Lecture 1: Introduction and Preliminaries

Zhenlin Yang

School of Economics, Singapore Management University
zlyang@smu.edu.sg
http://www.mysmu.edu.sg/faculty/zlyang/

ECON747: Spatial Econometric Models and Methods
Term I, 2024-25

Spatial econometrics

Spatial econometrics concerns spatial dependence among geographical units or social interaction among economic agents or social actors, e.g., neighbourhood effects, spillover, copy-catting, network, and peer effects.

- It has received an increased attention by regional scientists, economists, econometricians, and statisticians.
- Standard econometric techniques often fails in the presence of spatial interactions. Spatial econometric models and methods extend the standard ones to capture the spatial interactions.
- They have been applied not only in specialized fields such as regional science, urban economics, real estate and economic geography,
- but also increasingly in more traditional fields of economics, including demand analysis, labour economics, public economics, international economics, agricultural and environmental economics,
- **and finance** for issues, e.g., asset pricing with spatial interaction (Kou, et al., 2017), stock market comovements, and more.

Course Contents

This course introduces common spatial econometric models, and common methods of estimation and inference:

1. spatial linear regression models,
2. spatial panel data models,
3. dynamic spatial panel data models.

These models extend the classical linear regression models, panel data models, and dynamic panel data models, by adding the following terms:

- spatial lag dependence,
- spatial Durbin effect,
- spatial error dependence,

to capture the *endogenous social effects* (Manski, 1993), *contextual effects*, and *cross-sectional* (*spatial*) dependence that results from the 'interactions' among geographical units, economic agents or social actors.

Spatial effects are modeled by **spatial weight matrices** or **connectivity matrices**, and **spatial parameters**.

For each type model,

- common inference methods, such as quasi-maximum likelihood (QML), M-estimation, and GMM, are introduced,
- empirical illustrations are presented, and
- Matlab routines are provided so that one can conduct his/her own empirical analyses by adapting/extending these codes.

In addition, refined and robust inferences methods will be introduced.

Lecture 1 introduces common spatial econometric models and common spatial weight matrices, presents some background knowledge on QML, M-, and GMM estimation methods, and discuss some popular applications of spatial models and recent developments in spatial software tools.

Brief History of Spatial Econometrics

The term *spatial econometrics* was coined by **Jean Paelinck**, a Belgium Econometrician, in the early 1970s to designate a growing body of the regional science literature that dealt primarily with estimation and testing problems encountered in the implementation of *multiregional econometric models*. See page 7 of the book: Anselin (1988).

The dependence among the cross-sectional units can be considered to to lie at the core of the disciplines of the regional science and geography, as expressed in Tobler's (1979) *first law of geography*:

> **"everything is related to everything else, but near things are more related than distant things."**

See also the books of Paelinck and Klaassen (1979), LeSage and Pace (1999), and Elhorst (2014).

While the concept of spatial dependence originated from the notion of relative space or relative location, which emphasizes the effect of distance or neighborhood, the notation of space or location can easily be extended beyond the strict Euclidean space inducing

- economic distance,
- policy distance,
- inter-personal distance,
- financial networks,
- international trade,
- social networks, etc.

Hence, spatial dependence is a phenomenon that exists in a wide range of applications in the social sciences and economics.

Spatial Linear Regression (SLR) Model

The SLR model extends the classical linear regression model:

$$Y_n = X_n\beta + u_n, \tag{1.1}$$

by adding one or more of the following terms:

1. **spatial lag** (SL) term: $\lambda W_{1n} Y_n$,
2. **spatial Durbin** (SD) term: $W_{3n} X_n^* \gamma$,
3. **spatial error** (SE) term: $u_n = \rho W_{2n} u_n + v_n$,

where

- Y_n $(n \times 1)$, X_n $(n \times p)$, β, and u_n have the usual interpretations,
- X_n^*: a submatrix of X_n excluding the column of ones, etc,
- $W_{rn}, r = 1, 2, 3$: known $n \times n$ spatial weight matrices, and
- λ, ρ and γ are the parameters corresponding to, respectively, the SL, SE, and SD effects.

Spatial Panel Data (SPD) Model

The SPD model extends the classical panel data model:

$$Y_{nt} = X_{nt}\beta + Z_n\gamma + \mu_n + \alpha_t 1_n + u_{nt}, \quad t = 1, \ldots, T, \tag{1.2}$$

by adding one or more of the following: $\lambda W_{1n} Y_{nt}$, $\rho W_{2n} u_{nt}$ and $W_{3n} X_{nt}^{*}\gamma$, being, respectively, the SL, SE and SD effects, where

- X_{nt}: $n \times p$ matrix or time-varying regressors,
- Z_n: $n \times q$ matrix of time-invariant regressors,
- μ_n is an $n \times 1$ vector of unit-specific effects,
- $\{\alpha_t\}$ are the time-specific effects,
- 1_n is an n-vector of ones, and other quantities are similarly defined.

