                na.print = "",
                                symbolic.cor = p > 4,
                                signif.stars = getOption("show.signif.stars"),
                                ...)
{
  cat("\nCall:\n")
  cat(paste(deparse(x$call), sep = "\n", collapse = "\n"), "\n\n", sep = "")

  cat("\nCoefficients:\n")
  printCoefmat(x$CoefTable, digit = digits, P.value = TRUE, has.Pvalue = TRUE)
  invisible(NULL)
}

```

We test our function and compare it with `GMerrorsar`

```

b.fs2sls <- sem.sfgls(y ~ x1 + x2 + x3, data = data, M = M, verbose = FALSE)
summary(b.fs2sls)

##
## Call:
## sem.sfgls(formula = y ~ x1 + x2 + x3, data = data, M = M, verbose = FALSE)
##
##
## Coefficients:
##           Estimate Std.Error t-value Pr(>|t|)
## (Intercept) -0.192429  0.149854  -1.284    0.200
## x1          -1.079338  0.063535 -16.988 <2e-16 ***
## x2           0.006391  0.060054   0.106   0.915
## x3           0.975246  0.064009  15.236 <2e-16 ***
## ---

```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Check

```
sem_mm      <- GMerrorsar(y ~ x1 + x2 + x3,
                          data = data,
                          listw = mat2listw(M, style = "W"),
                          verbose = FALSE,
                          legacy = TRUE)

summary(sem_mm)

##
## Call:GMerrorsar(formula = y ~ x1 + x2 + x3, data = data, listw = mat2listw(M,
##      style = "W"), verbose = FALSE, legacy = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.748647 -0.946972 -0.050107  1.076211  4.413975
##
## Type: GM SAR estimator
## Coefficients: (GM standard errors)
##              Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -0.192429   0.149854  -1.2841  0.1991
## x1          -1.079338   0.063535 -16.9880 <2e-16
## x2           0.006391   0.060054   0.1064  0.9152
## x3           0.975246   0.064009  15.2362 <2e-16
##
## Lambda: 0.55602 (standard error): 0.10098 (z-value): 5.5061
## Residual variance (sigma squared): 2.3414, (sigma: 1.5302)
## GM argmin sigma squared: 2.3535
## Number of observations: 529
## Number of parameters estimated: 6
```

6.5 GMM Estimator for SEM

6.5.1 GMM Estimator Under Homoskedasticity

In this section, we provide the GMM estimator for the SEM model under homoskedasticity following [Lee and Liu \(2010\)](#) and [Liu et al. \(2010\)](#). It is important to note that both articles derive the OGMME for the SAC model, but [Lee and Liu \(2010\)](#) focuses on SAC model with higher orders.

Recall that the SEM model is given by:

$$\begin{aligned} \mathbf{y}_n &= \mathbf{X}\boldsymbol{\beta}_0 + \mathbf{u}_n \\ \mathbf{u}_n &= \lambda_0 \mathbf{M}_n \mathbf{u}_n + \boldsymbol{\varepsilon}_n, \end{aligned}$$

where $\mathbf{P}_{jn}, j = 1, 2$ are from \mathcal{P}_{1n} , and $\epsilon_{ni}, i = 1, \dots, n$, of $\boldsymbol{\varepsilon}_n$ are i.i.d. $(0, \sigma_0^2)$ with zero mean and variance σ_0^2 . For this model, denote $\boldsymbol{\varepsilon}_n = (\mathbf{I}_n - \lambda_0 \mathbf{M}_n) \mathbf{u}_n = \mathbf{R}_0 \mathbf{u}_n$, where $\mathbf{R}_0 = (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)$, and $\mathbf{u}_n = \mathbf{y}_n - \mathbf{X}_n \boldsymbol{\beta}_0$. The reduced form equation of this model is

$$\mathbf{y}_n = \mathbf{X}_n \boldsymbol{\beta}_0 + \mathbf{R}_0^{-1} \boldsymbol{\varepsilon}_n. \quad (6.75)$$

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \lambda)^\top$. The GMM estimator of this model can be based on the following linear and quadratic empirical moments

$$\mathbf{g}_n(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{X}_n^\top \boldsymbol{\varepsilon}_n \\ \boldsymbol{\varepsilon}_n(\boldsymbol{\theta})^\top \mathbf{P}_{1n}^\top \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) \\ \boldsymbol{\varepsilon}_n(\boldsymbol{\theta})^\top \mathbf{P}_{2n}^\top \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) \end{pmatrix} = \begin{pmatrix} \mathbf{X}_n^\top \mathbf{R}_n \mathbf{u}_n \\ \mathbf{u}_n^\top \mathbf{R}_n^\top \mathbf{P}_{1n}^\top \mathbf{R}_n \mathbf{u}_n \\ \mathbf{u}_n^\top \mathbf{R}_n^\top \mathbf{P}_{2n}^\top \mathbf{R}_n \mathbf{u}_n \end{pmatrix} = \begin{pmatrix} \mathbf{X}_n^\top \\ \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{1n}^\top \\ \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{2n}^\top \end{pmatrix} \mathbf{R}_n \mathbf{u}_n,$$

where $\mathbf{R}_n = (\mathbf{I}_n - \lambda \mathbf{M}_n)$. At $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, the population moments are $\mathbb{E}[\mathbf{g}_n(\boldsymbol{\theta}_0)] = \mathbb{E}[(\mathbf{X}_n, \mathbf{P}_{1n} \boldsymbol{\varepsilon}_n, \mathbf{P}_{2n} \boldsymbol{\varepsilon}_n)^\top \boldsymbol{\varepsilon}_n] = \mathbf{0}$, because under exogeneity $\mathbb{E}(\mathbf{X}_n^\top \boldsymbol{\varepsilon}_n) = \mathbf{X}_n^\top \mathbb{E}(\boldsymbol{\varepsilon}_n) = \mathbf{0}$ and $\mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{P}_{jn} \boldsymbol{\varepsilon}_n) = \sigma_0^2 \text{tr}(\mathbf{P}_{jn}) = 0$ for $j = 1, 2$.

The derivatives of $\mathbf{g}_n(\boldsymbol{\theta})$ with respect to $\boldsymbol{\beta}$ and λ are

$$\frac{\partial \mathbf{g}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} = \begin{pmatrix} \mathbf{X}_n^\top \frac{\partial \boldsymbol{\varepsilon}_n}{\partial \boldsymbol{\beta}^\top} \\ \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{1n}^s \frac{\partial \boldsymbol{\varepsilon}_n}{\partial \boldsymbol{\beta}^\top} \\ \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{2n}^s \frac{\partial \boldsymbol{\varepsilon}_n}{\partial \boldsymbol{\beta}^\top} \end{pmatrix} = \begin{pmatrix} \mathbf{X}_n^\top \\ \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{1n}^s \\ \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{2n}^s \end{pmatrix} \frac{\partial \boldsymbol{\varepsilon}_n}{\partial \boldsymbol{\beta}^\top},$$

where $\mathbf{P}_{jn}^s = \mathbf{P}_{jn} + \mathbf{P}_{jn}^\top$, $\boldsymbol{\varepsilon}_n = \mathbf{R}_n \mathbf{y}_n - \mathbf{R}_n \mathbf{X} \boldsymbol{\beta}$, and

$$\frac{\partial \boldsymbol{\varepsilon}_n}{\partial \boldsymbol{\theta}^\top} = \begin{pmatrix} \frac{\partial \boldsymbol{\varepsilon}_n}{\partial \boldsymbol{\beta}^\top} & \frac{\partial \boldsymbol{\varepsilon}_n}{\partial \lambda} \end{pmatrix} = \begin{pmatrix} -\mathbf{R}_n \mathbf{X}_n & -\mathbf{M}_n \mathbf{u}_n \end{pmatrix}.$$

Thus,

$$\frac{\partial \mathbf{g}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top} = - \begin{pmatrix} \mathbf{X}_n^\top \mathbf{R}_n \mathbf{X}_n & \mathbf{X}_n^\top \mathbf{M}_n \mathbf{u}_n \\ \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{1n}^s \mathbf{R}_n \mathbf{X}_n & \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{1n}^s \mathbf{M}_n \mathbf{u}_n \\ \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{2n}^s \mathbf{R}_n \mathbf{X}_n & \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{2n}^s \mathbf{M}_n \mathbf{u}_n \end{pmatrix}.$$

The variance-covariance matrix of the moments is:

$$\boldsymbol{\Omega}_n = \begin{pmatrix} \mathbf{O}_{(k \times k)} & \mu_3 \mathbf{X}_n^\top \boldsymbol{\omega}_n_{(k \times 2)} \\ \mu_3 \boldsymbol{\omega}_n^\top \mathbf{X}_n_{(2 \times k)} & (\mu_4 - 3\sigma_0^4) \boldsymbol{\omega}_n^\top \boldsymbol{\omega}_n_{(2 \times 2)} \end{pmatrix} + \mathbf{V}_n,$$

with $\boldsymbol{\omega}_n = [\text{diag}(\mathbf{P}_{1n}), \text{diag}(\mathbf{P}_{2n})]$ is $n \times 2$ and

$$\mathbf{V}_n = \sigma_0^4 \begin{pmatrix} \frac{1}{\sigma_0^2} \mathbf{X}_n^\top \mathbf{X}_n_{(k \times k)} & \mathbf{0}_{(k \times 1)} & \mathbf{0}_{(k \times 1)} \\ \mathbf{0}_{(1 \times k)} & \text{tr}(\mathbf{P}_{1n}^s \mathbf{P}_{1n})_{(1 \times 1)} & \text{tr}(\mathbf{P}_{1n}^s \mathbf{P}_{2n})_{(1 \times 1)} \\ \mathbf{0}_{(1 \times k)} & \text{tr}(\mathbf{P}_{2n}^s \mathbf{P}_{1n})_{(1 \times 1)} & \text{tr}(\mathbf{P}_{2n}^s \mathbf{P}_{2n})_{(1 \times 1)} \end{pmatrix}.$$

When $\boldsymbol{\varepsilon}_n$ is normally distributed, $\boldsymbol{\Omega}_n$ is simplified to \mathbf{V}_n because μ_3 and $\mu_4 = 3\sigma_0^4$.

If $\boldsymbol{\Upsilon}_n = \left(\frac{1}{n} \widehat{\boldsymbol{\Omega}}_n \right)^{-1}$, then the OGMM estimator has the following asymptotic distribution

$$\sqrt{n} \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \right) \xrightarrow{d} \mathbf{N}(\mathbf{0}, \boldsymbol{\Sigma}^0),$$

where

$$\Sigma^0 = \left[\lim_{n \rightarrow \infty} D_n^\top \Omega_n D_n \right]^{-1}.$$

At θ_0 , $\frac{\partial \mathbb{E}[g(\theta_0)]}{\partial \theta^\top} = -D_n$, where

$$D_n = \begin{pmatrix} X_n^\top R_0 X_n & \mathbf{0} \\ \mathbf{0} & \sigma_0^2 \text{tr}(P_{1n}^s Q_0) \\ \mathbf{0} & \sigma_0^2 \text{tr}(P_{2n}^s Q_0) \end{pmatrix} \quad (6.76)$$

$\begin{matrix} k \times k & k \times 1 \\ 1 \times k & 1 \times 1 \\ 1 \times k & 1 \times 1 \end{matrix}$

where $Q_0 = M_n R_0^{-1}$.

6.5.2 Coding GMM Estimator for SEM

First, we create the function `moments.lee2010` which creates the empirical moments

```
moments.lee2010 <- function(theta, y, X, M){
  # Theta: beta and lambda
  k      <- ncol(X)
  n      <- nrow(X)
  beta   <- theta[1:k]
  lambda <- tail(theta, n = 1L)
  I      <- diag(n)
  R      <- I - lambda * M
  u.n    <- y - crossprod(t(X), beta)
  epsi   <- R %*% u.n
  P1     <- M
  M2     <- crossprod(t(M), M)
  P2     <- M2 - (tr(M2) / n) * diag(n)
  g.lin  <- crossprod(X, epsi)           # k * 1
  g.q1   <- crossprod(epsi, P1) %*% epsi # 1*1
  g.q2   <- crossprod(epsi, P2) %*% epsi # 1*1
  g      <- rbind(g.lin, g.q1, g.q2)

  # Gradient
  P1s <- P1 + t(P1)
  P2s <- P2 + t(P2)
  D   <- -1 * rbind(t(X),
                    crossprod(epsi, P1s),
                    crossprod(epsi, P2s)
                    ) %*% cbind(R %*% X, M %*% u.n)

  # Return results (note that they are divided by n)
  out <- list(g = g / n , D = D / n)
  return(out)
}
```


The objective function to be minimized is the following:

```
Q.sem <- function(start, y, X, M, Psi, gradient){
  g.hat <- moments.lee2010(theta = start, y = y, X = X, M = M)
  Q <- -1 * crossprod(g.hat$g, Psi) %% g.hat$g
  if (gradient){
    D <- g.hat$D
    Gr <- -2 * crossprod(D, Psi) %% g.hat$g
    attr(Q, "gradient") <- as.vector(Gr)
  }
  return(Q)
}
```

The function that generates the variance-covariance of the moments is the following

```
make.vmom.sem <- function(b.hat, y, X, M){
  k <- ncol(X)
  n <- nrow(X)
  beta <- b.hat[1:k]
  lambda <- tail(b.hat, n = 1L)
  I <- diag(n)
  R <- I - lambda * M
  u.n <- y - crossprod(t(X), beta)
  epsi <- R %% u.n
  sigma2 <- as.numeric(crossprod(epsi) / n)
  P1 <- M
  M2 <- crossprod(t(M), M)
  P2 <- M2 - (tr(M2) / n) * diag(n)
  P1s <- P1 + t(P1)
  P2s <- P2 + t(P2)

  # Construct V: (2 + k) * (2 + k)
  V11 <- (1 / sigma2) * crossprod(X) # k + k
  Delta <- matrix(0, nrow = 2, ncol = 2)
  Delta[1, 1] <- tr(P1s %% P1)
  Delta[1, 2] <- tr(P1s %% P2)
  Delta[2, 1] <- tr(P2s %% P1)
  Delta[2, 2] <- tr(P2s %% P2)
  V <- matrix(0, nrow = (k + 2), ncol = (k + 2))
  V[1:k, 1:k] <- V11
  V[(k + 1):(k + 2), (k + 1):(k + 2)] <- Delta
  V <- sigma2^2 * V
  # Construct first part of Omega
  omega <- cbind(diag(P1), diag(P2)) # n * 2
  mu4.hat <- sum(epsi^4) / n
  mu3.hat <- sum(epsi^3) / n
  Vp1 <- matrix(0, nrow = (k + 2), ncol = (k + 2))
```

```

Vp1[(k + 1):(k + 2), (k + 1):(k + 2)] <- (mu4.hat - 3 * sigma2^2) * crossprod(omega)
Vp1[(k + 1):(k + 2), 1:k] <- mu3.hat * crossprod(omega, X)
Vp1[1:k, (k + 1):(k + 2)] <- mu3.hat * crossprod(X, omega)
Omega <- Vp1 + V
Omega <- Omega / n
}

```

The main function is the following

```

sem.gmm <- function(formula, data, M,
                    estimator = c("gmm", "ogmm"),
                    gradient = TRUE){
  # Model Frame
  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())

  # Estimator
  estimator <- match.arg(estimator)

  # Get variables and globals
  y <- model.response(mf)
  X <- model.matrix(formula, mf)
  n <- nrow(X)
  sn <- nrow(M)
  if (n != sn) stop("number of spatial units in W is different to the number of data")

  # Starting values for optimization
  ols.e <- lm(y ~ X - 1)
  ols.r <- residuals(ols.e)
  start <- c(coef(ols.e), cor(M %*% ols.r, ols.r))
  names(start) <- c(colnames(X), "Mu")

  # GMM estimator with weighting matrix using a identity matrix
  k <- ncol(X)
  Psi <- diag(k + 2)
  require("maxLik")
  opt <- maxLik(logLik = Q.sem,
               start = start,
               method = "bfgs",
               y = y,
               X = X,
               M = M,
               Psi = Psi,

```

```

        gradient = gradient,
        print.level = 3,
        finalHessian = FALSE)

