Time Series Class Project
Pooled Time Series Analysis For Autoregressive Models

Laura Lu
University of Notre Dame

December 18, 2009

Background and Objective

Time series analysis mainly focuses on the changes within a person, which is often referred to as the intra-individual change. As an opposite, the inter-individual change is to study the change between persons. There are many approaches to study the time series analysis. Generally speaking, multiple subjects approach is a practical approach to estimate the model parameters. First, multiple subjects will be measured, and then the subjects with the similar behavior trend will be picked out to study. In the multiple subjects study, some factors, such as the length of series T, the number of participants N, and the data type, are important in the design of experiments. If the length of series is too short or the number of participants is too small, then the researchers may can not obtain the accurate statistical inferences and therefore miss the intra-individual change; On the other hand, if the length is too long or the number of participants is too large, then the experiment will cost more than necessary, although it seems true that the longer the time series or the more participants, the more accurate the estimate. Also, the choose of T and N will be different for different data types. There are two data types, continuous and ordinal. The ordinal data are usually collected by the Likert table.

After we collected data, basically, there are two MLE estimation methods: exact MLE estimation method and conditional MLE estimation method. Both methods obtain the parameter estimates through maximizing the multiple subjects' likelihood function. For multiple subjects, theoretically we should use multiple likelihood functions, or separate time series. However, in practice there is a more popular and more convenient way to analyze data: using the pooled data, or the directly connected data. This method analyzes the connected data from multiple subjects as from a single subject. Upon these two estimations and two data organization forms, we have four methods to obtain estimates: pooled likelihood exact MLE, pooled likelihood conditional MLE, pooled data exact MLE, pooled data conditional MLE.

This study compares these four estimation methods under different scenarios. The goals of this study are (1) to investigate the estimation of four methods under different scenarios, such as different lengthes of series, different numbers of subjects, and different data types; (2) to find which estimation method provides accurate estimates for different scenarios; and (3) to provide some insights on experimental design, for example, how many subjects and how many occasions should be used in a study.

Model Description

We first consider a model for a single subject (or individual). And then we extend it to the model for multiple subjects.

The AR(1) Model

Suppose in the current study we are interested in the first-order autoregressive model, denoted AR(1), which satisfies the following difference equations,

$$\begin{aligned} &y_1: \text{ the initial value} \\ &y_t = \mu + \alpha\, y_{t-1} + z_t \quad (t > 1) \\ &z_t \sim i.i.d.\ N(0, \phi) \end{aligned} \tag{1}$$

where y_t is the observed value at time point t, α is the model autoregressive coefficient, μ is a parameter correlated with the mean of y, z is a shock variable, or a white noise sequence, satisfying a normal distribution with mean 0 and variance ϕ. In this case, the vector of population parameters to be estimated consists of $\boldsymbol{\theta} = (\mu, \alpha, \phi)'$. When $|\alpha| < 1$, there is a covariance stationary process for y_t satisfying Eq (1). Thus, the remainder of this discussion of AR(1) assumes that $|\alpha| < 1$.

Eq (1) is equivalent to the following equation

$$y_t = \alpha\, y_{t-1} + (\mu + z_t) \qquad (t > 1). \tag{2}$$

Let L be a lag operator which satisfying $L(y_t) = y_{t-1}$, then the solution of Eq (2) is

$$\begin{aligned} y_t &= (1 - \alpha L)^{-1}(\mu + z_t) \\ &= (\mu + z_t) + \alpha\,(\mu + z_{t-1}) + \alpha^2\,(\mu + z_{t-2}) + \alpha^3\,(\mu + z_{t-3}) + ... \\ &= \frac{\mu}{1-\alpha} + (z_t + \alpha\, z_{t-1} + \alpha^2 z_{t-2} + \alpha^3 z_{t-3} + ...) \end{aligned} \tag{3}$$