The μ_n and $\{\alpha_t\}$ are called: (*i*) **fixed effects** if they are allowed to be correlated with the time-varying regressors X_{nt} in an arbitrary manner; (*ii*) **random effects** if they are uncorrelated with X_{nt}; or (*iii*) **correlated random effects** if they are correlated with X_{nt} linearly.

Dynamic Spatial Panel Data (DSPD) Model

The DSPD model extends the classical dynamic panel data model:

$$Y_{nt} = \rho Y_{n,t-1} + X_{nt}\beta + Z_n\gamma + \mu_n + \alpha_t 1_n + u_{nt}, \quad t = 1, \ldots, T, \tag{1.3}$$

by adding the SL, SE, or SD terms as in the SPD models. It can be further extended by adding a space-time effect $\lambda W_n Y_{n,t-1}$.

Challenges with the DSPD model:

1. it incurs the **incidental parameters problem** (number of parameters increase with the increase of sample size),
2. it gives rise to the **initial values problem** (the distribution of the initial observations Y_{n0} depends on the past values which are unobservables).

Remarks: For all the spatial econometric models, the spatial weight matrices can be the same. The SL and SE effects can go for higher order.

Spatial Weight Matrix

Spatial weight matrix $W_n = \{w_{ij}, i, j = 1, \ldots, n\}$ is an $n \times n$ matrix that measures the 'connectivity' among the n spatial units. The construction of W_n is a subject matter, depending on the actual problem considered.

Contiguity-based Weights: $w_{ij} = 1$ when units i and j are 'neighbors', and $w_{i,j} = 0$ otherwise. Popular ones include:

- `Rook, Bishop, or Queen Contiguity:` in a regular grid, the neighbors are those with common edge, or common vertex, or combination of both, in analogy with the move of Rook, Bishop or Queen in the game of chess (Anselin, 1988, p.18);
- `Circular World:` units immediate ahead or behind of a given spatial unit are considered as neighbors of this unit, e.g., "5 ahead and 5 behind" (Kelejian and Prucha, 1999, p. 520);
- `Group Interaction:` when n spatial units are naturally from R groups, the members within each group are neighbors, but the members from different groups are not (Case, 1991; Lee, 2004).

Weights based on Physical Distance: The concept of neighbors can be extended to mean areal units with (*i*) common border, (*ii*) within a given centroid distance of each other, or (*iii*) closest in terms of centroid distance, etc., resulting W_n with elements 1 or 0 as above:

- `Common Border Weights:` $w_{ij} = 1$ if units i and j share common border, and 0 otherwise;
- `K-Nearest Neighbors Weights:` $w_{ij} = 1$ if units $j(= 1, \ldots, K)$ are the K units closest to the unit i in terms of centroid distance;
- `Radial Distance Weights:` $w_{ij} = 1$ for $d_{ij} \leq \delta$, where d_{ij} is the distance between units i and j, and δ is the distance cutoff value, giving the distance-based contiguity.

Spatial weights w_{ij} can also be defined directly in terms of the actual distance d_{ij} between units i and j, or in terms of the length ℓ_{ij} of the common boarder between units i and j:

- `Power Distance Weights:` $w_{ij} = d_{ij}^{-\alpha}$, where α is any positive exponent;
- `Exponential Distance Weights:` $w_{ij} = \exp(-\alpha d_{ij})$, where α is a positive exponent;
- Shared-Boundary Weights: $w_{ij} = \ell_{ij} / \sum_{k \neq i} \ell_{ik}$, proportion of $i - j$ boundary in the total boundary of unit i.

Weights based on Social or Economic Distance: Other specification of spatial weights are possible as well (Anselin and Bera, 1998., p.244).

- In sociometrics, the weights reflect whether or not two individual belong to the same social network (Doreian, 1980).
- In **economics applications**, the use of spatial weights based on "economic" distance has been suggested, among others, in Case et al. (1993). Specifically, they suggest to use weights of the form

$$w_{ij} = 1/|x_i - x_j|,$$

  where x_i and x_j are observations on "meaningful" socioeconomic characteristics such as per capita income or percentage of the population in a given racial or ethic group.

The final form of W_n is such that $w_{ii} = 0$ (excluding self-influence), and W_n is row normalized, i.e., each element is divided by its row sum.

Possible endogeneity in spatial weight matrix.

Anselin and Bera (1998, p.244) made an important note:

> *in the standard estimation and testing problems, the weights matrix is taken to be* exogenous*. Therefore, indicators for the socioeconomic weights should be chosen with great care to ensure their exogeneity, unless their endogeneity is considered explicitly.*

Qu and Lee (2015) started research on endogenous spatial weights, where the elements w_{ij} are constructed by some other economic variables $Z_n = X_{2n}\Gamma + \varepsilon_n$ with ε_n being correlated with v_n. As such, W_n is endogenous.