# OGMM: GMM estimator with weighting matrix using the inverse of the var-cov of moment
if (estimator == "ogmm"){
  Omega.hat <- make.vmom.sem(coef(opt), y = y, X = X, M = M)
  Psi       <- solve(Omega.hat)
  opt       <- maxLik(logLik = Q.sem,
                     start = coef(opt),
                     method = "bfgs",
                     y = y,
                     X = X,
                     M = M,
                     Psi = Psi,
                     gradient = gradient,
                     print.level = 3,
                     finalHessian = FALSE)
}

results <- structure(
  list(
    coefficients = coef(opt),
    call        = callT,
    X           = X,
    y           = y,
    Psi         = Psi,
    M           = M,
    estimator   = estimator
  ),
  class = "gmm.sem"
)
return(results)
}

```

The function that creates D is the following

```

make.D.sem <- function(b.hat, y, X, M){
  k      <- ncol(X)
  n      <- nrow(X)
  beta   <- b.hat[1:k]
  lambda <- tail(b.hat, n = 1L)
  I      <- diag(n)
  R      <- I - lambda * M
  u.n    <- y - crossprod(t(X), beta)
  epsi   <- R %*% u.n
  sigma2 <- as.numeric(crossprod(epsi) / n)
}

```

```

P1      <- M
M2      <- crossprod(t(M), M)
P2      <- M2 - (tr(M2) / n) * diag(n)
P1s     <- P1 + t(P1)
P2s     <- P2 + t(P2)
Q       <- M %*% solve(R)

# Generate D
D <- matrix(0, nrow = (k + 2), ncol = k + 1)
rownames(D) <- c(colnames(X), "q1", "q2")
colnames(D) <- c(colnames(X), "Mu")
D[k + 1, k + 1] <- sigma2 * tr(P1s %*% Q)
D[k + 2, k + 1] <- sigma2 * tr(P2s %*% Q)
D[1:k, 1:k] <- t(X) %*% R %*% X
return(D)
}

```

The function to construct the variance-covariance matrix is given by

```

vcov.gmm.sem <- function(object, D = c("population", "gradient"), ...){
  estimator <- object$estimator
  D.type <- match.arg(D)
  X <- object$X
  y <- object$y
  M <- object$M
  k <- ncol(X)
  n <- nrow(X)
  b.hat <- object$coefficients

  if (estimator == "gmm"){
    if (D.type == "population"){
      D <- make.D.sem(b.hat = b.hat, y = y, X = X, M = M)
      D <- D / n
    }
    if (D.type == "gradient"){
      D <- moments.lee2010(b.hat, y = y, X = X, M = M)$D
    }
    Omega <- make.vmom.sem(b.hat = b.hat, y = y, X = X, M = M)
    var <- solve(crossprod(D)) %*% t(D) %*% Omega %*% D %*% solve(crossprod(D)) / n
  }
  if (estimator == "ogmm"){
    if (D.type == "population"){
      D <- make.D.sem(b.hat = b.hat, y = y, X = X, M = M)
      D <- D / n
    }
    if (D.type == "gradient"){
      D <- moments.lee2010(b.hat, y = y, X = X, M = M)$D
    }
  }
}

```

```

}
Psi <- object$Psi
var <- solve(t(D) %*% Psi %*% D) / n
}
return(var)
}

```

Summary S3 functions

```

summary.gmm.sem <- function(object,
                             D = c("population", "gradient"),
                             table = TRUE,
                             digits = max(3, .Options$digits - 3),
                             ...){
  D.type <- match.arg(D)
  n      <- nrow(object$X)
  df     <- n - length(object$coefficients)
  b      <- object$coefficients
  std.err <- sqrt(diag(vcov(object, D = D.type)))
  z      <- b / std.err
  p      <- 2 * pt(-abs(z), df = df)
  CoefTable <- cbind(b, std.err, z, p)
  colnames(CoefTable) <- c("Estimate", "Std.Error", "t-value", "Pr(>|t|)")
  result <- structure(
    list(
      CoefTable = CoefTable,
      digits    = digits,
      call      = object$call),
    class = 'summary.gmm.sem'
  )
  return(result)
}

print.summary.gmm.sem <- function(x,
                                  digits = x$digits,
                                  na.print = "",
                                  symbolic.cor = p > 4,
                                  signif.stars = getOption("show.signif.stars"),
                                  ...){
  {
    cat("\nCall:\n")
    cat(paste(deparse(x$call), sep = "\n", collapse = "\n"), "\n\n", sep = "")

    cat("\nCoefficients:\n")
    printCoefmat(x$CoefTable, digit = digits, P.value = TRUE, has.Pvalue = TRUE)
    invisible(NULL)
  }
}

```

Now, we check our function

```
semgmm <- sem.gmm(y ~ x1 + x2 + x3, data = data,
                  M = M, estimator = "gmm")

## Initial function value: -0.002361787
## Initial gradient value:
##      (Intercept)          x1          x2          x3          Mu
## 0.0003839962 -0.0511733218 -0.0289716000 -0.0427860038 0.0226625302
## initial value 0.002362
## iter 2 value 0.001449
## iter 3 value 0.001204
## iter 4 value 0.001070
## iter 5 value 0.001006
## iter 6 value 0.000970
## iter 7 value 0.000964
## iter 8 value 0.000963
## iter 9 value 0.000963
## iter 10 value 0.000963
## iter 11 value 0.000963
## iter 12 value 0.000963
## iter 12 value 0.000963
## final value 0.000963
## converged

summary(semgmm, D = "population")

##
## Call:
## sem.gmm(formula = y ~ x1 + x2 + x3, data = data, M = M, estimator = "gmm")
##
##
## Coefficients:
##              Estimate Std.Error t-value Pr(>|t|)
## (Intercept) -0.19250    0.15049  -1.279    0.201
## x1          -1.06289    0.06607 -16.087   <2e-16 ***
## x2           0.01368    0.06197   0.221    0.825
## x3           0.98552    0.06651  14.818   <2e-16 ***
## Mu           0.55800    0.04690  11.898   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(semgmm, D = "gradient")

##
## Call:
## sem.gmm(formula = y ~ x1 + x2 + x3, data = data, M = M, estimator = "gmm")
```

```
##
##
## Coefficients:
##           Estimate Std. Error t-value Pr(>|t|)
## (Intercept) -0.19250    0.15049  -1.279    0.201
## x1          -1.06289    0.06622 -16.051   <2e-16 ***
## x2           0.01368    0.06186   0.221    0.825
## x3           0.98552    0.06657  14.804   <2e-16 ***
## Mu           0.55800    0.04923  11.334   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

semOgmm <- sem.gmm(y ~ x1 + x2 + x3, data = data,
                   M = M, estimator = "ogmm")

## Initial function value: -0.002361787
## Initial gradient value:
##      (Intercept)           x1           x2           x3           Mu
## 0.0003839962 -0.0511733218 -0.0289716000 -0.0427860038 0.0226625302
## initial value 0.002362
## iter  2 value 0.001449
## iter  3 value 0.001204
## iter  4 value 0.001070
## iter  5 value 0.001006
## iter  6 value 0.000970
## iter  7 value 0.000964
## iter  8 value 0.000963
## iter  9 value 0.000963
## iter 10 value 0.000963
## iter 11 value 0.000963
## iter 12 value 0.000963
## iter 12 value 0.000963
## final value 0.000963
## converged
## Initial function value: -0.001029206
## Initial gradient value:
##      (Intercept)           x1           x2           x3           Mu
## -6.818661e-07 -1.071751e-04 -2.275764e-03  5.728118e-04  6.167284e-03
## initial value 0.001029
## iter  2 value 0.001019
## iter  3 value 0.001014
## iter  4 value 0.001014
## iter  5 value 0.001014
## iter  6 value 0.001014
## iter  7 value 0.001014
## iter  7 value 0.001014
## iter  7 value 0.001014
```

```
## final value 0.001014
## converged

summary(semOgmm, D = "population")

##
## Call:
## sem.gmm(formula = y ~ x1 + x2 + x3, data = data, M = M, estimator = "ogmm")
##
##
## Coefficients:
##              Estimate Std.Error t-value Pr(>|t|)
## (Intercept) -0.19255    0.15183  -1.268    0.205
## x1          -1.06348    0.06607 -16.096   <2e-16 ***
## x2           0.01109    0.06196   0.179    0.858
## x3           0.98598    0.06652  14.823   <2e-16 ***
## Mu           0.56190    0.04561  12.319   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(semOgmm, D = "gradient")

##
## Call:
## sem.gmm(formula = y ~ x1 + x2 + x3, data = data, M = M, estimator = "ogmm")
##
##
## Coefficients:
##              Estimate Std.Error t-value Pr(>|t|)
## (Intercept) -0.19255    0.15183  -1.268    0.205
## x1          -1.06348    0.06621 -16.061   <2e-16 ***
## x2           0.01109    0.06175   0.180    0.857
## x3           0.98598    0.06658  14.810   <2e-16 ***
## Mu           0.56190    0.04909  11.445   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

sem_mm <- GMerrorsar(y ~ x1 + x2 + x3,
                     data = data,
                     listw = mat2listw(M, style = "W"),
                     verbose = FALSE,
                     se.lambda = TRUE)

summary(sem_mm)

##
## Call:GMerrorsar(formula = y ~ x1 + x2 + x3, data = data, listw = mat2listw(M,
## style = "W"), verbose = FALSE, se.lambda = TRUE)
```



```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.461416 -1.275682 -0.023397  1.244488  4.935045
##
## Type: GM SAR estimator
## Coefficients: (GM standard errors)
##              Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -0.192429   0.149968  -1.2831   0.1994
## x1          -1.079338   0.063584 -16.9751  <2e-16
## x2           0.006391   0.060099   0.1063   0.9153
## x3           0.975246   0.064057  15.2246  <2e-16
##
## Lambda: 0.55602 (standard error): 0.10114 (z-value): 5.4977
## Residual variance (sigma squared): 2.345, (sigma: 1.5313)
## GM argmin sigma squared: 2.3535
## Number of observations: 529
## Number of parameters estimated: 6

library("sphet")
sem.spreg <- spreg(y ~ x1 + x2 + x3,
                  data = data,
                  listw = mat2listw(M, style = "W"),
                  model = "error")

summary(sem.spreg)

##
## Generalized stsls
##
## Call:
## spreg(formula = y ~ x1 + x2 + x3, data = data, listw = mat2listw(M,
## style = "W"), model = "error")
##
## Residuals:
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -4.4614 -1.2756 -0.0234 -0.0007  1.2445  4.9350
##
## Coefficients:
##              Estimate Std. Error  t-value Pr(>|t|)
## (Intercept) -0.1924293  0.1521211  -1.2650   0.2059
## x1          -1.0793318  0.0633928 -17.0261  <2e-16 ***
## x2           0.0063952  0.0599154   0.1067   0.9150
## x3           0.9752499  0.0638769  15.2677  <2e-16 ***
## rho          0.5631889  0.0488276  11.5342  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6.6 Estimation of SAC Model: The Feasible Generalized Two Stage Least Squares estimator Procedure


6.6.1 Intuition Behind the Procedure

Consider the following SAC model:

$$\begin{aligned} \mathbf{y}_n &= \mathbf{X}_n \boldsymbol{\beta}_0 + \rho_0 \mathbf{W}_n \mathbf{y}_n + \mathbf{u}_n \\ &= \mathbf{Z}_n \boldsymbol{\delta}_0 + \mathbf{u}_n, \\ \mathbf{u}_n &= \lambda_0 \mathbf{M}_n \mathbf{u}_n + \boldsymbol{\varepsilon}_n, \end{aligned} \tag{6.77}$$

where $\mathbf{Z}_n = [\mathbf{X}_n, \mathbf{W}_n \mathbf{y}_n]$, $\boldsymbol{\delta}_0 = [\boldsymbol{\beta}_0^\top, \rho_0]^\top$, \mathbf{y}_n is the $n \times 1$ vector of observations of the dependent variables, \mathbf{X}_n is the $n \times k$ matrix of observations on **nonstochastic (exogenous)** regressors, \mathbf{W}_n and \mathbf{M}_n are the $n \times n$ non stochastic weights matrices, \mathbf{u}_n is the $n \times 1$ vector of regression disturbances, $\boldsymbol{\varepsilon}_n$ is an $n \times 1$ vector of innovations.

It is worth noting that, while the model allows for different spatial weight matrices for each process, in practice, there is rarely a strong justification for assuming different structures.

 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 using ML procedure (see [Anselin, 1988](#)). However, ML estimation requires computing the inverses of $(\mathbf{I}_n - \rho \mathbf{W}_n)$ and $(\mathbf{I}_n - \lambda \mathbf{M}_n)$, which is computationally expensive for large samples. Additionally, ML estimation relies on the assumption of normally distributed errors.

To address these challenges, we rely on estimation techniques from the the S2SLS and GMM. To see how this works, consider applying a spatial Cochrane-Orcutt transformation to the SAC model:

$$\begin{aligned} \mathbf{y}_n &= \mathbf{Z}_n \boldsymbol{\delta}_0 + (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^{-1} \boldsymbol{\varepsilon}_n, \\ (\mathbf{I}_n - \lambda_0 \mathbf{M}_n) \mathbf{y}_n &= (\mathbf{I}_n - \lambda_0 \mathbf{M}_n) \mathbf{Z}_n \boldsymbol{\delta}_0 + \boldsymbol{\varepsilon}_n, \\ \mathbf{y}_s(\lambda_0) &= \mathbf{Z}_s(\lambda_0) \boldsymbol{\delta}_0 + \boldsymbol{\varepsilon}_n, \end{aligned} \tag{6.78}$$

where the spatially filtered variables are given by:

$$\begin{aligned} \mathbf{y}_s(\lambda_0) &= \mathbf{y}_n - \lambda_0 \mathbf{M}_n \mathbf{y}_n \\ &= \mathbf{y}_n - \lambda_0 \mathbf{y}_L, \\ &= (\mathbf{I}_n - \lambda_0 \mathbf{M}_n) \mathbf{y}_n, \\ \mathbf{Z}_s(\lambda_0) &= \mathbf{Z}_n - \lambda_0 \mathbf{M}_n \mathbf{Z}_n, \\ &= \mathbf{Z}_n - \lambda_0 \mathbf{Z}_L, \\ &= (\mathbf{I}_n - \lambda_0 \mathbf{M}_n) \mathbf{Z}_n. \end{aligned}$$

If we knew λ_0 , we would be able to apply an **IV approach on the transformed model** (6.78). For the discussion below, assume that we know λ_0 . Note that the ideal instruments in this case will be:

$$\mathbb{E}(\mathbf{Z}_n(\lambda_0)) = \mathbb{E}(\mathbf{Z}_n) - \lambda_0 \mathbb{E}(\mathbf{M}_n \mathbf{Z}_n)$$

where

$$\begin{aligned}\mathbb{E}(\mathbf{Z}_n) &= \mathbb{E}[\mathbf{X}_n, \mathbf{W}_n \mathbb{E}(\mathbf{y}_n)], \\ \mathbb{E}(\mathbf{M}_n \mathbf{Z}_n) &= \mathbb{E}[\mathbf{M}_n \mathbf{X}_n, \mathbf{M}_n \mathbf{W}_n \mathbb{E}(\mathbf{y}_n)],\end{aligned}$$

Given that all the columns of $\mathbb{E}(\mathbf{Z}_n)$ and $\mathbb{E}(\mathbf{M}_n \mathbf{Z}_n)$ are linear in

$$\mathbf{X}_n, \mathbf{W}_n \mathbf{X}_n, \mathbf{W}_n^2 \mathbf{X}_n, \dots, \mathbf{M}_n \mathbf{X}_n, \mathbf{M}_n \mathbf{W}_n \mathbf{X}_n, \mathbf{M}_n \mathbf{W}_n^2 \mathbf{X}_n, \dots \quad (6.79)$$

the matrix of instruments \mathbf{H}_n is a subset of the linearly independent columns in (6.79), for example

$$\mathbf{H}_n = [\mathbf{X}_n, \mathbf{W}_n \mathbf{X}_n, \dots, \mathbf{W}_n^l \mathbf{X}_n, \mathbf{M}_n \mathbf{X}_n, \mathbf{M}_n \mathbf{W}_n \mathbf{X}_n, \dots, \mathbf{M}_n \mathbf{W}_n^l \mathbf{X}_n]_{LI},$$

where typically, $l \leq 2$.