Because of the expression of (3) and $z_t \sim i.i.d.\ N(0, \phi)$, we know y_t follows a normal distribution with a mean

$$\begin{aligned} E(y_t) &= E(\frac{\mu}{1-\alpha}) + E(z_t + \alpha\, z_{t-1} + \alpha^2 z_{t-2} + \alpha^3 z_{t-3} + ...) \\ &= \frac{\mu}{1-\alpha}, \end{aligned} \tag{4}$$

the variance

$$\begin{aligned} \text{Var}(y_t) &= E(y_t - \frac{\mu}{1-\alpha})^2 \\ &= E(z_t + \alpha\, z_{t-1} + \alpha^2 z_{t-2} + \alpha^3 z_{t-3} + ...)^2 \\ &= (1 + \alpha^2 + \alpha^4 + \alpha^6 + ...)\phi \\ &= \frac{\phi}{1-\alpha^2}, \end{aligned} \tag{5}$$

and the j^{th} autocovariance

$$\begin{aligned} \text{Cov}(y_t, y_{t-j}) &= E(y_t - \frac{\mu}{1-\alpha})(y_{t-j} - \frac{\mu}{1-\alpha}) \\ &= E(z_t + \alpha\, z_{t-1} + \alpha^2 z_{t-2} + ... + \alpha^j z_{t-j} + \alpha^{j+1} z_{t-j-1} + ...)(z_{t-j} + \alpha\, z_{t-j-1} + \alpha^2 z_{t-j-2} + ...) \\ &= (\alpha^j + \alpha^{j+2} + \alpha^{j+4} + ...)\phi \\ &= \alpha^j \frac{\phi}{1-\alpha^2}. \end{aligned}$$

Based on the discussion above and with Eq (1), (4) and (5), we have the following distribution of y_t

$$\begin{cases} y_1 \sim N(\frac{\mu}{1-\alpha}, \frac{\phi}{1-\alpha^2}), \\ y_t | y_{t-1} \sim N(\mu + \alpha\, y_{t-1}, \phi), \quad (t > 1). \end{cases}$$

The Constant Coefficient AR(1) Model

We discussed the distribution of AR(1) model for a single subject. For multiple subjects, there are two types of AR(1) model: (1) constant coefficient AR(1) model and (2) random coefficient AR(1) model.

Suppose there are N individuals, the constant coefficient AR(1) model can be expressed as follows:

$$y_{it} = \mu + \alpha\, y_{i(t-1)} + z_{it},\ (i = 1, ..., N;\ t = 2, ..., T)$$

where $z_{it}\ \ i.i.d. \sim N(0, \phi)$ and the parameters μ, α and ϕ are constants which keep the same values across all individuals. This model is very useful when the sample size (or number of participants) is small but with fairly large measurement occasions.

The Random Coefficient AR(1) Model

Let $\boldsymbol{\beta}_i = (\mu_i, \alpha_i)'$, $\boldsymbol{\beta} = (\mu, \alpha)'$, $\mathbf{v}_i = (v_{0i}, v_{1i})'$ with $E(\mathbf{v}_i) = 0$ and $\mathrm{Cov}(\mathbf{v}_i) = \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$. The random coefficient AR(1) model can be expressed as follows:

$$\begin{cases} y_{it} = \mu_i + \alpha_i y_{i(t-1)} + z_{it}, & (i = 1, ..., N;\ t = 2, ..., T) \\ \mu_i = \mu + v_{0i}, & \\ \alpha_i = \alpha + v_{1i}, & \end{cases}$$

where $z_{it}\ \ i.i.d. \sim N(0, \phi)$. In this model, μ_i and α_i are random variables which are allowed to be different for different individuals. Therefore, random coefficients model sounds more reasonable for multiple subjects, but it requires a relatively large sample size.

Let $\mathbf{Y}_{i(t-1)} = (1, y_{i(t-1)})'$, then the conditional distribution of $(y_{it}|y_{i(t-1)})$ is

$$(y_{it}|y_{i(t-1)}) \sim N(\mathbf{Y}'_{i(t-1)}\boldsymbol{\beta}, \psi_{it})$$

where

$$\psi_{it} = \mathrm{Var}(y_{it}|y_{i(t-1)}) = (\mathbf{Y}_{i(t-1)})'\boldsymbol{\Sigma}(\mathbf{Y}_{i(t-1)}) + \phi \tag{6}$$

because $E(\boldsymbol{\beta}_i) = \boldsymbol{\beta}$ and $\mathrm{Cov}(\boldsymbol{\beta}_i) = \mathrm{Cov}(\mathbf{v}_i) = \boldsymbol{\Sigma}$.