Roles of Spatial Terms

Consider a spatial contiguity-based weight matrix with $n = 12$ spatial units:

$W_n =$

	1	2	3	4	5	6	7	8	9	10	11	12
1	0	0	0	0	0	1/2	1/2	0	0	0	0	0
2	0	0	0	0	0	0	1/3	1/3	1/3	0	0	0
3	0	0	0	1/4	0	1/4	1/4	1/4	0	0	0	0
4	0	0	1/3	0	1/3	0	0	0	0	1/3	0	0
5	0	0	0	1/2	0	1/2	0	0	0	0	0	0
6	1/3	0	1/3	0	1/3	0	0	0	0	0	0	0
7	1/3	1/3	1/3	0	0	0	0	0	0	0	0	0
8	0	1/4	1/4	0	0	0	0	0	0	1/4	1/4	0
9	0	1/2	0	0	0	0	0	0	0	0	1/2	0
10	0	0	0	1/3	0	0	0	1/3	0	0	0	1/3
11	0	0	0	0	0	0	0	1/3	1/3	0	0	1/3
12	0	0	0	0	0	0	0	0	0	1/2	1/2	0

SLR Model with Spatial Error: $Y_n = X_n\beta + u_n, \quad u_n = \rho W_n u_n + v_n$.

Let $\{u_i\}$ and $\{v_i\}$ be the elements of u_n and v_n, respectively. From the SE structure and the given W_n, we have, e.g.,

$$u_1 = \rho(u_6 + u_7)/2 + v_1;$$
$$u_4 = \rho(u_3 + u_5 + u_{10})/3 + v_4.$$

In general, $u_n = (I_n - \rho W_n)^{-1} v_n$, where I_n is the $N \times n$ identity matrix.

- If $\{v_i\} \overset{iid}{\sim} (0, \sigma_v^2)$, then
$$\text{Var}(u_n) = \sigma_v^2[(I_n - \rho W_n)'(I_n - \rho W_n)]^{-1}.$$
$$\Rightarrow \text{Var}(Y_n) = \sigma_v^2[(I_n - \rho W_n)'(I_n - \rho W_n)]^{-1}.$$
- However, $\mathrm{E}(Y_n) = X_n\beta$, the same as the regular linear model.

Spatial error induces **cross-sectional dependence** among disturbances $\{u_i\}$, and thus changes the variance of the response Y_n, but does not induce changes in the mean of Y_n.

SLR Model with Spatial Lag: $Y_n = \lambda W_n Y_n + X_n\beta + u_n$.

Let x_i' be the ith row of X_n. From the model and the given W_n, we have,

- $Y_1 = \lambda(Y_6 + Y_7)/2 + x_1'\beta + u_1$,
- $Y_7 = \lambda(Y_1 + Y_2 + Y_3)/3 + x_7'\beta + u_7$, etc.

In general, $Y_n = (I_n - \lambda W_n)^{-1}(X_n\beta + u_n)$.

If $\{u_i\} \overset{iid}{\sim} (0, \sigma_u^2)$, then

- $E(Y_n) = (I_n - \lambda W_n)^{-1} X_n\beta$,
- $\mathrm{Var}(Y_n) = \sigma_u^2[(I_n - \lambda W_n)'(I_n - \lambda W_n)]^{-1}$.

Spatial lag induces *not only* the cross-sectional dependence in u_n and thus in Y_n, *but also* also the endogenous (interaction) effects on the means of the responses! The latter highlight the difference between a spatial model with SL effect and a spatial model with SE effect.

SLR Model with Spatial Durbin effect: $Y_n = \beta_0 1_n + X_n\beta_1 + W_n X_n \gamma + u_n$.

- For example, $Y_1 = \beta_0 + x_1'\beta_1 + [(x_6 + x_7)'/2]\gamma + u_1$.

The mean of the response Y_1 of the spatial unit 1 is affected not only by its own regressors' values x_1, but also by the regressors' values of the neighboring spatial units 6 and 7, and similarly for the means of the responses of other spatial units.

- The term **Spatial Durbin** first appeared in Anselin (1988, p. 40), for its analogy to the suggestion by Durbin (1960) for time series models.
- SD term induces the **local effects** (from the immediate neighbors), in contrast to the SL effect which induces the **global effects** (from the immediate neighbors, neighbors of the immediate neighbors, etc.).
- Change in, e.g., Y_1 caused by the change in x_1 is referred to as the **direct effect**; and changes in other $Y's$ caused by the change in x_1 are referred to as the **indirect effects** that could be local, or global.

Maximum Likelihood Estimation

The maximum likelihood (ML) estimator holds special place among estimators (Cameron and Trivedi, 2005, p.139), as

- it is the most efficient estimator among estimators which are **consistent** and **asymptotically normal**, and
- the principle behind the ML estimation lays out a foundation based on which many other more general methods are obtained such as **quasi-ML estimation** and **M-estimation**.

The **likelihood principle**, due to a pioneer statistician R. A. Fisher (1922), is to choose, as an estimator of parameter vector θ_0, a value of θ that maximizes the **likelihood** of observing the actual sample data.

The joint probability density function (pdf) or joint probability mass function (pmf) of random observations can be interpreted as "the likelihood' for a particular sample to be observed".