Since we have the instruments \mathbf{H}_n , and we assuming that we obtain a consistent estimate of λ_0 , we might apply a GMM-type procedure using the following moment conditions for the transformed model (6.78):

$$\mathbf{m}(\lambda_0, \delta_0) = \mathbb{E} \left[\frac{1}{\sqrt{n}} \mathbf{H}_n^\top \boldsymbol{\varepsilon}_n \right] = \mathbf{0}.$$

Now let $\tilde{\lambda}_n$ some consistent estimator for λ_0 which can be obtained in a previous step, then the sample moment vector is:

$$\mathbf{m}^\delta(\tilde{\lambda}_n, \delta_n) = \frac{1}{\sqrt{n}} \mathbf{H}_n^\top \underbrace{\left[\mathbf{y}_s(\tilde{\lambda}_n) - \mathbf{Z}_s(\tilde{\lambda}_n) \delta_n \right]}_{\tilde{\boldsymbol{\varepsilon}}_n},$$

where we explicitly state that the moments depends on δ_n —which will be estimated—and a consistent estimate of λ_0 . Under **homoskedasticity** the variance-covariance matrix of the moment vector $\mathbf{m}(\lambda_0, \delta_0)$ is given by:

$$\mathbb{V}[\mathbf{m}(\lambda_0, \delta_0)] = \mathbb{E}[\mathbf{m}(\lambda_0, \delta_0) \mathbf{m}(\lambda_0, \delta_0)^\top] = \sigma_0^2 n^{-1} \mathbf{H}_n^\top \mathbf{H}_n,$$

which motivates the following two-step GMM estimator for δ_0 :

$$\hat{\boldsymbol{\delta}}_n = \underset{\boldsymbol{\delta}_n}{\operatorname{argmin}} \quad \mathbf{m}_n^\delta(\tilde{\lambda}_n, \boldsymbol{\delta}_n)^\top \boldsymbol{\Upsilon}_n^{\delta\delta} \mathbf{m}_n^\delta(\tilde{\lambda}_n, \boldsymbol{\delta}_n)$$

with

$$\boldsymbol{\Upsilon}_n^{\delta\delta} = \left[\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right]^{-1}.$$

Note that:

$$\begin{aligned}Q_n &= \left[\frac{1}{\sqrt{n}} \mathbf{H}_n^\top \left[\mathbf{y}_s(\tilde{\lambda}_n) - \mathbf{Z}_s(\tilde{\lambda}_n) \boldsymbol{\delta}_n \right] \right]^\top \left[\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right]^{-1} \left[\frac{1}{\sqrt{n}} \mathbf{H}_n^\top \left[\mathbf{y}_s(\tilde{\lambda}_n) - \mathbf{Z}_s(\tilde{\lambda}_n) \boldsymbol{\delta}_n \right] \right], \\ &= \frac{1}{n} \left[\mathbf{y}_s(\tilde{\lambda}_n) - \mathbf{Z}_s(\tilde{\lambda}_n) \boldsymbol{\delta}_n \right]^\top \mathbf{H}_n \left[\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right]^{-1} \mathbf{H}_n^\top \left[\mathbf{y}_s(\tilde{\lambda}_n) - \mathbf{Z}_s(\tilde{\lambda}_n) \boldsymbol{\delta}_n \right], \\ &= \left[\mathbf{y}_s(\tilde{\lambda}_n) - \mathbf{Z}_s(\tilde{\lambda}_n) \boldsymbol{\delta}_n \right]^\top \mathbf{H}_n \left[\mathbf{H}_n^\top \mathbf{H}_n \right]^{-1} \mathbf{H}_n^\top \left[\mathbf{y}_s(\tilde{\lambda}_n) - \mathbf{Z}_s(\tilde{\lambda}_n) \boldsymbol{\delta}_n \right], \\ &= \left[\mathbf{y}_s(\tilde{\lambda}_n) - \mathbf{Z}_s(\tilde{\lambda}_n) \boldsymbol{\delta}_n \right]^\top \mathbf{P}_{H,n} \left[\mathbf{y}_s(\tilde{\lambda}_n) - \mathbf{Z}_s(\tilde{\lambda}_n) \boldsymbol{\delta}_n \right]\end{aligned}$$

Then, the estimator of δ_0 will be:

$$\hat{\delta}_n = \left[\widehat{\mathbf{Z}}_s^\top \mathbf{Z}_s \right]^{-1} \widehat{\mathbf{Z}}_s^\top \mathbf{y}_s$$

where $\widehat{\mathbf{Z}}_s = \mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n \mathbf{Z}_s$. 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 key question remains: How can we obtain a consistent estimator of λ_0 ? As one may anticipate, this consistent estimator is obtained in a preliminary step using the generalized moments (GM) approach.

6.6.2 Moment Conditions Revised

Since we require a consistent estimate of λ_0 , this section focuses on alternative formulations of the moment conditions under both homoskedasticity (Kelejian and Prucha, 1999) and heteroskedasticity (Kelejian and Prucha, 2010). It is important to recall that the GM approach proposed by Kelejian and Prucha (1999), as presented in Section 6.4.2, does not yield a consistent estimate of λ_0 in the presence of heteroskedasticity. Specifically, Theorem 6.22 is derived under the assumption of homoskedasticity. Extensions incorporating a generalized method of moments (GMM) framework have been developed by Kelejian and Prucha (2010), Arraiz et al. (2010), and Drukker et al. (2013).

The GMM approach provides three key improvements over the GM estimator. First, it is robust to heteroskedasticity. Second, it yields an asymptotic variance matrix for the parameter $\hat{\lambda}_n$. Finally, it allows for joint inference on the spatial lag coefficient $\hat{\rho}_n$ and the spatial error coefficient $\hat{\lambda}_n$.

For any nonstochastic matrix \mathbf{A}_n , the quadratic moments are given by:

$$\begin{aligned} \mathbb{E} (\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) &= \text{tr} \left[\mathbb{E} (\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) \right], \\ &= \mathbb{E} \left[\text{tr} (\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) \right], \\ &= \mathbb{E} \left[\text{tr} (\mathbf{A}_n \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^\top) \right], \\ &= \text{tr} \left[\mathbb{E} (\mathbf{A}_n \boldsymbol{\varepsilon}_n \boldsymbol{\varepsilon}_n^\top) \right], \\ &= \text{tr} (\mathbf{A}_n \boldsymbol{\Sigma}_0), \end{aligned}$$

where $\boldsymbol{\Sigma}_0 = \text{Diag}(\sigma_{i,n}^2, \dots, \sigma_{n,n}^2)$. Note that $\mathbb{E} (\boldsymbol{\varepsilon}_n^\top \mathbf{A}_n \boldsymbol{\varepsilon}_n) \neq 0$, unless $\text{tr}(\mathbf{A}_n) = 0$. But if $\text{Diag}(\mathbf{A}_n) = 0$, meaning that $a_{ii,n} = 0$ for all $i = 1, \dots, n$, then $\text{tr}(\mathbf{A}_n \boldsymbol{\Sigma}_0) = 0$ and $\mathbf{A}_n \boldsymbol{\varepsilon}_n$ is uncorrelated with $\boldsymbol{\varepsilon}_n$. Since we need $\text{Diag}(\mathbf{A}_n) = 0$, we can always construct such matrix as $\mathbf{A}_n = \mathbf{P}_n - \text{Diag}(\mathbf{P}_n)$.

Thus, we have the following two moment conditions

$$\begin{aligned} \frac{1}{n} \mathbb{E} [\boldsymbol{\varepsilon}_n^\top \mathbf{A}_{1n} \boldsymbol{\varepsilon}_n] &= \mathbf{0} \\ \frac{1}{n} \mathbb{E} [\boldsymbol{\varepsilon}_n^\top \mathbf{A}_{2n} \boldsymbol{\varepsilon}_n] &= \mathbf{0} \end{aligned} \tag{6.80}$$

with

$$\begin{aligned} \mathbf{A}_{1n} &= \mathbf{M}_n^\top \mathbf{M}_n - \text{Diag}(\mathbf{M}_n^\top \mathbf{M}_n), \\ \mathbf{A}_{2n} &= \mathbf{M}_n. \end{aligned}$$

The sample moments are obtained by replacing $\varepsilon_n(\theta) = (\mathbf{I}_n - \lambda \mathbf{M}_n)(\mathbf{y}_n - \mathbf{Z}_n \delta)$, with $\theta = (\delta^\top, \lambda)^\top$, by their counterpart expressed as a function of the regression residuals. Since $\mathbf{u}_n(\delta) = \mathbf{y}_n - \mathbf{Z}_n \delta$, and $\mathbf{u}_L(\delta) = \mathbf{M}_n \mathbf{u}_n(\delta)$, it follows that $\varepsilon_n(\theta) = \mathbf{u}_n(\delta) - \lambda \mathbf{u}_L(\delta) = \mathbf{y}_s(\lambda) - \mathbf{Z}_s(\lambda) \delta$, the spatially filtered residuals. Then, the set of quadratic moment functions are

$$\begin{aligned} \mathbf{m}_n(\theta) &= \frac{1}{n} \begin{pmatrix} \varepsilon_n(\theta)^\top \mathbf{A}_{1n} \varepsilon_n(\theta) \\ \varepsilon_n(\theta)^\top \mathbf{A}_{2n} \varepsilon_n(\theta) \end{pmatrix}, \\ &= \frac{1}{n} \begin{pmatrix} [\mathbf{u}_n(\delta) - \lambda \mathbf{u}_L(\delta)]^\top \mathbf{A}_{1n} [\mathbf{u}_n(\delta) - \lambda \mathbf{u}_L(\delta)] \\ [\mathbf{u}_n(\delta) - \lambda \mathbf{u}_L(\delta)]^\top \mathbf{A}_{2n} [\mathbf{u}_n(\delta) - \lambda \mathbf{u}_L(\delta)] \end{pmatrix}. \end{aligned} \quad (6.81)$$

We need to write the estimator of λ as a weighted non-linear LS estimator. Note that:

$$\begin{aligned} \frac{1}{n} \varepsilon_n^\top(\theta_0) \mathbf{A}_{qn} \varepsilon_n(\theta_0) &= [\mathbf{u}_n(\delta) - \lambda \mathbf{u}_L(\delta)]^\top \mathbf{A}_{qn} [\mathbf{u}_n(\delta) - \lambda \mathbf{u}_L(\delta)], \\ &= \frac{1}{n} \mathbf{u}_n^\top \mathbf{A}_{qn} \mathbf{u}_n - \frac{1}{n} \lambda (\mathbf{u}_n^\top \mathbf{A}_{qn} \mathbf{u}_L + \mathbf{u}_L^\top \mathbf{A}_{qn} \mathbf{u}_n) + \frac{1}{n} \lambda^2 \mathbf{u}_L^\top \mathbf{A}_{qn} \mathbf{u}_L, \\ &= \frac{1}{n} \mathbf{u}_n^\top \mathbf{A}_{qn} \mathbf{u}_n - 2 \frac{1}{n} \lambda \mathbf{u}_L^\top \mathbf{A}_{qn} \mathbf{u}_n + \frac{1}{n} \lambda^2 \mathbf{u}_L^\top \mathbf{A}_{qn} \mathbf{u}_L, \end{aligned} \quad (6.82)$$

for $q = 1, 2$. In the third line of Equation (6.82), we assume that \mathbf{A}_{qn} is symmetric such that:

$$\begin{aligned} \mathbf{u}_n^\top \mathbf{A}_{qn} \mathbf{u}_L + \mathbf{u}_L^\top \mathbf{A}_{qn} \mathbf{u}_n &= \mathbf{u}_L^\top \mathbf{A}_{qn}^\top \mathbf{u}_n + \mathbf{u}_L^\top \mathbf{A}_{qn} \mathbf{u}_n, \\ &= \mathbf{u}_L^\top (\mathbf{A}_{qn} + \mathbf{A}_{qn}^\top) \mathbf{u}_n, \\ &= 2 \mathbf{u}_L^\top \mathbf{A}_{qn} \mathbf{u}_n. \end{aligned}$$

Here it is important to note that in some cases $\mathbf{A}_{2n} = \mathbf{M}_n$ might not be symmetric. However, we can use Definition 3.9.1 and set:

$$\mathbf{A}_{2n} = (1/2) (\mathbf{M}_n + \mathbf{M}_n^\top) \quad (6.83)$$

Taking expectation over (6.82):

$$\begin{aligned} \frac{1}{n} \mathbb{E}(\varepsilon_n^\top \mathbf{A}_{qn} \varepsilon_n) &= n^{-1} \mathbb{E}(\mathbf{u}_n^\top \mathbf{A}_{qn} \mathbf{u}_n) - 2n^{-1} \lambda \mathbb{E}(\mathbf{u}_L^\top \mathbf{A}_{qn} \mathbf{u}_n) + \lambda^2 n^{-1} \mathbb{E}(\mathbf{u}_L^\top \mathbf{A}_{qn} \mathbf{u}_L) \\ 0 &= n^{-1} \mathbb{E}(\mathbf{u}_n^\top \mathbf{A}_{qn} \mathbf{u}_n) - (2n^{-1} \mathbb{E}(\mathbf{u}_L^\top \mathbf{A}_{qn} \mathbf{u}_n) - n^{-1} \mathbb{E}(\mathbf{u}_L^\top \mathbf{A}_{qn} \mathbf{u}_L)) \begin{pmatrix} \lambda \\ \lambda^2 \end{pmatrix} \end{aligned}$$

Then, we have the following system of equations for $q = 1, 2$ (see Kelejian and Prucha, 2010, pag 56):

$$\begin{aligned} &\begin{pmatrix} n^{-1} \mathbb{E}(\mathbf{u}_n^\top \mathbf{A}_{1q} \mathbf{u}_n) \\ n^{-1} \mathbb{E}(\mathbf{u}_n^\top \mathbf{A}_{2q} \mathbf{u}_n) \end{pmatrix} - \begin{pmatrix} 2n^{-1} \mathbb{E}(\mathbf{u}_L^\top \mathbf{A}_{1q} \mathbf{u}_n) & -n^{-1} \mathbb{E}(\mathbf{u}_L^\top \mathbf{A}_{1q} \mathbf{u}_L) \\ 2n^{-1} \mathbb{E}(\mathbf{u}_L^\top \mathbf{A}_{2q} \mathbf{u}_n) & -n^{-1} \mathbb{E}(\mathbf{u}_L^\top \mathbf{A}_{2q} \mathbf{u}_L) \end{pmatrix} \begin{pmatrix} \lambda \\ \lambda^2 \end{pmatrix} = \mathbf{0} \\ &\begin{pmatrix} n^{-1} \mathbb{E}(\mathbf{u}_n^\top \mathbf{A}_{1q} \mathbf{u}_n) \\ n^{-1} \mathbb{E}(\mathbf{u}_n^\top \mathbf{A}_{2q} \mathbf{u}_n) \end{pmatrix} - \begin{pmatrix} 2n^{-1} \mathbb{E}(\mathbf{u}_n^\top \mathbf{M}_n^\top \mathbf{A}_{1q} \mathbf{u}_n) & -n^{-1} \mathbb{E}(\mathbf{u}_n^\top \mathbf{M}_n^\top \mathbf{A}_{1q} \mathbf{M}_n \mathbf{u}_n) \\ n^{-1} \mathbb{E}(\mathbf{u}_n^\top \mathbf{M}_n^\top \mathbf{A}_{2q} \mathbf{u}_n) & -n^{-1} \mathbb{E}(\mathbf{u}_n^\top \mathbf{M}_n^\top \mathbf{A}_{2q} \mathbf{M}_n \mathbf{u}_n) \end{pmatrix} = \mathbf{0} \\ &\gamma_n - \Gamma_n \alpha_n = \mathbf{0}. \end{aligned} \quad (6.84)$$

where we use Equation (6.83) for the second moment. Now, we can express the **sample moment conditions** as in Section 6.4.2:

$$\widetilde{\mathbf{m}}_n(\widetilde{\delta}_n, \lambda) = \widetilde{\mathbf{g}}_n - \widetilde{\mathbf{G}}_n \begin{pmatrix} \lambda \\ \lambda^2 \end{pmatrix} = \mathbf{0}$$