Estimation Methods

Basically, we have two MLE estimation methods. Exact MLE estimation method: the parameter estimates are obtained by maximizing the exact log-likelihood function which includes the distribution of y_1, and conditional MLE estimation method: the parameter estimates are obtained by maximizing the conditional log-likelihood function which does not include the distribution of y_1. The differences between these two methods is that the exact MLE uses the information of first observation but requires stationarity assumption, but the conditional MLE does not uses the information of the the first observation and does not require stationarity assumption.

As described in the section of ”background and objectives”, for multiple subjects theoretically we should pool together all the likelihood functions for all individuals, then we use exact MLE and conditional MLE. But in practice pooling data is another method to deal with time series data. It assumes there is some relationship between y_{iT} and $y_{(i+1)1}$. Based on this data form, exact MLE and conditional MLE are obtained.

(1) Exact MLE

Exact MLE estimation method estimates the parameter by maximizing the exact log-likelihood function which includes the distribution of y_1.

(i) Pooled likelihood exact MLE for constant coefficients model.

The exact likelihood function of the stationary AR(1) model described in Eq (1) is

$$\begin{aligned} L_i(\alpha,\mu,\phi|\mathbf{y}_i) &= \prod_{t=1}^{T} p(y_{it}|\alpha,\mu,\phi) \\ &= \frac{1}{\sqrt{2\pi(\frac{\phi}{1-\alpha^2})}} \exp\left[-\frac{(y_{i1}-\frac{\mu}{1-\alpha})^2}{2\,(\frac{\phi}{1-\alpha^2})}\right] \left\{\prod_{t=2}^{T}\frac{1}{\sqrt{2\pi\phi}}\exp\left[-\frac{(y_{it}-\mu-\alpha\,y_{i(t-1)})^2}{2\phi}\right]\right\}, \end{aligned}$$

and its corresponding log likelihood function is

$$\begin{aligned} log(L) &= \sum_{i=1}^{N} log(L_i) \\ &= -\frac{N}{2}\,log(\frac{2\pi\phi}{1-\alpha^2}) - \frac{\sum_{i=1}^{N}(y_{i1}-\frac{\mu}{1-\alpha})^2}{2\,(\frac{\phi}{1-\alpha^2})} + \sum_{i=1}^{N}\sum_{t=2}^{T}\left[-\frac{1}{2}log(2\pi\phi) - \frac{(y_{it}-\mu-\alpha\,y_{i(t-1)})^2}{2\phi}\right] \\ &= \frac{N}{2}log(1-\alpha^2) - \frac{1-\alpha^2}{2\phi}\sum_{i=1}^{N}(y_{i1}-\frac{\mu}{1-\alpha})^2 - \frac{NT}{2}log(2\pi\phi) - \frac{1}{2\phi}\sum_{i=1}^{N}\sum_{t=2}^{T}(y_{it}-\mu-\alpha\,y_{i(t-1)})^2. \end{aligned}$$

In order to find the maximum likelihood estimates (MLE) of parameters μ, α and ϕ, we need to make all of their first order derivatives with respective to these parameters 0 and their corresponding second order derivatives negative. The MLE obtained through solving the exact likelihood function is called the exact MLE. Notice that

$$\begin{aligned} \frac{\partial log(L)}{\partial\mu} &= \frac{1+\alpha}{\phi}\sum_{i=1}^{N}(y_{i1}-\frac{\mu}{1-\alpha}) + \frac{1}{\phi}\sum_{i=1}^{N}\sum_{t=2}^{T}(y_{it}-\mu-\alpha\,y_{i(t-1)}) \\ \frac{\partial\,log(L)}{\partial\alpha} &= -\frac{N\alpha}{1-\alpha^2} + \sum_{i=1}^{N}\left\{\left(y_{i1}-\frac{\mu}{1-\alpha}\right)\frac{(1+\alpha)\mu}{(1-\alpha)\phi} + \left(y_{i1}-\frac{\mu}{1-\alpha}\right)^2\frac{\alpha}{\phi}\right\} \\ &\quad +\frac{1}{\phi}\sum_{i=1}^{N}\sum_{t=2}^{T}\left[(y_{it}-\mu-\alpha\,y_{i(t-1)})\,y_{i(t-1)}\right] \\ \frac{\partial\,log(L)}{\partial\phi} &= \frac{1-\alpha^2}{2\phi^2}\sum_{i=1}^{N}(y_{i1}-\frac{\mu}{1-\alpha})^2 - \frac{NT}{2\phi} + \frac{1}{2\phi^2}\sum_{i=1}^{N}\sum_{t=2}^{T}(y_{it}-\mu-\alpha\,y_{i(t-1)})^2, \end{aligned}$$