Let $f(\mathbf{y}|\theta)$ be the joint pdf of random n-vector Y indexed by a vector θ of unknown parameters, given the exogenous regressors' values X and exogenous spatial weight matrices.

- Now, $f(\mathbf{y}|\theta)$ is viewed as a function of θ and is denoted by $L_n(\theta|\mathbf{y})$ or simply $L_n(\theta)$.
- Then $L_n(\theta)$ is called the **likelihood function** of θ based on the observed vector $\mathbf{y}$ on Y.
- Maximizing $L_n(\theta)$ is equivalent to maximizing the loglikelihood $\ell_n(\theta) = \ln L_n(\theta)$, and

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} \ell_n(\theta), \tag{1.4}$$

  is the so-called **maximum likelihood estimator** (MLE) of θ, where Θ denotes the parameter space.
- A simple interpretation of the MLE $\hat{\theta}_n$ is: $\hat{\theta}_n$ is the best 'guess' of the value of θ based on the observed data in that the observed vector $\mathbf{y}$ most likely 'came' from the 'population' represented by $f(\mathbf{y}|\hat{\theta}_n)$.

There are some important properties of the ML principle.

- Let $S_n(\theta) = \frac{\partial}{\partial\theta}\ell_n(\theta)$, then, $S_n(\theta)$ is called **score function**.
- At the **true value** θ_0 of θ, $S_n(\theta_0)$ is called **efficient score**.
- For standard ML estimation problems, $\hat{\theta}_n$ defined in (1.4) is equivalent to

$$\hat{\theta}_n = \arg\{S_n(\theta) = 0\}. \tag{1.5}$$

- Let 'E' denote the expectation with respect to the true pdf $f(\mathbf{y}|\theta_0)$, then

$$\mathrm{E}[S_n(\theta_0)] = 0, \tag{1.6}$$

- The expectation of the outer product of the efficient score vector:

$$\mathcal{I}_n \equiv \mathcal{I}_n(\theta_0) = \mathrm{E}[S_n(\theta_0)S_n'(\theta_0)], \tag{1.7}$$

  is called the **Fisher Information**, commonly known as the (expected) **information matrix**.

The well-known **information matrix equality (IME)** holds under true $\ell(\theta)$:

$$-\mathrm{E}[H_n(\theta_0)] = \mathrm{E}[S_n(\theta_0)S_n'(\theta_0)], \tag{1.8}$$

where $H_n(\theta_0) = \frac{\partial}{\partial \theta'} S_n(\theta_0)$, called the **Hessian** matrix. The negative Hessiam matrix is referred to as the **Observed Information Matrix**.

The proof of (1.6) goes as follows:

$$\begin{aligned}
\mathrm{E}[S_n(\theta_0)] &= \int S_n(\theta_0) f(\mathbf{y}|\theta_0) d\mathbf{y} \\
&= \int \frac{\partial \ell_n(\theta_0)}{\partial \theta} f(\mathbf{y}|\theta_0) d\mathbf{y} \\
&= \int \frac{\partial \ln f(\mathbf{y}|\theta_0)}{\partial \theta} f(\mathbf{y}|\theta_0) d\mathbf{y} = \int \frac{\partial}{\partial \theta} f(\mathbf{y}|\theta_0) d\mathbf{y} \\
&= \frac{\partial}{\partial \theta} \int f(\mathbf{y}|\theta_0) d\mathbf{y} = \frac{\partial}{\partial \theta}(1) = 0,
\end{aligned}$$

where differentiation and integration are assumed to be interchangeable, which is typically valid if the range of **y** does not depend on θ.

The proof of (1.8) goes as follows. Note that $H_n(\theta) = \frac{\partial^2}{\partial\theta\partial\theta'}\ell_n(\theta)$. We have

$$
\begin{aligned}
\mathrm{E}[H_n(\theta_0)] &= \int \frac{\partial^2 \ell_n(\theta_0)}{\partial\theta\partial\theta'} f(\mathbf{y}|\theta_0)d\mathbf{y} \\
&= \int \frac{\partial S_n(\theta_0)}{\partial\theta'} f(\mathbf{y}|\theta_0)d\mathbf{y} \\
&= \int \frac{\partial}{\partial\theta'}\left[S_n(\theta_0)f(\mathbf{y}|\theta_0)\right]d\mathbf{y} - \int S_n(\theta_0)\left[\frac{\partial}{\partial\theta'}f(\mathbf{y}|\theta_0)\right]d\mathbf{y} \\
&= \int \frac{\partial}{\partial\theta'}\left[S_n(\theta_0)f(\mathbf{y}|\theta_0)\right]d\mathbf{y} - \int S_n(\theta_0)S_n'(\theta_0)f(\mathbf{y}|\theta_0)d\mathbf{y} \\
&= \frac{\partial}{\partial\theta'}\int S_n(\theta_0)f(\mathbf{y}|\theta_0)d\mathbf{y} - \int S_n(\theta_0)S_n'(\theta_0)f(\mathbf{y}|\theta_0)d\mathbf{y} \\
&= 0 - \mathrm{E}[S_n(\theta_0)S_n'(\theta_0)] \\
&= -\mathcal{I}_n.
\end{aligned}
$$