The elements of $\tilde{\mathbf{g}}_n$ the following:

$$\begin{aligned}\tilde{\mathbf{g}}_{1n} &= \frac{1}{n} \tilde{\mathbf{u}}_n^\top \mathbf{A}_{1n} \tilde{\mathbf{u}}_n \\ \tilde{\mathbf{g}}_{2n} &= \frac{1}{n} \tilde{\mathbf{u}}_n^\top \mathbf{A}_{2n} \tilde{\mathbf{u}}_n = \frac{1}{n} \tilde{\mathbf{u}}_n^\top \tilde{\mathbf{u}}_L\end{aligned}$$

The $\tilde{\mathbf{G}}_n$ matrix is given by:

$$\begin{aligned}\tilde{\mathbf{G}}_{11,n} &= 2n^{-1} \tilde{\mathbf{u}}_n^\top \mathbf{M}_n^\top \mathbf{A}_{1n} \tilde{\mathbf{u}}_n \\ \tilde{\mathbf{G}}_{12,n} &= -n^{-1} \tilde{\mathbf{u}}_n^\top \mathbf{M}_n^\top \mathbf{A}_{1n} \mathbf{M}_n \tilde{\mathbf{u}}_n \\ \tilde{\mathbf{G}}_{21,n} &= -n^{-1} \tilde{\mathbf{u}}_n^\top \mathbf{M}_n^\top (\mathbf{A}_{2n} + \mathbf{A}_{2n}^\top) \tilde{\mathbf{u}}_n \\ \tilde{\mathbf{G}}_{22,n} &= -n^{-1} \tilde{\mathbf{u}}_n^\top \mathbf{M}_n \mathbf{A}_{2n} \mathbf{M}_n \tilde{\mathbf{u}}_n\end{aligned}$$

A more compact notation is:

$$\begin{aligned}\tilde{\mathbf{G}}_n &= \frac{1}{n} \begin{pmatrix} \tilde{\mathbf{u}}_n^\top (\mathbf{A}_{1n} + \mathbf{A}_{1n}^\top) \tilde{\mathbf{u}}_s & -\tilde{\mathbf{u}}_s^\top \mathbf{A}_{1n} \tilde{\mathbf{u}}_s^\top \\ \vdots & \vdots \\ \tilde{\mathbf{u}}_n^\top (\mathbf{A}_{qn} + \mathbf{A}_{qn}^\top) \tilde{\mathbf{u}}_s & -\tilde{\mathbf{u}}_s^\top \mathbf{A}_{qn} \tilde{\mathbf{u}}_s^\top \end{pmatrix}, \\ \tilde{\mathbf{g}}_n &= \frac{1}{n} \begin{pmatrix} \tilde{\mathbf{u}}_n^\top \mathbf{A}_{1n} \tilde{\mathbf{u}}_n \\ \vdots \\ \tilde{\mathbf{u}}_n^\top \mathbf{A}_{qn} \tilde{\mathbf{u}}_n \end{pmatrix},\end{aligned}\tag{6.85}$$

for $q = 1, 2$.

6.6.3 Assumptions

In this section, we outline the key assumptions for the SAC model under heteroskedasticity, following [Arraiz et al. \(2010\)](#). These assumptions primarily concern the spatial weight matrices, error structure, regressors, and instruments.

Assumption 6.26 — Spatial Weights Matrices ([Arraiz et al., 2010](#)). Assume the following:

- (a) All diagonal elements \mathbf{W}_n and \mathbf{M}_n are zero.
- (b) $\lambda_n \in (-1, 1)$, $\rho_n \in (-1, 1)$.
- (c) The matrices $\mathbf{I}_n - \rho_n \mathbf{W}_n$ and $\mathbf{I}_n - \lambda_n \mathbf{M}_n$ are nonsingular for all $\lambda_n \in (-1, 1)$ and $\rho_n \in (-1, 1)$.

Assumption 6.26(a) is a normalization rule, ensuring that a region cannot be its own neighbor. Assumption 6.26(b) defines the parameter space, as discussed in [Kelejian and Prucha \(2010, section 2.2\)](#). Finally, Assumption 6.26(c) ensures that the spatial processes for \mathbf{y}_n and \mathbf{u}_n are uniquely defined. Under this assumption, the SAC model can be written as:

$$\begin{aligned}\mathbf{y}_n &= (\mathbf{I}_n - \rho_0 \mathbf{W}_n)^{-1} [\mathbf{X}_n \boldsymbol{\beta}_0 + \mathbf{u}_n] \\ \mathbf{u}_n &= (\mathbf{I}_n - \rho_0 \mathbf{M}_n)^{-1} \boldsymbol{\epsilon}_n.\end{aligned}$$

The reduced form is:

$$\mathbf{y}_n = (\mathbf{I}_n - \rho_0 \mathbf{W}_n)^{-1} \mathbf{X}_n \boldsymbol{\beta}_0 + (\mathbf{I}_n - \rho_0 \mathbf{W}_n)^{-1} (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^{-1} \boldsymbol{\epsilon}_n.$$

Assumption 6.27 — 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.

This assumption allows for heteroskedasticity in the innovations while ensuring uniformly bounded variances. It also accommodates triangular array structures where variances may depend on the sample size n .

Assumption 6.28 — 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 $(\mathbf{I}_n - \rho_0 \mathbf{W}_n)^{-1}$ and $(\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^{-1}$ are bounded uniformly in absolute value by some finite constant.

This assumption is a technical requirement for deriving large-sample properties of the estimators. It ensures that spatial dependence in \mathbf{y}_n and \mathbf{u}_n does not accumulate indefinitely, maintaining a “fading” memory property. Under this assumption:

$$\begin{aligned} \mathbb{E}[\mathbf{u}_n] &= \mathbb{E}[(\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^{-1} \boldsymbol{\epsilon}_n], \\ &= (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^{-1} \mathbb{E}[\boldsymbol{\epsilon}_n], \\ &= \mathbf{0} \quad \text{by Assumption 6.27 (Heteroskedastic Errors).} \end{aligned}$$

Additionally, the variance of \mathbf{u}_n is given by:

$$\begin{aligned} \mathbb{E}[\mathbf{u}_n \mathbf{u}_n^\top] &= \mathbb{E}\left[(\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^{-1} \boldsymbol{\epsilon}_n \boldsymbol{\epsilon}_n^\top (\mathbf{I}_n - \lambda_0 \mathbf{M}_n^\top)^{-1}\right], \\ &= (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^{-1} \mathbb{E}[\boldsymbol{\epsilon}_n \boldsymbol{\epsilon}_n^\top] (\mathbf{I}_n - \lambda_0 \mathbf{M}_n^\top)^{-1}, \\ &= (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^{-1} \boldsymbol{\Sigma}_0 (\mathbf{I}_n - \lambda_0 \mathbf{M}_n^\top)^{-1} \end{aligned}$$

where $\boldsymbol{\Sigma}_0 = \text{Diag}(\sigma_{i,n}^2)$.

Assumption 6.29 — 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 prevents multicollinearity and ensures that the regressors remain well-behaved in large samples.

Assumption 6.30 — 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 6.31 — 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 $\mathbf{Q}_{HZ,s}(\lambda_0) = \mathbf{Q}_{HZ} - \lambda_0 \mathbf{Q}_{HMZ}$ has full column rank.
- (c) $\mathbf{Q}_{H\Sigma H} = \lim_{n \rightarrow \infty} n^{-1} \mathbf{H}_n^\top \Sigma_n \mathbf{H}_n$ is finite and nonsingular, where $\Sigma_n = \text{diag}(\sigma_{i,n}^2)$

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 .

6.6.4 Estimators and Estimation Procedure in a Nutshell

Consider again the transformed model:

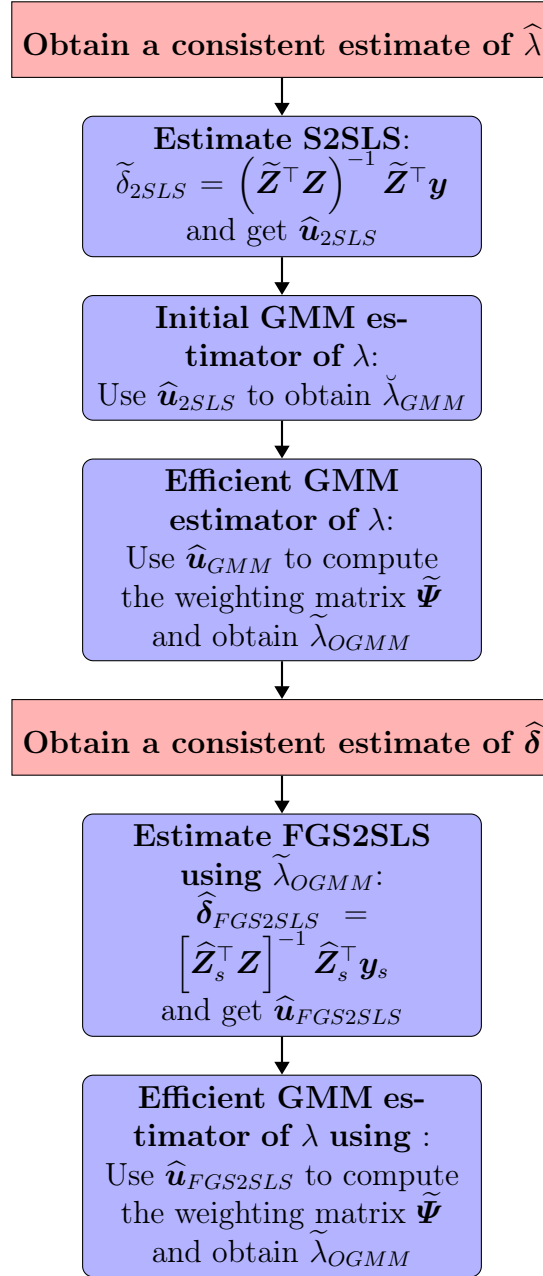
$$\mathbf{y}_{s,n}(\lambda_0) = \mathbf{Z}_{s,n}(\lambda_0) \boldsymbol{\delta}_0 + \epsilon_n,$$

where $\mathbf{y}_{s,n}(\lambda_0) = \mathbf{y}_n - \lambda_0 \mathbf{M}_n \mathbf{y}_n$ and $\mathbf{Z}_{s,n}(\lambda_0) = \mathbf{Z}_n - \lambda_0 \mathbf{M}_n \mathbf{Z}_n$. If λ_0 were known, we could directly apply the S2SLS estimator to the transformed model. However, since λ_0 is unknown, it must be estimated before $\boldsymbol{\delta}_0$. The estimation procedure consists of the following steps

- (a) Obtain an initial IV estimator of $\boldsymbol{\delta}_0$ to construct consistent residuals.
- (b) Use these residuals to derive the moment conditions that provide a consistent estimate of λ_0 via GMM.
- (c) Use the estimate of λ_0 to define a **weighting matrix** for the moment conditions to obtain a consistent and efficient estimator.
- (d) Estimate $\boldsymbol{\delta}_0$ 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. Now we will consider each step in detail.

Figure 6.1: Estimation steps for SAC model



Step 1a: S2SLS estimator

In the first step, δ_0 is estimated by 2SLS applied to **untransformed model** $y_n = Z_n \delta_0 + u_n$ using the instruments matrix H_n . Then:

$$\tilde{\delta}_{S2SLS,n} = (\tilde{Z}_n^\top Z_n)^{-1} \tilde{Z}_n^\top y_n,$$

where $\tilde{Z}_n = H_n (H_n^\top H_n)^{-1} H_n^\top Z_n = P_{H,n} Z_n = (X_n, \widetilde{W_n y_n})$. The estimates $\tilde{\delta}_{S2SLS,n}$ yield an initial vector of residuals, $u_{n,S2SLS}$ as:

$$\tilde{u}_{S2SLS,n} = y_n - Z_n \tilde{\delta}_{n,S2SLS}.$$

The following Theorem states that $\tilde{\delta}_{S2SL,n}$ is consistent:

Theorem 6.32 — Consistency of $\tilde{\delta}_{S2SL,n}$ (Kelejian and Prucha, 2010). Suppose the assumptions hold. Then $\tilde{\delta}_{S2SL,n} = \delta_0 + O_p(n^{-1/2})$, and hence $\tilde{\delta}_{n,S2SL}$ is consistent for δ_0 :

$$\tilde{\delta}_{S2SL,n} \xrightarrow{p} \delta_0.$$

Sketch of proof for Theorem 6.32. The model is:

$$\begin{aligned} \mathbf{y}_n &= \mathbf{Z}_n \delta_0 + \mathbf{u}_n, \\ \mathbf{u}_n &= \lambda_0 \mathbf{M}_n \mathbf{u}_n + \varepsilon_n. \end{aligned}$$

The sampling error is given by:

$$\begin{aligned} \hat{\delta}_n &= \delta_0 + \left(\hat{\mathbf{Z}}_n^\top \hat{\mathbf{Z}}_n \right)^{-1} \hat{\mathbf{Z}}_n^\top \mathbf{u}_n, \\ &= \delta_0 + \left[\left(\mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{Z}_n \right)^\top \left(\mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{Z}_n \right) \right]^{-1} \left(\mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{Z}_n \right)^\top \mathbf{u}_n, \\ &= \delta_0 + \left[\mathbf{Z}_n^\top \mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \mathbf{Z}_n \right]^{-1} \mathbf{Z}_n^\top \mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top (\mathbf{I} - \lambda_0 \mathbf{M}_n)^{-1} \varepsilon_n. \end{aligned}$$

Solving for $\hat{\delta}_n - \delta_0$ and multiplying by \sqrt{n} we obtain:

$$\begin{aligned} \sqrt{n}(\hat{\delta}_n - \delta_0) &= \left[\left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right)^\top \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right) \right]^{-1} \\ &\quad \times \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right)^\top \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \frac{1}{\sqrt{n}} \mathbf{H}_n^\top (\mathbf{I} - \lambda \mathbf{M}_n)^{-1} \varepsilon_n, \\ &= \left[\left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right)^\top \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right) \right]^{-1} \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right)^\top \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \frac{1}{\sqrt{n}} \mathbf{F}_n^\top \varepsilon_n, \end{aligned}$$

where:

$$\mathbf{F}_n^\top = \mathbf{H}_n^\top (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^{-1} \quad \text{whose elements are bounded in absolute value.}$$

Assumption 6.31 implies that:

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n &= \mathbf{Q}_{HH}, \\ \text{plim}_{n \rightarrow \infty} \frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n &= \mathbf{Q}_{HZ}, \end{aligned}$$

which are finite and nonsingular.