we can make them equal 0 at $\hat{\boldsymbol{\theta}}=(\hat{\mu},\hat{\alpha},\hat{\phi})$ to obtain the solution. Unfortunately, there is no simple solution for $\boldsymbol{\theta}$ in terms of $(\{y_{it}\}, 1\le i\le N, 1\le t\le T)$. But with the help of computers, we adopt iterative or numerical procedures to solve the equation.

(ii) Pooled data exact MLE for constant coefficients model.

The exact likelihood function of the connected stationary AR(1) model is

$$\begin{aligned} L(\alpha,\mu,\phi|\mathbf{y}) &= \prod_{t=1}^{NT} p(y_t|\alpha,\mu,\phi) \\ &= \frac{1}{\sqrt{2\pi(\frac{\phi}{1-\alpha^2})}} \exp\left[-\frac{(y_1-\frac{\mu}{1-\alpha})^2}{2\,(\frac{\phi}{1-\alpha^2})}\right] \left\{\prod_{t=2}^{NT} \frac{1}{\sqrt{2\pi\phi}} \exp\left[-\frac{(y_t-\mu-\alpha\, y_{t-1})^2}{2\phi}\right]\right\}, \end{aligned}$$

and its corresponding log likelihood function is

$$\begin{aligned} log(L) &= -\frac{1}{2}\, log(\frac{2\pi\phi}{1-\alpha^2}) - \frac{(y_1-\frac{\mu}{1-\alpha})^2}{2\,(\frac{\phi}{1-\alpha^2})} + \sum_{t=2}^{NT}\left[-\frac{1}{2}log(2\pi\phi) - \frac{(y_t-\mu-\alpha\, y_{t-1})^2}{2\phi}\right] \\ &= \frac{1}{2}log(1-\alpha^2) - \frac{1-\alpha^2}{2\phi}(y_1-\frac{\mu}{1-\alpha})^2 - \frac{NT}{2}log(2\pi\phi) - \frac{1}{2\phi}\sum_{t=2}^{NT}(y_t-\mu-\alpha\, y_{t-1})^2. \end{aligned}$$

Notice that

$$\begin{aligned} \frac{\partial log(L)}{\partial\mu} &= \frac{1+\alpha}{\phi}(y_1-\frac{\mu}{1-\alpha}) + \frac{1}{\phi}\sum_{t=2}^{NT}(y_t-\mu-\alpha\, y_{t-1}) \\ \frac{\partial\, log(L)}{\partial\alpha} &= -\frac{\alpha}{1-\alpha^2} + \left(y_1-\frac{\mu}{1-\alpha}\right)\frac{(1+\alpha)\mu}{(1-\alpha)\phi} + \left(y_1-\frac{\mu}{1-\alpha}\right)^2\frac{\alpha}{\phi} \\ &\quad + \frac{1}{\phi}\sum_{t=2}^{NT}\left[(y_t-\mu-\alpha\, y_{t-1})\, y_{t-1}\right] \\ \frac{\partial\, log(L)}{\partial\phi} &= \frac{1-\alpha^2}{2\phi^2}(y_1-\frac{\mu}{1-\alpha})^2 - \frac{NT}{2\phi} + \frac{1}{2\phi^2}\sum_{t=2}^{NT}(y_t-\mu-\alpha\, y_{t-1})^2, \end{aligned}$$

we make them equal to 0 at $\hat{\boldsymbol{\theta}} = (\hat{\mu},\hat{\alpha},\hat{\phi})$ to obtain the solution. Again, unfortunately, there is no simple solution for $\boldsymbol{\theta}$ in terms of $(\{y_{it}\}, 1\le i\le N, 1\le t\le T)$.