Under some 'regularity conditions', we have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N(0,\ \lim_{n\to\infty} n\mathcal{I}_n^{-1}). \tag{1.9}$$

Quasi Maximum Likelihood Estimation

In practice, the true distribution of Y, say $g(\mathbf{y}|\theta)$, is often unknown, the 'chosen' distribution $f(\mathbf{y}|\theta)$ can best be an approximation to $g(\mathbf{y}|\theta)$. Thus,

- $L_n(\theta)$ defined in terms of $f(\mathbf{y}|\theta)$ is a wrong (quasi) likelihood function. Would maximizing the quasi-likelihood $L_n(\theta)$ still lead to a consistent estimator for θ_0? Would the asymptotic normality result still hold?
- It is well known that for a maximum likelihood type estimator, or M-estimator, to be consistent, it is necessary that

$$\text{plim}\frac{1}{n}S_n(\theta_0) = 0, \tag{1.10}$$

  which boils down by requiring that $\lim \frac{1}{n}\text{E}[S_n(\theta_0)] = 0$.
- However, this is not guaranteed if $f(\mathbf{y}|\theta)$ is a **misspecification**.
- Very interestingly, we can show that for a **certain choice** of $f(\mathbf{y}|\theta)$ that partially specifies $g(\mathbf{y}|\theta)$ in the sense that its first two moments matches these of $g(\mathbf{y}|\theta)$, consistency can also be achieved.
- One such a choice is the Gaussian (normal) distribution.

Let $Y \sim (\mu_n, \Sigma_n)$, i.e., the true joint pdf $g(\mathbf{y}|\theta_0)$ is correctly specified for the first two moments only: $\mathrm{E}(Y) = \mu_n \equiv \mu_n(\beta_0)$ and $\mathrm{Var}(Y) = \Sigma_n \equiv \Sigma_n(\gamma_0)$.

If the ‘chosen’ distribution is a normal, i.e., $f(\mathbf{y}|\theta_0) \sim N[\mu(\beta_0), \Sigma(\gamma_0)]$, then the quasi loglikelihood is

$$\ell_n(\theta) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_n(\gamma)| - \frac{1}{2}\big[\mathbf{y} - \mu_n(\beta)\big]'\Sigma_n^{-1}(\gamma)\big[\mathbf{y} - \mu_n(\beta)\big], \quad (1.11)$$

where , and $\theta = (\beta', \gamma')'$. The quasi score function is

$$S_n(\theta) = \begin{cases} \dot{\mu}_n'(\beta)\Sigma_n^{-1}(\gamma)\big[\mathbf{y} - \mu_n(\beta)\big], \\ -\mathtt{tr}[\Sigma_n^{-1}(\gamma)\dot{\Sigma}_{nk}(\gamma)] + \big[\mathbf{y} - \mu_n(\beta)\big]'\Sigma_n^{-1}(\gamma)\dot{\Sigma}_{nk}(\gamma)\Sigma_n^{-1}(\gamma)\big[\mathbf{y} - \mu_n(\beta)\big], \end{cases}$$

for $k = 1, \ldots, \dim(\gamma)$, where $\dot{\mu}_n'(\beta) = \frac{\partial}{\partial\beta}\mu_n(\beta)$ and $\dot{\Sigma}_{nk}(\gamma) = \frac{\partial}{\partial\gamma_k}\Sigma_n(\gamma)$.

Note: for a matrix function $A(\rho)$ of scalar parameter ρ, positive definite (p.d.) $\forall\rho$,

(a) $\frac{\partial}{\partial\rho}A(\rho)^{-1} = -A(\rho)^{-1}[\frac{\partial}{\partial\rho}A(\rho)]A(\rho)^{-1}$,

(b) $\frac{\partial}{\partial\rho}\log|A(\rho)| = \mathrm{tr}[A(\rho)^{-1}\frac{\partial}{\partial\rho}A(\rho)]$,

where $\mathrm{tr}(\cdot)$ = trace of a matrix, see Horn and Johnson (1985).

Clearly,

$$\mathrm{E}[S_n(\theta_0)] = 0,$$

whether or not $f(\mathbf{y}|\theta) = g(\mathbf{y}|\theta)$, as the quasi score $S_n(\theta_0)$ is a linear-quadratic (LQ) function of $u_n = \mathbf{y} - \mu_n(\beta_0)$!

- As long as the two distribution have the same first two moments, the expectation of $S_n(\theta_0)$ equals zero.
- See Huber (1967) for the original ideas of QML estimation.
- See also White (1982) for a similar interpretation and for the related **information matrix test**.