Furthermore, note that $\mathbb{E}(n^{-1/2} \mathbf{F}_n^\top \varepsilon_n) = \mathbf{0}$ and, under homoskedasticity,

$$\begin{aligned} \mathbb{E} \left[(n^{-1/2} \mathbf{F}_n^\top \varepsilon_n) (n^{-1/2} \mathbf{F}_n^\top \varepsilon_n)^\top \right] &= \frac{1}{n} \mathbb{E} \left[\mathbf{H}_n^\top (\mathbf{I}_n - \lambda \mathbf{M}_n)^{-1} \varepsilon_n \varepsilon_n^\top (\mathbf{I}_n - \lambda \mathbf{M}_n^\top)^{-1} \mathbf{H}_n \right], \\ &= \sigma_0^2 \frac{1}{n} \mathbf{H}_n^\top (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^{-1} (\mathbf{I}_n - \lambda_0 \mathbf{M}_n^\top)^{-1} \mathbf{H}_n. \end{aligned}$$

Assume that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \mathbf{H}_n^\top (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^{-1} (\mathbf{I}_n - \lambda_0 \mathbf{M}_n^\top)^{-1} \mathbf{H}_n = \frac{1}{n} \mathbf{F}_n^\top \mathbf{F}_n = \Phi \quad \text{exists}$$

Then using Theorem 3.30:

$$n^{-1/2} \mathbf{F}_n^\top \boldsymbol{\varepsilon}_n \xrightarrow{d} N(\mathbf{0}, \sigma_0^2 \boldsymbol{\Phi}).$$

Therefore:

$$\sqrt{n}(\hat{\boldsymbol{\delta}}_n - \boldsymbol{\delta}_0) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Delta}),$$

and

$$\boldsymbol{\Delta} = \sigma_0^2 [\mathbf{Q}_{HZ}^\top \mathbf{Q}_{HH}^{-1} \mathbf{Q}_{HZ}]^{-1} \mathbf{Q}_{HZ}^\top \mathbf{Q}_{HH}^{-1} \boldsymbol{\Phi} \mathbf{Q}_{HH}^{-1} \mathbf{Q}_{HZ} [\mathbf{Q}_{HZ}^\top \mathbf{Q}_{HH}^{-1} \mathbf{Q}_{HZ}]^{-1}.$$

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

$$\mathbb{V}(n^{-1} \mathbf{F}_n^\top \boldsymbol{\varepsilon}_n) = \sigma_0^2 \frac{1}{n^2} \mathbf{H}_n^\top (\mathbf{I}_n - \lambda_0 \mathbf{M}_n)^{-1} (\mathbf{I}_n - \lambda_0 \mathbf{M}_n^\top)^{-1} \mathbf{H}_n,$$

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 $\tilde{\boldsymbol{\delta}}_{S2SLS,n}$ 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 S2SLS residuals

Using the consistent estimate \mathbf{u}_n in the previous step, now we create the sample moments for $q = 1, 2$ based on the estimated residuals. Define:

$$\begin{aligned} \tilde{\mathbf{u}}_{n,S2SLS} &= \mathbf{y}_n - \mathbf{Z}_n \tilde{\boldsymbol{\delta}}_{n,S2SLS}, \\ \tilde{\mathbf{u}}_{s,n} &= \mathbf{M}_n \tilde{\mathbf{u}}_{n,S2SLS}, \end{aligned}$$

and the sample moments are:

$$\begin{aligned} \mathbf{m}_n(\lambda_n, \tilde{\boldsymbol{\delta}}_{n,S2SLS}) &= \frac{1}{n} \begin{pmatrix} \tilde{\mathbf{u}}_{n,S2SLS}^\top (\mathbf{I}_n - \lambda_n \mathbf{M}_n^\top) \mathbf{A}_{1n} (\mathbf{I}_n - \lambda_n \mathbf{M}_n) \tilde{\mathbf{u}}_{n,S2SLS} \\ \tilde{\mathbf{u}}_{n,S2SLS}^\top (\mathbf{I}_n - \lambda_n \mathbf{M}_n^\top) \mathbf{A}_{2n} (\mathbf{I}_n - \lambda_n \mathbf{M}_n) \tilde{\mathbf{u}}_{n,S2SLS} \end{pmatrix}, \\ &= \tilde{\mathbf{G}}_n \begin{pmatrix} \lambda_n \\ \lambda_n^2 \end{pmatrix} - \tilde{\mathbf{g}}_n, \end{aligned} \tag{6.86}$$

where,

$$\begin{aligned} \tilde{\mathbf{G}}_n &= \frac{1}{n} \begin{pmatrix} \tilde{\mathbf{u}}_n^\top (\mathbf{A}_{1n} + \mathbf{A}_{1n}^\top) \tilde{\mathbf{u}}_{n,s} & -\tilde{\mathbf{u}}_{n,s}^\top \mathbf{A}_{1n} \tilde{\mathbf{u}}_{s,n}^\top \\ \vdots & \vdots \\ \tilde{\mathbf{u}}_n^\top (\mathbf{A}_{qn} + \mathbf{A}_{qn}^\top) \tilde{\mathbf{u}}_{n,s} & -\tilde{\mathbf{u}}_{n,s}^\top \mathbf{A}_{qn} \tilde{\mathbf{u}}_{s,n}^\top \end{pmatrix}, \\ \tilde{\mathbf{g}}_n &= \frac{1}{n} \begin{pmatrix} \tilde{\mathbf{u}}_n^\top \mathbf{A}_{1n} \tilde{\mathbf{u}}_n \\ \vdots \\ \tilde{\mathbf{u}}_n^\top \mathbf{A}_{qn} \tilde{\mathbf{u}}_n \end{pmatrix}. \end{aligned}$$

The initial GMM estimator for λ is then defined as

$$\check{\lambda}_{n,\text{GMM}} = \underset{\lambda_n \in \Lambda}{\operatorname{argmin}} \left\{ \left[\tilde{\mathbf{G}}_n \begin{pmatrix} \lambda_n \\ \lambda_n^2 \end{pmatrix} - \tilde{\mathbf{g}}_n \right]^\top \left[\tilde{\mathbf{G}}_n \begin{pmatrix} \lambda_n \\ \lambda_n^2 \end{pmatrix} - \tilde{\mathbf{g}}_n \right] \right\} = \underset{\lambda_n \in \Lambda}{\operatorname{argmin}} \left[\mathbf{m}_n(\lambda_n, \tilde{\boldsymbol{\delta}}_{n,\text{S2SLS}})^\top \mathbf{m}_n(\lambda_n, \tilde{\boldsymbol{\delta}}_{n,\text{S2SLS}}) \right],$$

where $\boldsymbol{\Upsilon}_n^{\lambda\lambda} = \mathbf{I}_2$. This estimator is consistent but not efficient. For efficiency we need to replace $\boldsymbol{\Upsilon}_n^{\lambda\lambda}$ 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 $\mathbf{A}_{n,s}$. [Drukker et al. \(2013\)](#) suggest, for the homoskedastic case, the following expressions:

$$\begin{aligned} \mathbf{A}_{1n} &= v \left[\mathbf{M}_n^\top \mathbf{M}_n - \frac{1}{n} \operatorname{tr}(\mathbf{M}_n^\top \mathbf{M}_n) \mathbf{I}_n \right], \\ \mathbf{A}_{2n} &= \mathbf{M}_n, \end{aligned}$$

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:

$$\begin{aligned} \mathbf{A}_{1n} &= \mathbf{M}_n^\top \mathbf{M}_n - \operatorname{Diag}(\mathbf{M}_n^\top \mathbf{M}_n), \\ \mathbf{A}_{2n} &= \mathbf{M}_n \end{aligned}$$

Step 1c: Efficient GMM estimator of λ_0 based on S2SLS residuals

The efficient GMM estimator of λ_0 is a weighted nonlinear least squares estimator. Specifically, this estimator is $\tilde{\lambda}_n$ where:

$$\tilde{\lambda}_{n,\text{EGMM}} = \underset{\lambda_n \in \Lambda}{\operatorname{argmin}} \left[\mathbf{m}_n(\lambda_n, \tilde{\boldsymbol{\delta}}_n)^\top \tilde{\boldsymbol{\Psi}}_n^{-1} \mathbf{m}_n(\lambda_n, \tilde{\boldsymbol{\delta}}_n) \right], \quad (6.87)$$

and where the weighting matrix is $\tilde{\boldsymbol{\Psi}}_n^{-1}$, where $\boldsymbol{\Psi}_n$ is the variance of the moment conditions $\mathbf{m}(\lambda_0, \tilde{\boldsymbol{\delta}}_n)$.

Let $\tilde{\mathbf{u}}_n = \mathbf{y} - \mathbf{Z} \tilde{\boldsymbol{\delta}}_{n,\text{S2SLS}}$. The matrix $\tilde{\boldsymbol{\Psi}}_n^{-1} = \tilde{\boldsymbol{\Psi}}_n^{-1}(\check{\lambda}_{n,\text{GMM}})$ is defined as follows. Let $\tilde{\boldsymbol{\Psi}}_n = [\tilde{\Psi}_{rs}]_{r,s=1,2}$ with

$$\tilde{\Psi}_{rs} = (2n)^{-1} \operatorname{tr} \left[(\mathbf{A}_{r,n} + \mathbf{A}_{r,n}^\top) \tilde{\boldsymbol{\Sigma}}_n (\mathbf{A}_{s,n} + \mathbf{A}_{s,n}^\top) \tilde{\boldsymbol{\Sigma}}_n \right] + n^{-1} \tilde{\mathbf{a}}_{r,n}^\top \tilde{\boldsymbol{\Sigma}}_n \tilde{\mathbf{a}}_{s,n}, \quad (6.88)$$

where:

$$\begin{aligned} \tilde{\boldsymbol{\Sigma}}_n &= \operatorname{Diag}(\tilde{\epsilon}_{in}^2), \\ \tilde{\boldsymbol{\epsilon}}_n &= \left(\mathbf{I}_n - \check{\lambda}_{n,\text{GMM}} \mathbf{M}_n \right) \tilde{\mathbf{u}}_n, \\ \tilde{\mathbf{a}}_{r,n} &= \left(\mathbf{I}_n - \check{\lambda}_{n,\text{GMM}} \mathbf{M}_n^\top \right)^{-1} \mathbf{H}_n \tilde{\mathbf{P}}_n \tilde{\boldsymbol{\alpha}}_{r,n}, \\ \tilde{\boldsymbol{\alpha}}_{r,n} &= -n^{-1} \left[\mathbf{Z}_n^\top \left(\mathbf{I}_n - \check{\lambda}_{n,\text{GMM}} \mathbf{M}_n^\top \right) (\mathbf{A}_{r,n} + \mathbf{A}_{r,n}^\top) \left(\mathbf{I}_n - \check{\lambda}_{n,\text{GMM}} \mathbf{M}_n \right) \tilde{\mathbf{u}}_n \right] \\ \tilde{\mathbf{P}}_n &= \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right) \left[\left(\frac{1}{n} \mathbf{H}^\top \mathbf{Z}_n \right)^\top \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n \right) \right]^{-1} \end{aligned}$$

It is important to note that this step is not necessary since the previous estimator of λ is already consistent.

Step 2a: GS2SLS Estimator

Using $\check{\lambda}_{ogmm}$ from step 1c (or the consistent estimator from step 1b) in the transformed model we have:

$$\hat{\delta}_n(\tilde{\lambda}_{n,EGMM}) = \left[\hat{\mathbf{Z}}_{s,n}^\top(\tilde{\lambda}_{n,EGMM}) \mathbf{Z}(\tilde{\lambda}_{n,EGMM}) \right]^{-1} \hat{\mathbf{Z}}_{s,n}^\top(\tilde{\lambda}_{n,EGMM}) \mathbf{y}_{s,n}(\tilde{\lambda}_{n,EGMM})$$

where

$$\begin{aligned} \mathbf{y}_{s,n}(\tilde{\lambda}_{n,EGMM}) &= \mathbf{y}_n - \tilde{\lambda}_{n,EGMM} \mathbf{M}_n \mathbf{y}_n, \\ \mathbf{Z}_{s,n}(\tilde{\lambda}_{n,EGMM}) &= \mathbf{Z}_n - \tilde{\lambda}_{n,EGMM} \mathbf{M}_n \mathbf{Z}_n, \\ \hat{\mathbf{Z}}_{s,n}(\tilde{\lambda}_{n,EGMM}) &= \mathbf{P}_{H,n} \mathbf{Z}_{s,n}(\tilde{\lambda}_{n,EGMM}), \\ \mathbf{P}_{H,n} &= \mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top. \end{aligned}$$

Step 2b: Efficient GMM estimator of λ using GS2SLS residuals

In this last step, the **efficient** GMM estimator of λ_0 based on the GS2SLS residuals is obtained by minimizing the following expression:

$$\hat{\lambda}_n = \underset{\lambda_n \in \Lambda}{\operatorname{argmin}} \left\{ \left[\hat{\mathbf{G}}_n \begin{pmatrix} \lambda_n \\ \lambda_n^2 \end{pmatrix} - \hat{\mathbf{g}}_n \right]^\top (\hat{\Psi}_n^{\hat{\lambda}\hat{\lambda}})^{-1} \left[\hat{\mathbf{G}}_n \begin{pmatrix} \lambda_n \\ \lambda_n^2 \end{pmatrix} - \hat{\mathbf{g}}_n \right] \right\}$$

where $\hat{\Psi}_n^{\hat{\lambda}\hat{\lambda}}$ 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.

Note that the residuals are $\hat{\mathbf{u}}_n = \mathbf{y}_n - \mathbf{Z}_n \hat{\delta}_n(\tilde{\lambda}_{n,EGMM})$. For the **heteroskedastic** case the r, s (with $r, s = 1, 2$) element of $\hat{\Psi}_n^{\hat{\lambda}\hat{\lambda}}$ is given by:

$$\hat{\Psi}_{rs}^{\hat{\lambda}\hat{\lambda}} = (2n)^{-1} \operatorname{tr} \left[(\mathbf{A}_{n,r} + \mathbf{A}_{n,r}^\top) \hat{\Sigma}_n (\mathbf{A}_{n,s} + \mathbf{A}_{n,s}^\top) \hat{\Sigma}_n \right] + n^{-1} \hat{\mathbf{a}}_{n,r}^\top \hat{\Sigma}_n \hat{\mathbf{a}}_{n,s}, \quad (6.89)$$

where:

$$\begin{aligned} \hat{\Sigma}_n &= \operatorname{Diag}(\hat{\epsilon}_{in}^2), \\ \hat{\epsilon}_n &= (\mathbf{I}_n - \tilde{\lambda}_{n,EGMM} \mathbf{M}_n) \hat{\mathbf{u}}_n, \\ \hat{\mathbf{a}}_{r,n} &= \mathbf{H}_n \hat{\mathbf{P}}_n \hat{\alpha}_{r,n}, \\ \hat{\alpha}_{r,n} &= -n^{-1} \left[\mathbf{Z}_n^\top (\mathbf{I}_n - \tilde{\lambda}_{n,EGMM} \mathbf{M}_n^\top) (\mathbf{A}_{r,n} + \mathbf{A}_{r,n}^\top) (\mathbf{I}_n - \tilde{\lambda}_{n,EGMM} \mathbf{M}_n) \hat{\mathbf{u}}_n \right] \\ \hat{\mathbf{P}}_n &= \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_{n,s} \right) \left[\left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_{n,s} \right)^\top \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right)^{-1} \left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_{n,s} \right) \right]^{-1} \end{aligned}$$

6.6.5 Asymptotic Distribution

The following Theorem provides the joint limiting distribution of $\hat{\lambda}_n$ and $\hat{\delta}_n$ following [Arraiz et al. \(2010\)](#).

Theorem 6.33 — Asymptotic Distribution of GS2SLS and EGMM estimators: $\hat{\lambda}_n$ and $\hat{\delta}_n$. Under the assumptions made. Then, $\hat{\lambda}_n$ is efficient among the class of GMM estimators based on GS2SLS residuals, and

$$\begin{pmatrix} \sqrt{n}(\hat{\delta}_n - \delta_0) \\ \sqrt{n}(\hat{\lambda}_n - \lambda_0) \end{pmatrix} \xrightarrow{d} N \left[\mathbf{0}, \text{plim}_{n \rightarrow \infty} \hat{\Omega}_n(\hat{\lambda}_n) \right],$$

where $\text{plim}_{n \rightarrow \infty} \hat{\Omega}_n(\hat{\lambda}_n)$ is a positive definite matrix.