(2) Conditional MLE

An alternative approach to maximize the exact likelihood function is to regard the value of y_1 as deterministic and maximize the conditional likelihood function. The MLE obtained through solving the conditional likelihood function is called the conditional MLE.

(i) Pooled likelihood conditional MLE for constant coefficients model.

The conditional likelihood function of the stationary AR(1) model does not take the distribution of y_1 into consideration, so the function is

$$L_i(\alpha,\mu,\phi|\mathbf{y}_i) = \prod_{t=2}^{T} p(y_{it}|\alpha,\mu,\phi) = \prod_{t=2}^{T}\frac{1}{\sqrt{2\pi\phi}}\exp\left[-\frac{(y_{it}-\mu-\alpha\, y_{i(t-1)})^2}{2\phi}\right],$$

and its log likelihood function is

$$log(L(\alpha,\mu,\phi|\mathbf{y})) = \sum_{i=1}^{N} log(L_i) = -\frac{N(T-1)}{2}log(2\pi\phi) - \frac{1}{2\phi}\sum_{i=1}^{N}\sum_{t=2}^{T}(y_{it}-\mu-\alpha\, y_{i(t-1)})^2. \quad (7)$$

Similarly, to find the MLE of parameters μ, α and ϕ, first we need to obtain their first and second order derivatives. Since

$$\begin{aligned}
\frac{\partial log(L)}{\partial \mu} &= \frac{1}{\phi}\sum_{i=1}^{N}\sum_{t=2}^{T}(y_{it}-\mu-\alpha\, y_{i(t-1)}) \\
\frac{\partial\, log(L)}{\partial \alpha} &= \frac{1}{\phi}\sum_{i=1}^{N}\sum_{t=2}^{T}\left[(y_{it}-\mu-\alpha\, y_{i(t-1)})\, y_{i(t-1)}\right] \\
\frac{\partial\, log(L)}{\partial \phi} &= -\frac{N(T-1)}{2\phi}+\frac{1}{2\phi^2}\sum_{i=1}^{N}\sum_{t=2}^{T}(y_{it}-\mu-\alpha\, y_{i(t-1)})^2,
\end{aligned}$$

then by making them equal 0 at $\hat{\boldsymbol{\theta}} = (\hat{\mu}, \hat{\alpha}, \hat{\phi})$, we have

$$\hat{\mu} = \frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=2}^{T}(y_{it}-\hat{\alpha}\, y_{i(t-1)}) \tag{8}$$

$$\hat{\alpha} = \frac{\sum_{i=1}^{N}\sum_{t=2}^{T}\left[(y_{it}-\hat{\mu})y_{i(t-1)}\right]}{\sum_{i=1}^{N}\sum_{t=2}^{T}y_{i(t-1)}^2} \tag{9}$$

$$\hat{\phi} = \frac{1}{N(T-1)}\sum_{i=1}^{N}\sum_{t=2}^{T}(y_{it}-\hat{\mu}-\hat{\alpha}\, y_{i(t-1)})^2 \tag{10}$$

By combing Eq (8) and (9), the solution for (μ, α) is obtained as

$$\hat{\mu} = \frac{(\sum_{i=1}^{N}\sum_{t=2}^{T}y_{i(t-1)}^2)(\sum_{i=1}^{N}\sum_{t=2}^{T}y_{it}) - (\sum_{i=1}^{N}\sum_{t=2}^{T}y_{i(t-1)})(\sum_{i=1}^{N}\sum_{t=2}^{T}y_{i(t-1)}y_{it})}{N(T-1)\sum_{i=1}^{N}\sum_{t=2}^{T}y_{i(t-1)}^2 - (\sum_{i=1}^{N}\sum_{t=2}^{T}y_{i(t-1)})^2} \tag{11}$$

$$\hat{\alpha} = \frac{N(T-1)(\sum_{i=1}^{N}\sum_{t=2}^{T}y_{i(t-1)}y_{it}) - (\sum_{i=1}^{N}\sum_{t=2}^{T}y_{i(t-1)})(\sum_{i=1}^{N}\sum_{t=2}^{T}y_{it})}{N(T-1)\sum_{i=1}^{N}\sum_{t=2}^{T}y_{i(t-1)}^2 - (\sum_{i=1}^{N}\sum_{t=2}^{T}y_{i(t-1)})^2}. \tag{12}$$

And then by inserting (11) and (12) into Eq (10), we have the MLE of ϕ.