Maximizing $\ell_n(\theta)$ gives the quasi (Gaussian) MLE, or QMLE $\hat{\theta}_n$ of θ if $f(\mathbf{y}|\theta) \neq g(\mathbf{y}|\theta)$, which become MLE if $f(\mathbf{y}|\theta) = g(\mathbf{y}|\theta)$. The Gaussian QMLE $\hat{\theta}_n$ would be still consistent and asymptotically normal. The difference is that its asymptotic variance becomes

$$\mathrm{Var}(\hat{\theta}_n) \overset{a}{=} \mathcal{J}_n^{-1}(\theta_0)\mathcal{I}_n(\theta_0)\mathcal{J}_n^{-1}(\theta_0), \tag{1.12}$$

where $\mathcal{J}_n(\theta_0) = -\mathrm{E}[H_n(\theta_0)]$.

- Thus, in case that the true underlining distribution of a model is nonnormal, but normal (Gaussian) likelihood is used, the resulting estimator would still be consistent and asymptotically normal. This makes the Gaussian QML estimation method very attractive.
- A **limitation** of the QML method is that it cannot be applied to models with additional endogenous regressors where endogeneity is of **implicit form**, unlike the endogenous 'spatial' regressor WY.
- More general results hold for all distributions from the **exponential family** (see Cameron and Trivedi, 2005, Sec. 5.7).

M-Estimation

The term 'M-estimator' was coined by Huber (1964) to mean the maximum likelihood type estimator (see also Huber, 1969). An M-estimator $\hat{\theta}_n$ of a $p \times 1$ vector of parameters θ is defined is defined in two ways:

(a) as the solution of a maximization problem and
(b) as the root of a set of estimating equations.

In case (a), $\hat{\theta}_n$ is defined as

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} Q_n(\theta), \quad (1.13)$$

for real valued, random criterion function $Q_n(\theta)$.

In case (b), $\hat{\theta}_n$ is defined as

$$\hat{\theta}_n = \arg\{\Psi_n(\theta) = 0\}, \quad (1.14)$$

for $p \times 1$ vector-valued **estimating function** $\Psi_n(\theta)$.

The asymptotic behavior of the M-estimator $\hat{\theta}_n$ is similar to that of the QML estimator. In particular, it is consistent and asymptotically normal.

Under regularity conditions, the M-estimator $\hat{\theta}_n$ is such that $\hat{\theta}_n \xrightarrow{p} \theta_0$, and

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N[0, \lim_{n\to\infty} n\Sigma_n^{-1}(\theta_0)\Gamma_n(\theta_0)\Sigma_n^{-1}(\theta_0)],$$

where $\Gamma_n(\theta_0) = \mathrm{Var}[\Psi_n(\theta_0)]$ and $\Sigma_n(\theta_0) = -\mathrm{E}[\frac{\partial}{\partial\theta'}\Psi_n(\theta_0)]$.

- The original formulation of the M-estimation is in terms of Q_n or Ψ_n which is the sum of n independent quantities. However, this can be generalized to allow either function to have more general forms to suit for the estimation of more general models such as the spatial econometric models considered in this course.
- See van der Vaart (1998) for more general formulations of M-estimation strategy, including the methods for proving its asymptotic properties.
- See also White (1994) for the relevant asymptotic methods.

Generalized Method of Moments Estimation

The generalized method of moments (GMM), developed by Lars Peter Hansen in 1982 as a generalization of the method of moments which was introduced by Karl Pearson in 1894, is a generic method for estimating parameters in statistical models.

- Usually it is applied in the context of semiparametric models, where the parameter of interest is finite-dimensional, whereas the full shape of the distribution function of the data may not be known, and therefore maximum likelihood estimation is not applicable.
- The method requires that certain **moment conditions** be specified for the model. These moment conditions are functions of the model parameters and the data, such that their expectation is zero at the true values of the parameters.
- The GMM method then minimizes a certain norm of the sample moment conditions.

In its most compact form, the GMM estimator $\hat{\theta}_n$ minimizes the following object function:

$$Q_n(\theta) = \mathbf{g}_n'(\theta)\Omega_n\mathbf{g}_n(\theta), \tag{1.15}$$

where $\mathbf{g}_n(\theta)$ is a $r \times 1$ vector of sample moments, θ is a $p \times 1$ vector of parameters ($p \leq r$), and Ω_n is the weights matrix.

Under certain regularity conditions, the GMM estimator $\hat{\theta}_n$ defined above is consistent, i.e., $\hat{\theta}_n \xrightarrow{p} \theta_0$, and asymptotically normal, i.e.,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} N[0, \lim_{n\to\infty} n(\Sigma_n'\Omega_n\Sigma_n)^{-1}(\Sigma_n'\Omega_n\Gamma_n\Omega_n\Sigma_n)(\Sigma_n'\Omega_n\Sigma_n)^{-1}],$$

where $\Gamma_n = \mathrm{Var}[\mathbf{g}_n(\theta_0)]$ and $\Sigma_n = -\mathrm{E}[\frac{\partial}{\partial\theta'}\mathbf{g}_n(\theta_0)]$.