The estimator $\hat{\Omega}_n$ is given by

$$\hat{\Omega}_n = \begin{pmatrix} \hat{P}_n^\top & \mathbf{0} \\ \mathbf{0} & (\hat{J}_n^\top \hat{\Psi}_n^{-1} \hat{J}_n)^{-1} \hat{J}_n^\top \hat{\Psi}_n^{-1} \end{pmatrix} \hat{\Psi}_0 \begin{pmatrix} \hat{P}_n & \mathbf{0} \\ \mathbf{0} & \hat{\Psi}_n^{-1} \hat{J}_n (\hat{J}_n^\top \hat{\Psi}_n^{-1} \hat{J}_n)^{-1} \end{pmatrix} \quad (6.90)$$

with

$$\begin{aligned} \hat{\Psi}_0 &= \begin{pmatrix} \hat{\Psi}_{n,\delta\delta} & \hat{\Psi}_{n,\delta\lambda} \\ \hat{\Psi}_{n,\delta\lambda}^\top & \hat{\Psi}_n \end{pmatrix}, \\ \hat{\Psi}_n &= \begin{pmatrix} \hat{\Psi}_{n,11}^{\lambda\lambda} & \hat{\Psi}_{n,12}^{\lambda\lambda} \\ \hat{\Psi}_{n,12}^{\lambda\lambda} & \hat{\Psi}_{n,22}^{\lambda\lambda} \end{pmatrix}, \\ \hat{\Psi}_{n,\delta\delta} &= n^{-1} \mathbf{H}_n^\top \hat{\Sigma}_n \mathbf{H}_n, \\ \hat{\Psi}_{n,\delta\lambda} &= n^{-1} \mathbf{H}_n^\top \hat{\Sigma}_n [\hat{a}_{n,1}, \hat{a}_{n,2}], \\ \hat{J}_n &= \hat{G}_n \begin{pmatrix} 1 \\ 2\hat{\lambda}_n \end{pmatrix}, \\ \hat{G}_n &= \begin{pmatrix} 2n^{-1} \hat{\mathbf{u}}_{n,s}^\top \mathbf{A}_{n,1} \hat{\mathbf{u}}_n & -n^{-1} \hat{\mathbf{u}}_{n,s}^\top \mathbf{A}_{n,1} \hat{\mathbf{u}}_{n,s} \\ n^{-1} \hat{\mathbf{u}}_{n,s}^\top (\mathbf{A}_{n,2} + \mathbf{A}_{n,2}^\top) \hat{\mathbf{u}}_n & -n^{-1} \hat{\mathbf{u}}_{n,s}^\top \mathbf{A}_{n,2} \hat{\mathbf{u}}_{n,s} \end{pmatrix} \end{aligned}$$

6.6.6 Coding FG2SLS Estimator Procedure

We begin by generating a DGP for the SAC (SARAR) model which will be used to test our function. The DGP assumes homoskedasticity, that is, ϵ_n are i.i.d $N(0, \sigma^2)$.

```
# Generate Data Generating Process
set.seed(666)
n      <- 529
rho    <- 0.6
lambda <- 0.6
W.nb2  <- cell2nb(sqrt(n), sqrt(n))
W      <- nb2mat(W.nb2)
M.nb2  <- cell2nb(sqrt(n), sqrt(n), type = "queen")
M      <- nb2mat(M.nb2)

# Exogenous variables
x1     <- rnorm(n)
```

```

x2      <- rnorm(n)
x3      <- rnorm(n)

# DGP parameters
b0      <- 0 ; b1 <- -1; b2 <- 0; b3 <- 1
sigma2  <- 1
epsilon <- rnorm(n, mean = 0, sd = sqrt(sigma2))

# Simulate the dependent variable
u <- solve(diag(n) - lambda * M) %*% epsilon
y <- solve(diag(n) - rho * W) %*% (b0 + b1 * x1 + b2 * x2 + b3 * x3 + u)

# Data as data.frame
data <- as.data.frame(cbind(y, x1, x2, x3))
names(data) <- c("y", "x1", "x2", "x3")

```

We generate several functions. The next function, `moments.kp`, generates the moment function $\mathbf{m}_n(\lambda_n)$ given in Equation (6.81). This function is needed later for the optimization of the GMM estimator of λ_0 .

```

# This function provides the KP's 2010 moments conditions
moments.kp <- function(lambda, u, M, A1, A2){
  n      <- length(u)
  I      <- diag(n)
  R      <- I - lambda * M
  epsi   <- R %*% u
  m.q1   <- crossprod(epsi, A1) %*% epsi
  m.q2   <- crossprod(epsi, A2) %*% epsi
  m      <- rbind(m.q1, m.q2)
  return(m)
}

```

The following function provides the objective function to be optimize: $Q_n(\lambda) = \mathbf{m}_n^\top \Upsilon_n \mathbf{m}_n$, where \mathbf{m}_n is a 2×1 vector and Υ_n is some 2×2 weighting-moment matrix.

```

# This function generates the function Q(theta) to be minimized.
Q.min <- function(lambda, u, M, A1, A2, Upsilon, verbose){
  m <- moments.kp(lambda = lambda, u = u, M = M, A1 = A1, A2 = A2)
  Q <- crossprod(m, Upsilon) %*% m / n
  if (verbose) cat("function:", Q, "lambda:", lambda, "\n")
  return(Q)
}

```

The next function `make.Upsilon` generates either $\tilde{\Psi}_n$ given in Equation (6.88) using $\hat{\mathbf{u}}_{n,\text{S2SLS}}$ or $\hat{\Psi}_n$ in Equation (6.89) using $\hat{\mathbf{u}}_{n,\text{GS2SLS}}$, based on heteroskedastic error terms.

```

# This function generates the VC matrix for
make.Upsilon <- function(lambda, HH.i, H, Z, A1, A2, u, M,
                          step = c("first", "second")){
  step <- match.arg(step)
  n <- nrow(H)
  A1s <- A1 + t(A1)
  A2s <- A2 + t(A2)
  R <- diag(n) - lambda * M
  Zs <- R %%% Z
  us <- R %%% u
  alpha1 <- - (1 / n) * (crossprod(Zs, A1s) %%% us)
  alpha2 <- - (1 / n) * (crossprod(Zs, A2s) %%% us)
  if (step == "first"){
    # Based on 2SLS residuals
    HZ <- crossprod(H, Z) / n
    P <- HH.i %%% HZ %%% solve(t(HZ) %%% HH.i %%% HZ)
    R.inv <- solve(t(R))
    a1 <- crossprod(R.inv, H) %%% P %%% alpha1
    a2 <- crossprod(R.inv, H) %%% P %%% alpha2
  } else {
    # Based on G2SLS residuals
    HZ <- crossprod(H, Zs) / n
    P <- HH.i %%% HZ %%% solve(t(HZ) %%% HH.i %%% HZ)
    a1 <- H %%% P %%% alpha1
    a2 <- H %%% P %%% alpha2
  }
  epsi <- R %%% u
  Sigma <- diag(drop(epsi^2))
  Psi.11 <- (1 / (2 * n)) * tr(A1s %%% Sigma %%% A1s %%% Sigma) +
    (1 / n) * crossprod(a1, Sigma) %%% a1
  Psi.12 <- (1 / (2 * n)) * tr(A1s %%% Sigma %%% A2s %%% Sigma) +
    (1 / n) * crossprod(a1, Sigma) %%% a2
  Psi.22 <- (1 / (2 * n)) * tr(A2s %%% Sigma %%% A2s %%% Sigma) +
    (1 / n) * crossprod(a2, Sigma) %%% a2
  Upsi <- matrix(0, nrow = 2, ncol = 2)
  Upsi[1, 1] <- Psi.11
  Upsi[1, 2] <- Upsi[2, 1] <- Psi.12
  Upsi[2, 2] <- Psi.22
  out <- list(Upsi = Upsi, a1 = a1, a2 = a2, epsi = epsi, P = P, Sigma = Sigma)
  return(out)
}

```

The function `make.G` generates the function \hat{G}_n given in Equation (6.85).

```

# This function generates matrix G
make.G <- function(u, M, A1, A2){
  n <- length(u)

```



```

u.l      <- M %*% u
G        <- matrix(0, 2, 2)
G[1, 1] <- 2 * t(u.l) %*% (A1 + t(A1)) %*% u
G[1, 2] <- -1 * t(u.l) %*% A1 %*% u.l
G[2, 1] <- t(u.l) %*% (A2 + t(A2)) %*% u
G[2, 2] <- -1 * t(u.l) %*% A2 %*% u.l
G <- G / n
return(G)
}

```

The main function to estimate the GS2SLS estimator for the SAC model is the following:

```

gstsls.sac <- function(formula, data, W, M = NULL, instruments = 2,
                       verbose = FALSE){
  # Model Frame
  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())

  # Get variables and globals
  y <- model.response(mf)
  X <- model.matrix(formula, mf)
  n <- nrow(X)

  # Check spatial weight matrices
  if (!is.null(M)){
    if(identical(M, W)) {
      M <- W
      has.M <- FALSE
    } else {
      has.M <- TRUE
    }
  } else {
    M <- W
    has.M <- FALSE
  }

  #----- #
  # Step 1a: S2SLS estimator
  #----- #
  # Make instruments
  WXI <- make.H(W = W, X = X, l = instruments)
  H    <- cbind(X, WXI)

```

```

if(has.M){
  H <- cbind(H, M %*% X[, -1], M %*% WXI)
}

# Compute S2SLSE
Wy      <- W %*% y
Z        <- cbind(X, Wy)
colnames(Z) <- c(colnames(X), "Wy")
HH        <- crossprod(H)
HH.i      <- solve(HH)
PH        <- H %*% tcrossprod(HH.i, H)
Z_hat     <- PH %*% Z
b_2sls    <- solve(crossprod(Z_hat)) %*% crossprod(Z_hat, y)
y_hat     <- Z %*% b_2sls
u_2sls    <- y - y_hat

#----- #
# Step 1b: Initial GMM estimator of lambda
#----- #
# Initial value for optimization
init.1b <- crossprod(M %*% u_2sls, u_2sls) / crossprod(u_2sls)

# Optimization
Upsilon.1b <- diag(2)
MM          <- crossprod(M)
A1.1b       <- MM - diag(diag(MM))
A2.1b       <- M
opt.1b <- nlminb(start      = init.1b,
                 objective  = Q.min,
                 u          = u_2sls,
                 M          = M,
                 A1         = A1.1b,
                 A2         = A2.1b,
                 Upsilon    = Upsilon.1b,
                 verbose    = verbose,
                 lower      = -0.9 + .Machine$double.eps,
                 upper      = 0.9 - .Machine$double.eps)

#----- #
# Step 1c: Efficient GMM estimator of lambda
#----- #

# Initial value for optimization
init.1c <- opt.1b$par

# Optimization

```

```

Upsilon.1c <- make.Upsilon(lambda = init.1c,
                          HH.i = HH.i,
                          H = H,
                          Z = Z,
                          A1 = A1.1b,
                          A2 = A2.1b,
                          u = u_2sls,
                          M = M,
                          step = "first")$Upsi
opt.1c <- nlminb(start = init.1c,
                objective = Q.min,
                u = u_2sls,
                M = M,
                A1 = A1.1b,
                A2 = A2.1b,
                Upsilon = solve(Upsilon.1c),
                verbose = verbose,
                lower = -0.9 + .Machine$double.eps,
                upper = 0.9 - .Machine$double.eps)

#----- #
# Step 2a: GS2SLS
#----- #
lambda.hat <- opt.1c$par
ys <- y - lambda.hat * M %*% y
Zs <- Z - lambda.hat * M %*% Z

# Compute S2SLSE
Zs_hat <- PH %*% Zs
b_g2sls <- solve(crossprod(Zs_hat)) %*% crossprod(Zs_hat, ys)
y_hat <- Z %*% b_g2sls
u_g2sls <- y - y_hat

#----- #
# Step 2b: Efficient GMM
#----- #
# Initial value for optimization
init.2b <- lambda.hat

# Optimization
Upsilon.2b <- make.Upsilon(lambda = init.2b,
                          HH.i = HH.i,
                          H = H,
                          Z = Z,
                          A1 = A1.1b,
                          A2 = A2.1b,

```

```

                                u = u_g2sls,
                                M = M,
                                step = "second")$Upsi
opt.2b <- nlminb(start          = init.2b,
                 objective     = Q.min,
                 u              = u_g2sls,
                 M              = M,
                 A1             = A1.1b,
                 A2             = A2.1b,
                 Upsilon       = solve(Upsilon.2b),
                 verbose        = verbose,
                 lower          = -0.9 + .Machine$double.eps,
                 upper          = 0.9 - .Machine$double.eps)

b.hat.f <- c(b_g2sls, opt.2b$par)
names(b.hat.f) <- c(rownames(b_g2sls), "Wu")

# Generate matrix G
G <- make.G(u = u_g2sls, M = M, A1 = A1.1b, A2 = A2.1b)

# Save results
results <- structure(
  list(
    coefficients = b.hat.f,
    call         = callT,
    X             = X,
    H             = H,
    HH.i          = HH.i,
    Z             = Z,
    y             = y,
    W             = W,
    M             = M,
    A1            = A1.1b,
    A2            = A2.1b,
    G             = G
  ),
  class = 'mygs2sls'
)
}

```

The `vcov` function is based on Equation (6.90).

```

vcov.mygs2sls <- function(object, ...){
  theta.hat <- object$coefficients
  Z          <- object$Z
  k          <- ncol(Z)
  lambda.hat <- theta.hat[k + 1]

```

```

delta.hat <- theta.hat[1:k]
H <- object$H
HH.i <- object$HH.i
A1 <- object$A1
A2 <- object$A2
M <- object$M
y <- object$y
p <- ncol(H)
u_g2sls <- y - Z %*% delta.hat
fitL <- make.Upsilon(lambda = lambda.hat,
                     HH.i = HH.i,
                     H = H,
                     Z = Z,
                     A1 = A1,
                     A2 = A2,
                     u = u_g2sls,
                     M = M,
                     step = "second")

## Make Omega
Psi.rr <- (1 / n) * crossprod(H, fitL$Sigma) %*% H
Psi.rd <- (1 / n) * crossprod(H, fitL$Sigma) %*% cbind(fitL$a1, fitL$a2)
Psi.f10 <- cbind(Psi.rr, Psi.rd)
Psi <- fitL$Upsi
Psi.f20 <- cbind(t(Psi.rd), Psi)
Psi.0 <- rbind(Psi.f10, Psi.f20)
Psi.inv <- solve(fitL$Upsi)
G <- object$G
J <- G %*% rbind(1, 2 * lambda.hat)
bread.lower <- solve(t(J) %*% Psi.inv %*% J) %*% t(J) %*% Psi.inv
bread <- matrix(0, nrow = (k + 1), ncol = (p + 2))
bread[1:k, 1:p] <- t(fitL$P)
bread[k + 1, (p + 1):(p + 2)] <- bread.lower
Omega <- bread %*% Psi.0 %*% t(bread) / n
return(Omega)
}

```

The S3 Methods are the following:

```

summary.mygs2sls <- function(object,
                             table = TRUE,
                             digits = max(3, .Options$digits - 3),
                             ...){
  n <- nrow(object$Z)
  df <- n - (ncol(object$Z) + 1)
  b <- object$coefficients
  std.err <- sqrt(diag(vcov(object)))
  z <- b / std.err

```

```

p      <- 2 * pt(-abs(z), df = df)
CoefTable <- cbind(b, std.err, z, p)
colnames(CoefTable) <- c("Estimate", "Std.Error", "t-value", "Pr(>|t|)")
result <- structure(
  list(
    CoefTable = CoefTable,
    digits    = digits,
    call      = object$call),
  class = 'summary.mygs2sls'
)
return(result)
}

print.summary.mygs2sls <- function(x,
                                   digits = x$digits,
                                   na.print = "",
                                   symbolic.cor = p > 4,
                                   signif.stars = getOption("show.signif.stars"),
                                   ...)
{
  cat("\nCall:\n")
  cat(paste(deparse(x$call), sep = "\n", collapse = "\n"), "\n\n", sep = "")

  cat("\nCoefficients:\n")
  printCoefmat(x$CoefTable, digit = digits, P.value = TRUE, has.Pvalue = TRUE)
  invisible(NULL)
}

```

Now, we test our function and compare the results with those obtained using `spregr` function from **sphet** package.

```

out <- gstspls.sac(y ~ x1 + x2 + x3,
                  data = data,
                  M = M,
                  W = W)
summary(out)

##
## Call:
## gstspls.sac(formula = y ~ x1 + x2 + x3, data = data, W = W, M = M)
##
##
## Coefficients:
##              Estimate Std.Error t-value Pr(>|t|)
## (Intercept)  0.05118   0.06762   0.757   0.4495
## x1          -1.00304   0.04398 -22.805  <2e-16 ***
## x2           0.07097   0.04275   1.660   0.0975 .