There is another simple way to obtain the estimate of $\boldsymbol{\theta} = (\mu, \alpha, \phi)$ using the ordinal least square (OLS) estimation method. Notice that maximizing (7) with respect to μ and α is equivalent to minimizing

$$\sum_{i=1}^{N}\sum_{t=2}^{T}(y_{it}-\mu-\alpha\, y_{i(t-1)})^2 \tag{13}$$

with respect to μ and α. Let $\mathbf{Y}_t$ be a $[N \times (T-1)]$-dimensional vector $\mathbf{Y}_t = (y_{12}, y_{13}, ..., y_{1T}, y_{22}, y_{23}, ..., y_{1T}, ..., y_{N2}, y_{N3}, ..., , y_{NT})'$, $\mathbf{Y}_{t-1}$ be a $[N \times (T-1)] \times 2$ matrix,

$$\mathbf{Y}_{t-1} = \begin{pmatrix} 1 & 1 & ... & 1 & 1 & 1 & ... & 1 & ... & 1 \\ y_{11} & y_{12} & ... & y_{1(T-1)} & y_{21} & y_{22} & ... & y_{2(T-1)} & ... & y_{N(T-1)} \end{pmatrix}'$$

and $\boldsymbol{\beta} = (\mu, \alpha)'$. Then minimizing (13) becomes minimizing

$$\mathbf{Y}_t - \mathbf{Y}_{t-1}\boldsymbol{\beta}$$

Thus, we have the OLS estimate of μ and α as follows,

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \begin{bmatrix}\hat{\mu}\\ \hat{\alpha}\end{bmatrix} = (\mathbf{Y}_{t-1}'\mathbf{Y}_{t-1})^{-1}(\mathbf{Y}_{t-1}'\mathbf{Y}_t) \\
&= \begin{bmatrix} N(T-1) & \sum_{i=1}^{N}\sum_{t=2}^{T}y_{i(t-1)} \\ \sum_{i=1}^{N}\sum_{t=2}^{T}y_{i(t-1)} & \sum_{i=1}^{N}\sum_{t=2}^{T}y_{i(t-1)}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{N}\sum_{t=2}^{T}y_{it} \\ \sum_{i=1}^{N}\sum_{t=2}^{T}y_{i(t-1)}y_{it} \end{bmatrix}.
\end{aligned}$$

Again, the MLE of ϕ can be obtained by inserting the MLE of (μ, α) into Eq (10).

The OLS solution is exactly the same as the MLE solution.

(ii) Pooled data conditional MLE for constant coefficients model.

The conditional likelihood function of the connected stationary AR(1) model is

$$\begin{aligned} L(\alpha,\mu,\phi|\mathbf{y}) &= \prod_{t=2}^{NT} p(y_t|\alpha,\mu,\phi) \\ &= \prod_{t=2}^{NT} \frac{1}{\sqrt{2\pi\phi}} \exp\left[-\frac{(y_t-\mu-\alpha\, y_{t-1})^2}{2\phi}\right], \end{aligned}$$

and its corresponding log likelihood function is

$$\begin{aligned} log(L) &= \sum_{t=2}^{NT}\left[-\frac{1}{2}log(2\pi\phi) - \frac{(y_t-\mu-\alpha\, y_{t-1})^2}{2\phi}\right] \\ &= -\frac{NT-1}{2}log(2\pi\phi) - \frac{1}{2\phi}\sum_{t=2}^{NT}(y_t-\mu-\alpha\, y_{t-1})^2. \end{aligned}$$

Notice that

$$\begin{aligned} \frac{\partial log(L)}{\partial\mu} &= \frac{1}{\phi}\sum_{t=2}^{NT}(y_t-\mu-\alpha\, y_{t-1}) \\ \frac{\partial\, log(L)}{\partial\alpha} &= \frac{1}{\phi}\sum_{t=2}^{NT}[(y_t-\mu-\alpha\, y_{t-1})\, y_{t-1}] \\ \frac{\partial\, log(L)}{\partial\phi} &= -\frac{NT-1}{2\phi} + \frac{1}{2\phi^2}\sum_{t=2}^{NT}(y_t-\mu-\alpha\, y_{t-1})^2, \end{aligned}$$

we make them equal to 0 at $\hat{\boldsymbol{\theta}} = (\hat{\mu},\hat{\alpha},\hat{\phi})$ to obtain the solution.

