2. Statistical Modelling (3)

Statistical Modelling & Machine Learning

Jaejik Kim

Department of Statistics
Sungkyunkwan University

STA3036

Data with Time Dependency

- Data: $(y_t, \boldsymbol{x}_t),\ t = 1, \ldots, T$.
  - $(y_t, \boldsymbol{x}_t)$ are measured for the same object at discrete time points (e.g., hourly, weekly, monthly, yearly data).

- Model: $Y_t = f(\boldsymbol{X}_t; \boldsymbol{\theta}) + \epsilon_t,\ t = 1, \ldots, T)$.
  - $\epsilon_t,\ t = 1, \ldots, T$ have constant variance.
  - ϵ_t's are correlated (time dependency) $\Rightarrow$ Y_t's are correlated.
  - ϵ_t's have a stationary process (i.e., covariance between ϵ_t's depends only on time difference).
  - ARMA (p,q) time series modelling for ϵ_t.

$$\epsilon_t = \sum_{j=1}^{p} \alpha_j \epsilon_{t-j} + \sum_{j=1}^{q} \phi_j \eta_{t-j} + \eta_t,$$

where $\eta_t \sim N(0, \sigma^2)$.

Regression Model with AR(1) Error

- Regression Model with AR(1) Error:

$$\begin{aligned} Y_t &= f(\boldsymbol{X}_t; \boldsymbol{\theta}) + \epsilon_t, \\ \epsilon_t &= \alpha \epsilon_{t-1} + \eta_t, \end{aligned}$$

where α is an autocorrelation parameter satisfying $|\alpha| < 1$ (stationary condition), and $\eta_t \overset{iid}{\sim} N(0, \sigma^2)$.

- AR(1) error: From the AR(1) model and recursive calculations, we obtain

$$\epsilon_t = \sum_{j=0}^{\infty} \alpha^j \eta_{t-j}.$$

Properties of AR(1) Error

- Since $\epsilon_t = \sum_{j=0}^{\infty} \alpha^j \eta_{t-j}$ and $E(\eta_t) = 0$ for all t,

$$E(\epsilon_t) = 0.$$

- Since η_t's are independent and $Var(\eta_t) = \sigma^2$ for all t,

$$Var(\epsilon_t) = \frac{\sigma^2}{1-\alpha^2}.$$

- Covariance of ϵ_t and ϵ_{t-j}:

$$Cov(\epsilon_t, \epsilon_{t-j}) = \alpha^j \left(\frac{\sigma^2}{1-\alpha^2} \right), \; j \neq 0.$$

Maximum Likelihood Estimation of AR(1) Error Model

Method 1: Likelihood function from multivariate normal density.

- $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_T)^\top \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \frac{\sigma^2}{1-\alpha^2}\begin{pmatrix} 1 & \alpha & \cdots & \alpha^{T-1} \\ \alpha & 1 & \cdots & \alpha^{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha^{T-1} & \alpha^{T-2} & \cdots & 1 \end{pmatrix}.$$

- $\boldsymbol{y} = (y_1, \dots, y_T)^\top \sim MVN(\boldsymbol{f}, \boldsymbol{\Sigma})$.
- Log-likelihood function:

$$l(\boldsymbol{\theta}; \boldsymbol{y}, \alpha, \sigma^2) = -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{f})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{f}).$$

Maximum Likelihood Estimation of AR(1) Error Model

Method 2: Likelihood function from conditional density.

- Relationship between joint density and conditional densities:

$$\begin{aligned} p(x_1, x_2, \ldots, x_n) &= p(x_n|x_1, \ldots, x_{n-1})p(x_{n-1}|x_1, \ldots, x_{n-2}) \\ &\quad \cdots p(x_2|x_1)p(x_1). \end{aligned}$$

- AR(1) structure:
  - Current status depends only on the previous status (Markovian property).
  - i.e., X_t depends only on X_{t-1}
    $\Rightarrow X_t$ is independent of $X_{t-2}, X_{t-3}, \ldots, X_1$.