```

```
## x3          0.98948    0.04216   23.468    <2e-16 ***
## Wy          0.71231    0.04321   16.485    <2e-16 ***
## Wu          0.35028    0.03940    8.890    <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
test1 <- spreg(y ~ x1 + x2 + x3,
              data = data,
              listw = mat2listw(W, style = "W"),
              listw2 = mat2listw(M, style = "W"),
              model = "sarar",
              het = TRUE,
              step1.c = TRUE)

## Error in dimnames(x) <- dn: length of 'dimnames' [1] not equal to array extent

summary(test1)

## Error in h(simpleError(msg, call)): error in evaluating the argument 'object'
in selecting a method for function 'summary': object 'test1' not found
```

6.7 GMM Estimator for the SAC model

Consider the SAC model

$$\begin{aligned} \mathbf{y}_n &= \rho_0 \mathbf{W}_n \mathbf{y}_n + \mathbf{X}_n \boldsymbol{\beta}_0 + \mathbf{u}_n, \\ \mathbf{u}_n &= \lambda_0 \mathbf{M}_n \mathbf{u}_n + \boldsymbol{\varepsilon}_n, \end{aligned}$$

where $\varepsilon_{ni}, i = 1, \dots, n$ of $\boldsymbol{\varepsilon}_n$ are i.i.d. $(0, \sigma_0^2)$ with zero mean and variance σ_0^2 . Let $\boldsymbol{\theta} = (\lambda, \rho, \boldsymbol{\beta}^\top)^\top$, and denote at the true $\boldsymbol{\theta} = \boldsymbol{\theta}_0$

$$\begin{aligned} \mathbf{S}_0 &= \mathbf{I}_n - \rho_0 \mathbf{W}_n, \\ \mathbf{R}_0 &= \mathbf{I}_n - \lambda_0 \mathbf{M}_n, \\ \boldsymbol{\varepsilon}_n &= \mathbf{R}_0 \mathbf{u}_n. \end{aligned}$$

The reduced form equation of this model can be written as

$$\begin{aligned} \mathbf{y}_n &= \mathbf{S}_0^{-1} \mathbf{X}_n \boldsymbol{\beta}_0 + \mathbf{S}_0^{-1} \mathbf{R}_0^{-1} \boldsymbol{\varepsilon}_n, \\ &= \mathbf{S}_0^{-1} \mathbf{X}_n \boldsymbol{\beta}_0 + \mathbf{F}_0^{-1} \boldsymbol{\varepsilon}_n, \end{aligned}$$

where $\mathbf{F}_0 = \mathbf{R}_0 \mathbf{S}_0$. The empirical moment conditions are

$$\mathbf{g}_n(\boldsymbol{\theta}) = \begin{pmatrix} (\mathbf{P}_{1n} \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}))^\top \\ \vdots \\ (\mathbf{P}_{mn} \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}))^\top \\ \mathbf{H}_n^\top \end{pmatrix} \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) = \begin{pmatrix} (\mathbf{P}_{1n} \mathbf{R}_n(\lambda) \mathbf{u}_n(\boldsymbol{\theta}))^\top \\ \vdots \\ (\mathbf{P}_{mn} \mathbf{R}_n(\lambda) \mathbf{u}_n(\boldsymbol{\theta}))^\top \\ \mathbf{H}_n^\top \end{pmatrix} \mathbf{R}_n(\lambda) \mathbf{u}_n(\boldsymbol{\theta}),$$

where $\mathbf{P}_{jn}, j = 1, \dots, m$ are from \mathcal{P}_{1n} , the IV matrix \mathbf{H}_n is constructed from $\mathbf{X}_n, \mathbf{W}_n$ and \mathbf{M}_n .

6.8 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 Asymptotic Distribution of GMME for SEM Model

The error term for the SEM model can be expressed as:

$$\begin{aligned} \boldsymbol{\varepsilon}_n(\boldsymbol{\theta}) &= (\mathbf{I}_n - \lambda \mathbf{M}_n) \mathbf{u}_n, \\ &= \mathbf{R}_n (\mathbf{y}_n - \mathbf{X}\boldsymbol{\beta}), \\ &= \mathbf{R}_n (\mathbf{X}_n \boldsymbol{\beta}_0 + \mathbf{R}_0^{-1} \boldsymbol{\varepsilon}_n - \mathbf{X}\boldsymbol{\beta}), \quad \text{replacing } \mathbf{y}_n \\ &= \mathbf{R}_n \mathbf{X}_n \boldsymbol{\beta}_0 + \mathbf{R}_n \mathbf{R}_0^{-1} \boldsymbol{\varepsilon}_n - \mathbf{R}_n \mathbf{X}_n \boldsymbol{\beta}, \\ &= \mathbf{R}_n \mathbf{X}_n (\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + \mathbf{R}_n \mathbf{R}_0^{-1} \boldsymbol{\varepsilon}_n, \\ &= \mathbf{R}_n \mathbf{X}_n (\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + (\mathbf{I}_n + (\lambda_0 - \lambda) \mathbf{Q}_0) \boldsymbol{\varepsilon}_n, \quad \text{using the result in (6.45)} \\ &= \mathbf{R}_n \mathbf{d}_n(\boldsymbol{\beta}) + (\mathbf{I}_n + (\lambda_0 - \lambda) \mathbf{Q}_0) \boldsymbol{\varepsilon}_n, \end{aligned}$$

where in this case $\mathbf{Q}_0 = \mathbf{M}_n \mathbf{R}_0^{-1}$.

We next find \mathbf{D}_n . For the first element of the Jacobian, we note that $\mathbf{R}_n = \mathbf{R}_0 + (\lambda_0 - \lambda) \mathbf{W}_n$:

$$\frac{1}{n} \mathbf{X}_n^\top \mathbf{R}_n \mathbf{X}_n \xrightarrow{p} \frac{1}{n} \mathbf{X}_n^\top \mathbf{R}_0 \mathbf{X}_n + \frac{1}{n} (\lambda_0 - \lambda) \mathbf{X}_n^\top \mathbf{M}_n \mathbf{X}_n.$$

In addition,

$$\begin{aligned}
\frac{1}{n} \mathbf{X}_n^\top \mathbf{M}_n \mathbf{u}_n &= \frac{1}{n} \mathbf{X}_n^\top \mathbf{M}_n \mathbf{R}_0^{-1} \boldsymbol{\varepsilon}_n, \\
&= \frac{1}{n} \mathbf{X}_n^\top \mathbf{M}_n \mathbf{R}_0^{-1} [\mathbf{R}_n \mathbf{X}_n (\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + (\mathbf{I}_n + (\lambda_0 - \lambda) \mathbf{Q}_0) \boldsymbol{\varepsilon}_n] \\
&= \frac{1}{n} \mathbf{X}_n^\top \mathbf{M}_n \mathbf{R}_0^{-1} \mathbf{R}_n \mathbf{X}_n (\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + \frac{1}{n} (\lambda_0 - \lambda) \mathbf{X}_n^\top \mathbf{M}_n \mathbf{R}_0^{-1} \mathbf{Q}_0 \boldsymbol{\varepsilon}_n, \\
&\xrightarrow{p} \frac{1}{n} \mathbf{X}_n^\top \mathbf{M}_n \mathbf{R}_0^{-1} \mathbf{R}_n \mathbf{X}_n (\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + \frac{1}{n} (\lambda_0 - \lambda) \mathbf{X}_n^\top \mathbf{M}_n \mathbf{R}_0^{-1} \mathbf{Q}_0 \mathbb{E}(\boldsymbol{\varepsilon}_n), \\
&\xrightarrow{p} \frac{1}{n} \mathbf{X}_n^\top \mathbf{M}_n \mathbf{R}_0^{-1} \mathbf{R}_n \mathbf{X}_n (\boldsymbol{\beta}_0 - \boldsymbol{\beta}).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{jn}^s \mathbf{M}_n \mathbf{u}_n &= \frac{1}{n} (\mathbf{R}_n \mathbf{X}_n (\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + \mathbf{R}_n \mathbf{R}_0^{-1} \boldsymbol{\varepsilon}_n)^\top \mathbf{P}_{jn}^s \mathbf{M}_n \mathbf{u}_n, \\
&= \frac{1}{n} (\boldsymbol{\beta}_0 - \boldsymbol{\beta})^\top \mathbf{X}_n^\top \mathbf{R}_n^\top \mathbf{P}_{jn}^s \mathbf{M}_n \mathbf{u}_n + \frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{R}_0^{-1\top} \mathbf{R}_n \mathbf{P}_{jn}^s \mathbf{M}_n \mathbf{u}_n.
\end{aligned} \tag{6.91}$$

For the first element

$$\begin{aligned}
\frac{1}{n} (\boldsymbol{\beta}_0 - \boldsymbol{\beta})^\top \mathbf{X}_n^\top \mathbf{R}_n^\top \mathbf{P}_{jn}^s \mathbf{M}_n \mathbf{u}_n &= \frac{1}{n} (\boldsymbol{\beta}_0 - \boldsymbol{\beta})^\top \mathbf{X}_n^\top \mathbf{R}_n^\top \mathbf{P}_{jn}^s \mathbf{M}_n \mathbf{R}_0^{-1} \boldsymbol{\varepsilon}_n, \\
&\xrightarrow{p} \frac{1}{n} (\boldsymbol{\beta}_0 - \boldsymbol{\beta})^\top \mathbf{X}_n^\top \mathbf{R}_n^\top \mathbf{P}_{jn}^s \mathbf{M}_n \mathbf{R}_0^{-1} \mathbb{E}(\boldsymbol{\varepsilon}_n), \\
&\xrightarrow{p} \mathbf{0}.
\end{aligned}$$

For the second element

$$\begin{aligned}
\frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{R}_0^{-1\top} \mathbf{R}_n \mathbf{P}_{jn}^s \mathbf{M}_n \mathbf{u}_n &= \frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{R}_0^{-1\top} \mathbf{R}_n \mathbf{P}_{jn}^s \mathbf{M}_n \mathbf{R}_0^{-1} \boldsymbol{\varepsilon}_n, \\
&= \frac{1}{n} \boldsymbol{\varepsilon}_n^\top [\mathbf{I}_n + (\lambda_0 - \lambda) \mathbf{Q}_0]^\top \mathbf{P}_{jn}^s \mathbf{M}_n \mathbf{R}_0^{-1} \boldsymbol{\varepsilon}_n, \\
&= \frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{jn}^s \mathbf{Q}_0 \boldsymbol{\varepsilon}_n + (\lambda_0 - \lambda) \frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{Q}_0^\top \mathbf{P}_{jn}^s \mathbf{Q}_0 \boldsymbol{\varepsilon}_n, \\
&\xrightarrow{p} \frac{1}{n} \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{P}_{jn}^s \mathbf{Q}_0 \boldsymbol{\varepsilon}_n) + \frac{1}{n} (\lambda_0 - \lambda) \mathbb{E}(\boldsymbol{\varepsilon}_n^\top \mathbf{Q}_0^\top \mathbf{P}_{jn}^s \mathbf{Q}_0 \boldsymbol{\varepsilon}_n), \\
&\xrightarrow{p} \sigma_0^2 \text{tr}(\mathbf{P}_{jn}^s \mathbf{Q}_0) + (\lambda_0 - \lambda) \frac{\sigma_0^2}{n} \text{tr}(\mathbf{Q}_0^\top \mathbf{P}_{jn}^s \mathbf{G}_0).
\end{aligned}$$

Now:

$$\begin{aligned}
\frac{1}{n} \boldsymbol{\varepsilon}_n^\top \mathbf{P}_{jn}^s \mathbf{R}_n \mathbf{X}_n &= \frac{1}{n} (\mathbf{R}_n \mathbf{X}_n (\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + \mathbf{R}_n \mathbf{R}_0^{-1} \boldsymbol{\varepsilon}_n)^\top \mathbf{P}_{jn}^s \mathbf{R}_n \mathbf{X}_n, \\
&= \frac{1}{n} (\boldsymbol{\beta}_0 - \boldsymbol{\beta})^\top \mathbf{X}_n^\top \mathbf{R}_n^\top \mathbf{P}_{jn}^s \mathbf{R}_n \mathbf{X}_n + \frac{1}{n} \boldsymbol{\varepsilon}_n^\top (\mathbf{R}_n \mathbf{R}_0^{-1})^\top \mathbf{P}_{jn}^s \mathbf{R}_n \mathbf{X}_n, \\
&= \frac{1}{n} (\boldsymbol{\beta}_0 - \boldsymbol{\beta})^\top \mathbf{X}_n^\top \mathbf{R}_n^\top \mathbf{P}_{jn}^s \mathbf{R}_n \mathbf{X}_n + \frac{1}{n} \boldsymbol{\varepsilon}_n^\top (\mathbf{I}_n + (\lambda_0 - \lambda) \mathbf{Q}_0)^\top \mathbf{P}_{jn}^s \mathbf{R}_n \mathbf{X}_n, \\
&\xrightarrow{p} \frac{1}{n} (\boldsymbol{\beta}_0 - \boldsymbol{\beta})^\top \mathbf{X}_n^\top \mathbf{R}_n^\top \mathbf{P}_{jn}^s \mathbf{R}_n \mathbf{X}_n.
\end{aligned}$$

As a result,

$$\frac{1}{n} \frac{\partial \mathbb{E}[\mathbf{g}_n(\boldsymbol{\theta})]}{\partial \boldsymbol{\theta}^\top} = \begin{pmatrix} \frac{1}{n} \mathbf{X}_n^\top \mathbf{R}_0 \mathbf{X}_n + \frac{1}{n} (\lambda_0 - \lambda) \mathbf{X}_n^\top \mathbf{M}_n \mathbf{X}_n^\top & \frac{1}{n} \mathbf{X}_n^\top \mathbf{M}_n \mathbf{R}_0^{-1} \mathbf{R}_n \mathbf{X}_n (\boldsymbol{\beta}_0 - \boldsymbol{\beta}) \\ \frac{1}{n} (\boldsymbol{\beta}_0 - \boldsymbol{\beta})^\top \mathbf{X}_n^\top \mathbf{R}_n^\top \mathbf{P}_{1n}^\top \mathbf{R}_n \mathbf{X}_n & \sigma_0^2 \text{tr}(\mathbf{P}_{jn}^s \mathbf{Q}_0) + (\lambda_0 - \lambda) \frac{\sigma_0^2}{n} \text{tr}(\mathbf{G}_0^\top \mathbf{P}_{jn}^s \mathbf{Q}_0) \\ \frac{1}{n} (\boldsymbol{\beta}_0 - \boldsymbol{\beta})^\top \mathbf{X}_n^\top \mathbf{R}_n^\top \mathbf{P}_{2n}^\top \mathbf{R}_n \mathbf{X}_n & \sigma_0^2 \text{tr}(\mathbf{P}_{jn}^s \mathbf{Q}_0) + (\lambda_0 - \lambda) \frac{\sigma_0^2}{n} \text{tr}(\mathbf{G}_0^\top \mathbf{P}_{jn}^s \mathbf{Q}_0) \end{pmatrix} \quad (6.92)$$

At $\boldsymbol{\theta} = \boldsymbol{\theta}_0$,

$$\mathbf{D}_n = \begin{pmatrix} \mathbf{X}_n^\top \mathbf{R}_0 \mathbf{X}_n & \mathbf{0} \\ k \times k & k \times 1 \\ \mathbf{0} & \sigma_0^2 \text{tr}(\mathbf{P}_{1n}^s \mathbf{Q}_0) \\ 1 \times k & 1 \times 1 \\ \mathbf{0} & \sigma_0^2 \text{tr}(\mathbf{P}_{2n}^s \mathbf{Q}_0) \\ 1 \times k & 1 \times 1 \end{pmatrix}_{(k+2) \times (k+1)}$$

6.B Proof Theorem 3 in KP 1998

Recall that the GS2SLS is given by:

$$\hat{\boldsymbol{\delta}}_n = \left[\hat{\mathbf{Z}}_s(\lambda)^\top \hat{\mathbf{Z}}_s(\lambda) \right]^{-1} \hat{\mathbf{Z}}_s(\lambda)^\top \mathbf{y}_s(\lambda) \quad (6.93)$$