(iii) Pooled likelihood conditional MLE for random coefficients model.

The likelihood function for $(y_{it}|y_{i(t-1)})$ is

$$L_{it} = \frac{1}{\sqrt{2\pi\psi_{it}}}\exp[-\frac{(y_{it}-\mathbf{Y}'_{i(t-1)}\boldsymbol{\beta})^2}{2\psi_{it}}]$$

where ψ is defined in (6). Then the likelihood function for all participants at all time points is

$$\prod_{i=1}^{N}\prod_{i=2}^{T} L_{it} = \prod_{i=1}^{N}\prod_{i=2}^{T}\frac{1}{\sqrt{2\pi\psi_{it}}}\exp[-\frac{(y_{it}-\mathbf{Y}'_{i(t-1)}\boldsymbol{\beta})^2}{2\psi_{it}}]$$

We can also express it as a multivariate form. Let $\mathbf{Y}_i = (y_{i2},...,y_{iT})'$, $\mathbf{Y}_{i.} = \begin{pmatrix} 1 & ... & 1 \\ y_{i,1} & ... & y_{i(T-1)} \end{pmatrix}'$, then $E(\mathbf{Y}_i) = \mathbf{Y}_{i.}\boldsymbol{\beta}$, and $\mathrm{Var}(\mathbf{Y}_i) = \mathbf{\Psi}_i = \mathbf{Y}_{i.}\mathbf{\Sigma}\mathbf{Y}'_{i.} + \mathbf{\Phi}$ where $\mathbf{\Phi}$ is a $(T-1)\times(T-1)$ diagonal matrix with equal diagonal element ϕ. The conditional distribution of $(\mathbf{Y}_i|\mathbf{Y}_{i.})$ is

$$(\mathbf{Y}_i|\mathbf{Y}_{i.}) \sim MN_{T-1}(\mathbf{Y}_{i.}\boldsymbol{\beta},\mathbf{\Psi}_i)$$

The likelihood function for $(\mathbf{Y}_i|\mathbf{Y}_{i.})$ is

$$L_i = \frac{1}{(\sqrt{2\pi})^{T-1}} |\mathbf{\Psi}_i|^{-1/2} \exp[-\frac{1}{2}(\mathbf{Y}_i - \mathbf{Y}_{i.}\boldsymbol{\beta})' \mathbf{\Psi}_i^{-1} (\mathbf{Y}_i - \mathbf{Y}_{i.}\boldsymbol{\beta})],$$

and the likelihood function for all $\mathbf{Y}_i$ is

$$\prod_{i=1}^{N} L_i = \prod_{i=1}^{N} \frac{1}{(\sqrt{2\pi})^{T-1}} |\mathbf{\Psi}_i|^{-1/2} \exp[-\frac{1}{2}(\mathbf{Y}_i - \mathbf{Y}_{i.}\boldsymbol{\beta})' \mathbf{\Psi}_i^{-1} (\mathbf{Y}_i - \mathbf{Y}_{i.}\boldsymbol{\beta})]$$

The parameters we want to estimate are ϕ, $\boldsymbol{\beta}$ and $\mathbf{\Sigma}$. Let $\boldsymbol{\theta}$ be the all parameters vestor $\boldsymbol{\theta} = (\mu, \alpha, \phi, \sigma_1^2, \sigma_2^2, \sigma_{12})$. Like the ML method for the constant coefficient model, we can obtain its maximum likelihood estimates by making its first derivatives with respect to $\boldsymbol{\theta}$ zero and the second derivatives negative.

Simulation Studies

To investigate the performance of the exact MLE estimation method and conditional MLE estimation method fitting different models under different sceneries, we conduct three simulation studies: continuous constant coefficients model simulation study; categorical constant coefficients model simulation study; and continuous random coefficients model simulation study.