GMM estimator is efficient in the class of all estimators that do not use any extra information aside from that contained in the moment conditions.

- Choice of weights matrix Ω_n is an issue of concern.
- Choice of moment conditions is another issue of concern.

Optimal GMM. If Γ_n is 'known', then one can choose $\Omega_n = \Gamma_n^{-1}$. In this case, the asymptotic variance-covariance (VC) matrix of the GMM estimator simplifies to

$$(\Sigma_n \Gamma_n^{-1} \Sigma_n)^{-1},$$

a similar form to the asymptotic VC matrix of an M-estimator.

When $r = p$, the model is said to be **just-identified**. In this case, Ω_n is simply taken to be the identity matrix, and the GMM estimator becomes the **method of moments** (MM) estimator defined as

$$\hat{\theta}_n = \arg\{\mathbf{g}_n'(\theta) = 0\}. \tag{1.16}$$

Clearly, the MM estimator is also the M-estimator defined earlier. In this case, Σ_n is also invertible, and the asymptotic VC matrix can be written as $\Sigma_n^{-1} \Gamma_n \Sigma_n^{-1}$.

Neighborhood Crime. In illustrating the applications of spatial linear regression models, Anselin (1988, p.187) used neighborhood crime data (readme) corresponding to 49 contiguous neighborhood in Columbus, Ohio, in 1980. These neighborhood correspond to census tracts, or aggregates of a small number of census tracts, where

- **Crime:** the combined total of residential burglaries and vehicle thefts per thousand household in the neighborhood (the response variable).
- **Income** and **House:** the explanatory variables representing income and housing values in thousand dollars.
- **East:** a dummy variable indicates whether the 'neighborhood' in the east or west of a main north-south transportation axis.
- In addition, the neighborhood centroid coordinates (X and Y) are also given, as well as the list of neighbors of each spatial unit (neighborhood) that gives a first-order contiguity spatial weight matrix.

Boston House Price. The data, given by Harrison and Rubinfeld (1978), and corrected and augmented with longitude and latitude by Gilley and Pace (1996), contains 506 observations (1 observation per census tract) from Boston Metropolitan Statistical Area, and can be found in R (spdep). The response variable **MEDV** and 13 explanatory variables are:

- **MEDV:** median value (corrected) of owner-occupied homes in 1000's;
- **crime:** per capita crime rate by town;
- **zoning:** proportion of residential land zoned for lots over 25,000 square feet;
- **industry:** proportion of non-retail business acres per town;
- **charlesr:** Charles River dummy variable (= 1 if tract bounds river);
- **nox:** nitric oxides concentration (parts per 10 million);
- **room:** average number of rooms per dwelling;
- **houseage:** proportion of owner-occupied units built prior to 1940;

- **distance:** weighted distances to five Boston employment centres;
- **access:** index of accessibility to radial highways;
- **taxrate:** full-value property-tax rate per 10,000;
- **ptratio:** pupil-teacher ratio by town;
- **blackpop:** $1000(\mathrm{Bk} - 0.63)^2$ where Bk is the proportion of blacks by town;
- **lowclass:** lower status of the population proportion.

The spatial weight matrix is constructed using the Euclidean distance in terms of longitude and latitude. A threshold distance, e.g., 0.05, is chosen, which gives a W_n matrix with 19.08% non-zero elements.

Public Capital Productivity. The statewide capital productivity data of Munnell (1990) give indicators related to public capital productivity for 48 US states observed over 17 years (1970-1986). .

In Munnel (1990), the empirical model specified is a Cobb-Douglas production function, with state-specific fixed effects μ:

$$\ln Y = \beta_0 + \beta_1 \ln K_1 + \beta_2 \ln K_2 + \beta_3 \ln L + \beta_4 \mathrm{Unemp} + \mu + \epsilon,$$

where Y is the gross social product of a given state,

- K_1 is public capital,
- K_2 is private capital,
- L is labour input and Unemp is the state unemployment rate.

The model can be extended by adding SL, SE, SD, or dynamic effect. The spatial weights matrix W takes a contiguity form with its (i, j)th element being 1 if states i and j share a common border, otherwise 0.

Cigarette Demand. This is another well known panel data that has been applied under various panel data model frameworks, non-spatial or spatial, fixed effects or random effects, static or dynamic. In particular, the demand equations for cigarettes for United States were estimated, based on a panel of 46 states over 30 time periods (1963-1992), given as `CIGAR.TXT` on the Wiley web site associated with book of Baltagi (2005).

- Y = Cigarette sales in packs per capita (the response variable).
- X_1 = Price per pack of cigarettes;
- X_2 = Population (Pop);
- X_3 = Population above the age of 16;
- X_4 = Consumer price index with (1983=100);
- X_5 = Per capita disposable income;
- X_6 = Minimum price in adjoining states per pack of cigarettes.
- Several time dummies corresponding to the major policy interventions in 1965, 1968 and 1971 can be added into the model.
- Spatial weights can be the first-order contiguity matrix.