$$\begin{aligned} p(x_1, x_2, \ldots, x_n) &= p(x_n|x_{n-1})p(x_{n-1}|x_1, \ldots, x_{n-2}) \\ &\quad \cdots p(x_2|x_1)p(x_1) \\ &= \left[\prod_{t=2}^{n} p(x_t|x_{t-1})\right] p(x_1). \end{aligned}$$

Maximum Likelihood Estimation of AR(1) Error Model

- AR(1) error: $\epsilon_t = \alpha\epsilon_{t-1} + \eta_t$, $\eta_t \sim N(0, \sigma^2)$.
- $\epsilon_t | \epsilon_{t-1} \ \sim \ N(\alpha\epsilon_{t-1}, \sigma^2)$.
- Since $E(\epsilon_t) = 0$ and $Var(\epsilon_t) = \frac{\sigma^2}{1-\alpha^2}$, $\epsilon_1 \ \sim \ N\left(0, \frac{\sigma^2}{1-\alpha^2}\right)$.
- $Y_t | Y_{t-1} \ \sim \ N(f(\boldsymbol{X}_t; \boldsymbol{\theta}) + \alpha\epsilon_{t-1}, \sigma^2)$.
- $Y_1 \ \sim \ N\left(f(\boldsymbol{X}_1; \boldsymbol{\theta}), \frac{\sigma^2}{1-\alpha^2}\right)$.
- Log-likelihood function:

$$l(\boldsymbol{\theta}; \boldsymbol{y}, \alpha, \sigma^2) = \sum_{t=2}^{T} \log p(Y_t | Y_{t-1}) + \log p(Y_1).$$

Maximum Likelihood Estimation of AR(1) Error Model

Estimation Algorithm:

1. Set the initial parameter vectors $\hat{\boldsymbol{\theta}}$.
2. Compute residuals $r_t = y_t - f(\boldsymbol{x}_t; \hat{\boldsymbol{\theta}}),\ t = 1, \ldots, T$.
3. Estimate the AR(1) model parameters α and σ^2 using the residuals $r_1, \ldots, r_T$.
4. Construct $\boldsymbol{\Sigma}$ using $\hat{\alpha}$ and $\hat{\sigma}^2$ obtained from Step 3.
5. Find $\hat{\boldsymbol{\theta}}$ minimizing $(\boldsymbol{y} - \boldsymbol{f})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{f})$.
6. Repeat Steps 2–5 until $\boldsymbol{\theta}$ is converged.

Data with Spatial Correlations

- Data are observed at spatial points in 2 or 3 dimensional space (e.g., house price in a city, house income in a city, the number of infectious persons in an area, etc.)
- There exist correlations between spatial points.
- Basically, as distance between two spatial points increases, the correlation decreases.
- There are various approaches for spatial prediction problems (spatial autoregressive model, spatial error model, kriging, etc.)

Spatial Autoregressive Model (SAR)

- Data: $(y_s, \boldsymbol{x}_s),\ s = 1, \ldots, S$.
- Spatial autoregressive model:

$$\boldsymbol{y} = \rho \boldsymbol{W} \boldsymbol{y} + \boldsymbol{X} \beta + \epsilon.$$

  - $\boldsymbol{y}$: $S \times 1$ output variable vector.
  - ρ: Spatial autocorrelation parameter.
  - $\boldsymbol{W}$: $S \times S$ weight matrix that accounts for the spatial dependencies among spatial units.
  - $\boldsymbol{X}$: $S \times p$ input matrix.
  - β: Coefficient vector of $\boldsymbol{X}$.
  - $\epsilon \sim MVN(\mathbf{0}, \sigma^2 \boldsymbol{I})$.

Spatial Weight Matrix

- Spatial weight matrix $\boldsymbol{W} = (w_{ij};\ i, j = 1, \ldots, S)$:
  - w_{ij}: Spatial influence of unit j on unit i.
  - $w_{ii} = 0$ (i.e., all diagonal elements of $\boldsymbol{W}$ are 0).

- Construction of $\boldsymbol{W}$:
  1. Weights based on distance:
     - k-Nearest neighbor weights.
     - Radial distance weights.
     - Power distance weights.
     - Exponential distance weights.
     - Double power distance weights.
  2. Weights based on boundaries:
     - Spatial contiguity weights.
     - Shared-boundary weights.
  3. Combined distance-boundary weights.