Whereas, the FGS2SLS is given by:

$$\hat{\boldsymbol{\delta}}_{F,n} = \left[\hat{\mathbf{Z}}_s(\hat{\lambda})^\top \hat{\mathbf{Z}}_s(\hat{\lambda}) \right]^{-1} \hat{\mathbf{Z}}_s(\hat{\lambda})^\top \mathbf{y}_s(\hat{\lambda}) \quad (6.94)$$

where

$$\begin{aligned} \hat{\mathbf{Z}}_s(\hat{\lambda}_n) &= \mathbf{P}_{H_n} \mathbf{Z}_s(\hat{\lambda}_n) \\ \mathbf{Z}_s(\hat{\lambda}_n) &= \mathbf{Z}_n - \hat{\lambda}_n \mathbf{M}_n \mathbf{Z}_n \\ \mathbf{y}_s(\hat{\lambda}_n) &= \mathbf{y}_n - \hat{\lambda}_n \mathbf{M}_n \mathbf{y}_n \\ \hat{\mathbf{Z}}_s(\hat{\lambda}_n) &= \left(\mathbf{X}_n - \hat{\lambda}_n \mathbf{M}_n \mathbf{X}_n, \mathbf{W}_n \mathbf{y}_n - \hat{\lambda}_n \mathbf{M}_n \mathbf{W}_n \mathbf{y}_n \right) \\ \widehat{\lambda_n \mathbf{M}_n \mathbf{W}_n \mathbf{y}_n} &= \mathbf{P}_{H_n} \left(\mathbf{W}_n \mathbf{y}_n - \hat{\lambda}_n \mathbf{M}_n \mathbf{W}_n \mathbf{y}_n \right). \end{aligned} \quad (6.95)$$

The sampling error is:

$$\hat{\boldsymbol{\delta}}_{F,n} - \boldsymbol{\delta} = \left[\hat{\mathbf{Z}}_s(\hat{\lambda})^\top \hat{\mathbf{Z}}_s(\hat{\lambda}) \right]^{-1} \hat{\mathbf{Z}}_s(\hat{\lambda})^\top \mathbf{u}_s(\hat{\lambda}_n) \quad (6.96)$$

where:

$$\begin{aligned} \mathbf{u}_s(\hat{\lambda}_n) &= (\mathbf{I} - \hat{\lambda}_n) \mathbf{u} \\ &= (\mathbf{I} - \hat{\lambda}_n) \mathbf{u} + \boldsymbol{\varepsilon}_n - \boldsymbol{\varepsilon}_n \\ &= \boldsymbol{\varepsilon}_n + (\mathbf{I} - \hat{\lambda}_n \mathbf{M}_n) \mathbf{u} - (\mathbf{I} - \lambda \mathbf{M}_n) \mathbf{u} \\ &= \boldsymbol{\varepsilon}_n + \mathbf{u} - \hat{\lambda}_n \mathbf{M}_n \mathbf{u} - \mathbf{u} + \lambda \mathbf{M}_n \mathbf{u} \\ &= \boldsymbol{\varepsilon}_n - (\hat{\lambda}_n - \lambda) \mathbf{M}_n \mathbf{u}_n \end{aligned} \quad (6.97)$$

Then:

$$\begin{aligned}
 \hat{\delta}_{F,n} - \delta &= \left[\hat{\mathbf{Z}}_s(\hat{\lambda})^\top \hat{\mathbf{Z}}_s(\hat{\lambda}) \right]^{-1} \hat{\mathbf{Z}}_s(\hat{\lambda})^\top \left[\boldsymbol{\varepsilon}_n - (\hat{\lambda}_n - \lambda) \mathbf{M}_n \mathbf{u}_n \right] \\
 &= \left[\hat{\mathbf{Z}}_s(\hat{\lambda})^\top \hat{\mathbf{Z}}_s(\hat{\lambda}) \right]^{-1} \hat{\mathbf{Z}}_s(\hat{\lambda})^\top \boldsymbol{\varepsilon}_n - \left[\hat{\mathbf{Z}}_s(\hat{\lambda})^\top \hat{\mathbf{Z}}_s(\hat{\lambda}) \right]^{-1} \hat{\mathbf{Z}}_s(\hat{\lambda})^\top (\hat{\lambda}_n - \lambda) \mathbf{M}_n \mathbf{u}_n \\
 &= \left[\frac{1}{n} \hat{\mathbf{Z}}_s(\hat{\lambda})^\top \hat{\mathbf{Z}}_s(\hat{\lambda}) \right]^{-1} \frac{1}{n} \hat{\mathbf{Z}}_s(\hat{\lambda})^\top \boldsymbol{\varepsilon}_n - \left[\frac{1}{n} \hat{\mathbf{Z}}_s(\hat{\lambda})^\top \hat{\mathbf{Z}}_s(\hat{\lambda}) \right]^{-1} (\hat{\lambda}_n - \lambda) \frac{1}{n} \hat{\mathbf{Z}}_s(\hat{\lambda})^\top \mathbf{M}_n \mathbf{u}_n \\
 \sqrt{n}(\hat{\delta}_{F,n} - \delta) &= \left[\frac{1}{n} \hat{\mathbf{Z}}_s(\hat{\lambda})^\top \hat{\mathbf{Z}}_s(\hat{\lambda}) \right]^{-1} \frac{1}{\sqrt{n}} \hat{\mathbf{Z}}_s(\hat{\lambda})^\top \boldsymbol{\varepsilon}_n - \left[\frac{1}{n} \hat{\mathbf{Z}}_s(\hat{\lambda})^\top \hat{\mathbf{Z}}_s(\hat{\lambda}) \right]^{-1} (\hat{\lambda}_n - \lambda) \frac{1}{\sqrt{n}} \hat{\mathbf{Z}}_s(\hat{\lambda})^\top \mathbf{M}_n \mathbf{u}_n
 \end{aligned} \tag{6.98}$$

By consistency $\hat{\lambda}_n - \lambda = o_p(1)$. Now, we need to show that:

$$\frac{1}{n} \hat{\mathbf{Z}}_s(\hat{\lambda})^\top \hat{\mathbf{Z}}_s(\hat{\lambda}) \xrightarrow{p} \frac{1}{n} \hat{\mathbf{Z}}_s(\lambda)^\top \hat{\mathbf{Z}}_s(\lambda) = \bar{\mathbf{Q}} \tag{6.99}$$

$$\frac{1}{\sqrt{n}} \hat{\mathbf{Z}}_s(\hat{\lambda})^\top \boldsymbol{\varepsilon}_n \xrightarrow{d} \mathbf{N}(\mathbf{0}, \sigma_\varepsilon^2 \bar{\mathbf{Q}}), \tag{6.100}$$

$$(\hat{\lambda}_n - \lambda) \frac{1}{\sqrt{n}} \hat{\mathbf{Z}}_s(\hat{\lambda})^\top \mathbf{M}_n \mathbf{u}_n \xrightarrow{p} 0 \tag{6.101}$$

where:

$$\bar{\mathbf{Q}} = [\mathbf{Q}_{HZ} - \lambda \mathbf{Q}_{mHZ}]^\top \mathbf{Q}_{HH}^{-1} [\mathbf{Q}_{HZ} - \lambda \mathbf{Q}_{mHZ}] \tag{6.102}$$

is finite and nonsingular. For 6.99, note that:

$$\begin{aligned}
 \frac{1}{n} \hat{\mathbf{Z}}_s(\hat{\lambda})^\top \hat{\mathbf{Z}}_s(\hat{\lambda}) &= \frac{1}{n} \left(\mathbf{Z}_n - \hat{\lambda}_n \mathbf{M}_n \mathbf{Z}_n \right)^\top \mathbf{P}_{H_n} \left(\mathbf{Z}_n - \hat{\lambda}_n \mathbf{M}_n \mathbf{Z}_n \right) \\
 &= \frac{1}{n} \left(\mathbf{Z}_n - \hat{\lambda}_n \mathbf{M}_n \mathbf{Z}_n \right)^\top \mathbf{H}_n (\mathbf{H}_n^\top \mathbf{H}_n)^{-1} \mathbf{H}_n^\top \left(\mathbf{Z}_n - \hat{\lambda}_n \mathbf{M}_n \mathbf{Z}_n \right) \\
 &= \left(\underbrace{\frac{1}{n} \mathbf{Z}_n^\top \mathbf{H}_n}_{\xrightarrow{p} \mathbf{Q}_{HZ}^\top} - \underbrace{\hat{\lambda}_n}_{\xrightarrow{p} \lambda} \underbrace{\frac{1}{n} \mathbf{Z}_n^\top \mathbf{M}_n^\top \mathbf{H}_n}_{\xrightarrow{p} \mathbf{Q}_{mHZ}^\top} \right) \underbrace{\left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right)^{-1}}_{\rightarrow \mathbf{Q}_{HH}^{-1}} \underbrace{\left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{Z}_n - \hat{\lambda}_n \frac{1}{n} \mathbf{H}_n^\top \mathbf{M}_n \mathbf{Z}_n \right)}_{\xrightarrow{p} \mathbf{Q}_{HZ} - \lambda \mathbf{Q}_{mHZ}}
 \end{aligned} \tag{6.103}$$

For 6.100, note that:

$$\begin{aligned}
 \frac{1}{\sqrt{n}} \hat{\mathbf{Z}}_s(\hat{\lambda})^\top \boldsymbol{\varepsilon}_n &= \frac{1}{\sqrt{n}} \left(\mathbf{Z}_n - \hat{\lambda}_n \mathbf{M}_n \mathbf{Z}_n \right)^\top \mathbf{P}_{H_n} \boldsymbol{\varepsilon} \\
 &= \left(\underbrace{\frac{1}{n} \mathbf{Z}_n^\top \mathbf{H}_n}_{\xrightarrow{p} \mathbf{Q}_{HZ}^\top} - \underbrace{\hat{\lambda}_n}_{\xrightarrow{p} \lambda} \underbrace{\frac{1}{n} \mathbf{Z}_n^\top \mathbf{M}_n^\top \mathbf{H}_n}_{\xrightarrow{p} \mathbf{Q}_{mHZ}^\top} \right) \underbrace{\left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right)^{-1}}_{\rightarrow \mathbf{Q}_{HH}^{-1}} \underbrace{\frac{1}{\sqrt{n}} \mathbf{H}_n^\top \boldsymbol{\varepsilon}_n}_{\xrightarrow{d} \mathbf{N}(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{Q}_{HH})}
 \end{aligned} \tag{6.104}$$

For 6.101 note that:

$$\begin{aligned}
 (\hat{\lambda}_n - \lambda) \frac{1}{\sqrt{n}} \hat{\mathbf{Z}}_s(\hat{\lambda})^\top \mathbf{M}_n \mathbf{u}_n &= \underbrace{(\hat{\lambda}_n - \lambda)}_{o_p(1)} \left(\underbrace{\frac{1}{n} \mathbf{Z}_n^\top \mathbf{H}_n}_{\xrightarrow{p} \mathbf{Q}_{HZ}^\top} - \underbrace{\hat{\lambda}_n}_{\xrightarrow{p} \lambda} \underbrace{\frac{1}{n} \mathbf{Z}_n^\top \mathbf{M}_n^\top \mathbf{H}_n}_{\xrightarrow{p} \mathbf{Q}_{HMZ}^\top} \right) \underbrace{\left(\frac{1}{n} \mathbf{H}_n^\top \mathbf{H}_n \right)^{-1}}_{\rightarrow \mathbf{Q}_{HH}^{-1}} \frac{1}{\sqrt{n}} \mathbf{H}_n^\top \mathbf{M}_n \mathbf{u}_n \\
 &\quad (6.105)
 \end{aligned}$$

Note that $\mathbb{E} (n^{-1/2} \mathbf{H}_n^\top \mathbf{M}_n \mathbf{u}_n) = 0$ and $\mathbb{E} (n^{-1} \mathbf{H}_n^\top \mathbf{M}_n \mathbf{u}_n \mathbf{u}_n^\top \mathbf{M}_n^\top \mathbf{H}_n^\top) = n^{-1} \mathbf{H}_n^\top \mathbf{M}_n \boldsymbol{\Sigma}_{u_n} \mathbf{M}_n^\top \mathbf{H}_n^\top$, 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

$$\sqrt{n}(\hat{\boldsymbol{\delta}}_{F,n} - \boldsymbol{\delta}) \xrightarrow{d} \text{N}(\mathbf{0}, \sigma_\epsilon^2 \bar{\mathbf{Q}}^{-1}) \quad (6.106)$$

The small sample approximation is

$$\hat{\boldsymbol{\delta}}_{F,n} \sim \text{N} \left(\boldsymbol{\delta}, \hat{\sigma}^2 \left[\hat{\mathbf{Z}}_s(\hat{\lambda})^\top \hat{\mathbf{Z}}_s(\hat{\lambda}) \right]^{-1} \right) \quad (6.107)$$

where:

$$\hat{\sigma}^2 = \hat{\boldsymbol{\varepsilon}}^\top \hat{\boldsymbol{\varepsilon}} / n \quad (6.108)$$

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. (2021). Spatial models in econometric research.
- 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 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.
- 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.
- Ertur, C. and Koch, W. (2007). Growth, technological interdependence and spatial externalities: theory and evidence. *Journal of applied econometrics*, 22(6):1033–1062.
- 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.
- Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the econometric society*, pages 1029–1054.
- Henningsen, A. and Toomet, O. (2011). maxlik: A package for maximum likelihood estimation in r. *Computational Statistics*, 26(3):443–458.
- 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. (2001). Generalized method of moments estimation of spatial autoregressive processes. *Unpublished manuscript*.

- 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.
- Lee, L.-f. and Liu, X. (2010). Efficient gmm estimation of high order spatial autoregressive models with autoregressive disturbances. *Econometric Theory*, 26(1):187–230.
- LeSage, J. and Pace, R. K. (2010). *Introduction to Spatial Econometrics*. CRC press.
- 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.
- Lin, X. and Lee, L.-f. (2010). Gmm estimation of spatial autoregressive models with unknown heteroskedasticity. *Journal of Econometrics*, 157(1):34–52.
- Liu, T., Xu, X., and Lee, L.-f. (2022). Consistency without compactness of the parameter space in spatial econometrics. *Economics Letters*, 210:110224.
- Liu, X., Lee, L.-f., and Bollinger, C. R. (2010). An efficient gmm estimator of spatial autoregressive models. *Journal of Econometrics*, 159(2):303–319.
- Liu, X. and Saraiva, P. (2015). Gmm estimation of sar models with endogenous regressors. *Regional Science and Urban Economics*, 55:68–79.
- 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.
- Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1):439–446.

- Pebesma, E. and Bivand, R. S. (2023). *Spatial Data Science With Applications in R*. Chapman & Hall.
- 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, 76
- BS2SLS, 213
- Cochrane-Orcutt transformation, 128
- consistent estimator, 80
- Convergence
 - bounded sequences, 75
 - deterministic sequences, 73
- convergence in probability, 77
- eigen values, 41
- Endogeneity
 - additional endogenous variables, 203
 - error in variables, 203
- Generalized method of moments, 191
 - Moment conditions, 247
- Hessian
 - SLM, 123
- Heteroskedasticity
 - error term, 199
- Instrumental Variables
 - instruments, 197
 - optimal instruments, 213
 - S2SLS, 197
- Leontief expansion, 42
- Maximum likelihood, 120
 - concentrated log-likelihood, 122
 - Jacobian, 120
 - SLM, 119
- Moment conditions, 247
- Moran's I test, 24
 - `moran.mc` function, 30
 - `moran.test` function, 29
 - Monte carlo, 27
 - Normality, 27
 - Randomization, 27
- Multiplier effect, 42
- Parameter space, 41
- quadratic moment conditions, 247
- Reduced form
 - Spatial lag model, 40
- S2SLS
 - Asymptotic distribution, 205
 - consistency, 204
 - example, 207
 - `stsls` function, 207
- SAC model
 - FGS2SLS, 272
- score function
 - SLM, 123
- Spatial autocorrelation, 4
- Spatial autoregressive process, 36
- Spatial dependence, 4
- Spatial durbin model, 36
- Spatial error model, 37
- Spatial lag model, 35
 - pure, 115
- Spillover effects
 - Direct effects, 51
 - example, 57

- Global spillovers, 50
- Indirect effects, 51
- Local spillovers, 50
- Marginal effects, 50

Tobler’s law, 3

Weight matrix, 6

- `coords` function, 21
- `knearneigh` function, 20
- `lag.listw` function, 24
- `nb2listw` function, 17
- `poly2nb` function, 16
- Based on distance, 9
- Bishop contiguity, 8
- Definition, 7
- Higher order, 12
- Invertibility, 41
- Minkowski, 9
- Queen contiguity, 8
- Rook contiguity, 7
- Row-standardization, 11
- Spatial lag, 12