Spatial Software Tools

Major software include Matlab, STATA, Python, and R. See Computer Labs for more information on their installations and use.

The Matlab codes provided in this course are sufficient for the applications of common spatial econometrics models and methods introduced.

Additional useful sources include:

- GeoDa under The Center of Spatial Data Science, University of Chicago, is a free software package for spatial data analysis developed by Luc Anseline.
- Spatial Econometrics Toolbox by James LaSage privide some matlab codes for spatial analyses.
- spdep is an R package containing some R functions for spatial analyses.
- Paul Elhorst lists some matlab codes for spatial panel data analyses.
- The book, Fischer and Getis (2010), gives a comprehensive coverage on spatial software, methods and applications.

Basics on Econometrics and Matrix Algebra

Some background knowledge on multiple linear regression models, panel data models, and matrix theory would help learning of spatial econometric models and methods.

- For basics on multiple linear regression models required for the learning of SLR models, read Chapters 1-5 of Greene (2012).
- For basics on standard panel data models, read Chapters 1-5, and 8 of Baltagi (2005).
- For basics on matrix algebra, read Appendix A of Anderson (2003);
- or Appendix A of Greene (2022, 8th Ed.).

References

- Anselin, L., 1988. *Spatial econometrics: Methods and models*. Dordrecht: Kluwer Academic Publishers.
- Anselin L., Bera, A. K., 1998. Spatial dependence in linear regression models with an introduction to spatial econometrics. In: *Handbook of Applied Economic Statistics, Edited by Aman Ullah and David E. A. Giles*. New York: Marcel Dekker.
- Baltagi, B. H., 2013. *Econometric Analysis of Panel Data*, 5th edition. John Wiley & Sons Ltd.
- Case, A. C., 1991. Spatial Patterns in Household Demand. *Econometrica* 59, 953-965.
- Case, A. C., Rosen, H. S., Hines, J. R., 1993. Budget Spillovers and Fiscal Policy Interdependence: Evidence from the States. *Journal of Public Economics* 52, 285-307.
- Cameron, A. C., Trivedi, P. K., 2005. *Microeconometrics: Methods and Applications*. Cambridge University Press.

- Doreian, P., 1980. Linear models with spatially distributed data: spatial disturbances and spatial effects. *Sociology Methods and Research* 9, 29-60.
- Durbin, J. 1960. Estimation of parameters in time-series regression models. *Journal of the Royal Statistical Soceity B* 22, 139-153.
- Elhorst, J. P., 2014. *Spatial Econometrics From Cross-Sectional Data to Spatial Panels*. Springer.
- Fisher, R. A., 1922. On mathematical foundation of theoretical statistics. *Philosophical Transactions of the Royal Society of London A* 222, 309-368.
- Gilley, O. W., Pace, R. K., 1996. On the Harrison and Rubinfeld Data. *Journal of Environmental Economics and Management* 31, 403-405.
- Hansen, L. P., 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50, 1029-1054.
- Harrison, D., Rubinfeld, D. L., 1978. Hedonic housing prices and he demand for clear air. *Journal of Environmental Economics and Management* 5, 81-102.

- Huber, P. J., 1964. Robust estimation of a location parameter. *Annals of Mathematical Statistics* 35, 73-101.
- Huber, P. J., 1967. The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symp. on Mathematical Statistics and Probability*, vol. 1, pp. 221-233.
- Kelejian H. H., Prucha, I. R., 1999. A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review* 40, 509-533.
- Kou S., Peng, X, Zhong, H., 2017. Asset Pricing with Spatial Interaction. *Management Science*, Published online in Articles in Advance 06 Feb 2017.
- Lee, L. F., 2004. Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica* 72, 1899-1925.
- LeSage, J., Pace, R.K., 2009. *Introduction to Spatial Econometrics*. London: CRC Press, Taylor & Francis Group.
- Manski, C. F., 1993. Identification of endogenous social effects: the reflection problem. *The Review of Economic Studies* 60, 531-542.

- Munnell, A.H., 1990. Why has productivity growth declined? Productivity and public investment. *New England Economic Review* Jan/Feb, 3-22.
- Paelinck, J. H. P.; Klaassen, L. H., 1979. *Spatial econometrics*. Farnborough: Saxon House.
- Pearson, K., 1894. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Soceity of London (A)* 185, 71-110.
- Qu, X.; L. F. Lee (2015). Estimating a spatial autoregressive model with an endogenous spatial weight matrix. *Journal of Econometrics* 184, 209-232.
- Tobler, W., 1979. Celluar geography. In *Philosophy in Geography*, edited by S. Gale and G. Olsson, pp. 379-86. Dordrecht: Reidel.
- White, H., 1982. Maximum likelihood estimation of misspecified models. *Econometrica* 50, 1-25.
- White, H., 1994. *Estimation, Inference and Specification Analysis*. New York: Cambridge University Press.
- van der Vaart, A. W., 1998. *Asymptotic Statistics.* Cambridge: Cambridge University Press.