Construction of $\boldsymbol{W}$

Weights based on distance (1):

- k-Nearest neighbor weights:
  - d_{ij}: Distance between unit i and unit j.
  - $N_k(i)$: A set containing the k closet units to unit i based on $d_{ij},\ j = 1, \ldots, S,\ i \neq j$.
  - If $j \in N_k(i)$, then $w_{ij} = 1$. Otherwise, $w_{ij} = 0$.
  - For the symmetric matrix of $\boldsymbol{W}$, if $j \in N_k(i)$ or $i \in N_k(j)$, then $w_{ij} = 1$. Otherwise, $w_{ij} = 0$.

- Radial distance weights:
  - d: Threshold distance.
  - If d_{ij} is larger than d, units i and j have no spatial influence.
  - No diminishing effect of spatial influence up to d.
  - If $d_{ij} \leq d$, then $w_{ij} = 1$. Otherwise, $w_{ij} = 0$.

Construction of W

Weights based on distance (2):

- Power distance weights:
  - It considers diminishing effect of spatial influence.
  - $w_{ij} = d_{ij}^{-\alpha}$.
  - $\alpha > 0$. Typical choice of α is 1 or 2.
- Exponential distance weights:
  - Diminishing effect of spatial influence.
  - $w_{ij} = \exp(-\alpha d_{ij})$.
  - $\alpha > 0$.
- Double-power distance weights:
  - Bell-shaped function & threshold distance d.
  - If $d_{ij} \leq d$, then $w_{ij} = \left[1 - (d_{ij}/d)^k\right]^k$. Otherwise $w_{ij} = 0$.
  - Typical choice of k is 2, 3, or 4.

Construction of W

Weights based on boundaries: The boundaries shared between spatial units play in important role in determining degree of spatial influence.

- Spatial Contiguity weights:
  - If units i and j share their boundary, $w_{ij} = 1$. Otherwise $w_{ij} = 0$.
  - However, even if two units have a shared corner point, this weight returns 1.
  - l_{ij}: Length of shared boundary.
  - If $l_{ij} > 0$, then $w_{ij} = 1$. If $l_{ij} = 0$, $w_{ij} = 0$.

- Shared-boundary weights:
  - Proportional boundary length between unit i and j.
  - l_i: Total boundary length that unit i is shared with all other units (i.e., $\sum_{j=1,\dots,S,\ j\neq i} l_{ij}$).
  - $w_{ij} = l_{ij}/i_i$.

Construction of W

Combined distance-boundary weights:

- Spatial influence represented by both distance and boundary relations.
- Cliff and Ord (1969) proposed the weight by the combination of power distance and boundary-shares as follows:

$$w_{ij} = \frac{l_{ij} d_{ij}^{-\alpha}}{\sum_{k=1,\dots,S,\ k\neq i} l_{ik} d_{ik}^{-\alpha}}.$$

where $\alpha > 0$. Typical choice of α is 1.

Normalization of $\boldsymbol{W}$

- Normalization: Normalization of spatial effect for removing scale effects.
- Row normalized weights:
  - The sum of each row is 1 (i.e., $\sum_{j=1}^{S} w_{ij} = 1$).

$$w_{ij} \leftarrow \frac{w_{ij}}{\sum_{k=1,\ldots,S,k\neq i} w_{ij}}.$$

- Scalar normalized weights:
  - Row normalization is not appropriate for comparison between rows.
  - Scalar normalization: $\gamma \boldsymbol{W}$, where γ is a positive scalar.
  - $\gamma = 1/\max(w_{ij}) \Rightarrow$ All normalized w_{ij} has a value between 0 and 1 (relative influence intensity).
  - $\gamma = 1/\lambda_{max}$, where λ_{max} is the largest eigenvalue of $\boldsymbol{W}$.

Maximum Likelihood Estimation of SAR

- SAR Model: Let $\boldsymbol{A} = \boldsymbol{I} - \rho \boldsymbol{W}$

$$
\begin{aligned}
& \boldsymbol{y} = \rho \boldsymbol{W} \boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\
\Rightarrow \quad & (\boldsymbol{I} - \rho \boldsymbol{W})\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\
\Rightarrow \quad & \boldsymbol{\epsilon} = (\boldsymbol{I} - \rho \boldsymbol{W})\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} \\
\Rightarrow \quad & \boldsymbol{\epsilon} = \boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}.
\end{aligned}
$$

- Since $\boldsymbol{\epsilon} \sim MVN(\mathbf{0}, \sigma^2 \boldsymbol{I})$, the pdf of $\boldsymbol{\epsilon}$ is

$$
\begin{aligned}
p(\boldsymbol{\epsilon}) &= (2\pi\sigma^2)^{-S/2} \exp\left[-\frac{1}{2\sigma^2} \boldsymbol{\epsilon}^\top \boldsymbol{\epsilon} \right] \\
&= (2\pi\sigma^2)^{-S/2} \exp\left[-\frac{1}{2\sigma^2} (\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \right].
\end{aligned}
$$

Maximum Likelihood Estimation of SAR

- To construct the likelihood function of $(\rho, \boldsymbol{\beta}, \sigma^2)$, we need the pdf of $\boldsymbol{y}$.
- The pdf of $\boldsymbol{y}$ can be obtained by the transformation of the random vector $\boldsymbol{\epsilon}$ ($\because \boldsymbol{\epsilon} = \boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\beta$).
- Since $\boldsymbol{y} = \boldsymbol{A}^{-1}\boldsymbol{X}\beta + \boldsymbol{A}^{-1}\boldsymbol{\epsilon}$ is differentiable and monotone within the range of $\boldsymbol{\epsilon}$,

$$\begin{aligned} p(\boldsymbol{y}) &= p(\boldsymbol{\epsilon}) \left| \frac{d\boldsymbol{\epsilon}}{d\boldsymbol{y}} \right| \\ &= (2\pi\sigma^2)^{-S/2} \exp\left[-\frac{1}{2\sigma^2} (\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\beta)^\top (\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\beta) \right] |\boldsymbol{A}|. \end{aligned}$$

Maximum Likelihood Estimation of SAR

- Log-likelihood function:

$$
\begin{aligned}
l(\rho, \boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}) &= -\frac{S}{2}\log(\sigma^2) + \log|\boldsymbol{A}| \\
&\quad -\frac{1}{2\sigma^2}(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\beta)^\top(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\beta).
\end{aligned}
$$

- MLE of $\boldsymbol{\beta}$ and σ^2: By solving $\frac{\partial l(\rho,\boldsymbol{\beta},\sigma^2|\boldsymbol{y})}{\partial \boldsymbol{\beta}} = 0$ and $\frac{\partial l(\rho,\boldsymbol{\beta},\sigma^2|\boldsymbol{y})}{\partial \sigma^2} = 0$, respectively,

$$
\begin{aligned}
\hat{\beta} &= (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{A}\boldsymbol{y}, && (1) \\
\hat{\sigma}^2 &= \frac{1}{S}(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\hat{\beta})^\top(\boldsymbol{A}\boldsymbol{y} - \boldsymbol{X}\hat{\beta}). && (2)
\end{aligned}
$$

Maximum Likelihood Estimation of SAR

- MLE of ρ: By replacing $(\boldsymbol{\beta}, \sigma^2)$ with $(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$,

$$\max_{|\rho|<1} l(\rho|\boldsymbol{y}) = \max_{|\rho|<1} \log|\boldsymbol{A}| \\ -\frac{S}{2}\log(\boldsymbol{A}\boldsymbol{y})^\top(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)^\top(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top)(\boldsymbol{A}\boldsymbol{y}). \quad (3)$$

- ML Estimation procedure:
  1. Find $\hat{\rho}$ by solving the maximization problem (3).
  2. Compute $\boldsymbol{A} = \boldsymbol{I} - \hat{\rho}\boldsymbol{W}$.
  3. Obtain $\hat{\beta}$ and $\hat{\sigma^2}$ using (1) and (2), respectively.