The R language is used to implement all these simulation studies. First, the simulation data as required are generated. Second, all parameters through four estimation methods are estimated. For the equations which have no explicit solutions, we adopt iterative or numerical procedures to solve them. Finally, estimation results obtained from all simulation replications are summarized, so that the conclusions are drawn.

Simulation Study 1: Continuous Constant Coefficients Model

Simulation Design.

In this simulation study, we generate continuous data and use them to fit the constant coefficients AR(1) model. We focus on the influence of initial value y_1 on the parameter estimation. There are three cases for the initial value y_1:

- a fixed $y_1 = 0$;
- a random y_1 from $N(0, \phi)$;
- a random y_1 from $N(\frac{\mu}{1-\alpha}, \frac{\phi}{1-\alpha^2})$.

Estimates are obtained using four estimation methods: pooled likelihood exact MLE; pooled likelihood conditional MLE; pooled data exact MLE; and pooled data conditional MLE.

In order to catch the change patterns, we investigate and compare estimates through different lengths of series and different number of subjects. The lengths of series and the number of subjects are $T = (5, 10, 15, 20, 30)$ and $N = (10, 20, 30, 40, 50)$, respectively.

The true values are $\mu = 0$, $\alpha = 0.5$, $\phi = 0.25$, and the replication number is 1000.

Simulation Results.

To save space, we only show several of all result tables which present the changing trend:

Insert Table 1 - 8 here.

Conclusions.

Through the simulation, we have the following conclusions: (1) The method of pooled data does not perform very well. The pooled likelihood function methods perform better; (2)For the case of random $y_1 \sim N(\frac{\mu}{1-\alpha}, \frac{\phi}{1-\alpha^2})$, the exact MLE method is the best. For the other two initial values, the conditional MLE method is the best; (3) The longer the time series, the more accurate the estimate; and (4) The more individual participated, the more accurate the estimate.

Simulation Study 2: Categorical Constant Coefficients Model

Simulation Design.

In this simulation study, we use the categorical data to fit the constant coefficients AR(1) model. Let nc be the number of categories we are studying. We use 3 steps to generate the categorical data,

- Step 1: Generate the continuous data according to the constant coefficients AR(1) model $y_{it} = \mu + \alpha y_{i(t-1)} + z_{it}$.
- Step 2: Generate thresholds $\boldsymbol{\tau} = (\tau_1, ..., \tau_{nc-1})$. With the assumption of the normality distribution of y_t, the thresholds $\boldsymbol{\tau}$ are created by (1) obtaining the standardized thresholds $\boldsymbol{\tau}_z$ by dividing the segment $[-2, 2]$ into $nc - 2$ parts evenly, and then (2) transform $\boldsymbol{\tau}_z$ to $\boldsymbol{\tau}$ according to the original data scale. For example, if $nc = 5$, then the standardized thresholds are $\boldsymbol{\tau}_z = (\tau_{z1}, \tau_{z2}, \tau_{z3}, \tau_{z4}) = (-2, -2/3, 2/3, 2)$, then

$$\boldsymbol{\tau} = \boldsymbol{\tau}_z \sigma_y + \mu_y$$

where $\mu_y = \frac{\mu}{1-\alpha}$ and $\sigma_y = \sqrt{\frac{\phi}{1-\alpha^2}}$.

- Step 3: Generate the categorical data yc_{it} by

$$\begin{cases} yc_{it} = 1, & \text{when } y_{it} \le \tau_1; \\ yc_{it} = k, & \text{when } \tau_{k-1} < y_{it} \le \tau_k; \\ yc_{it} = nc, & \text{when } y_{it} > \tau_{nc-1}. \end{cases}$$

Obviously, the scale of yc_{it} is from 1 to nc.

Let $\boldsymbol{\pi}$ be a nc-dimensional vector $\boldsymbol{\pi} = (\pi_1, ..., \pi_{nc})$ which is defined as

$$\begin{aligned} \pi_1 &= \Phi(\tau_1) \\ \pi_k &= \Phi(\tau_k) - \Phi(\tau_{k-1}) \quad (2 \le k \le nc - 1) \\ \pi_{nc} &= 1 - \Phi(\tau_{nc-1}). \end{aligned}$$

So each π_k $(1 \le k \le nc)$ is defined to be the probability of corresponding k^{th} category. With $\boldsymbol{\pi}$, the mean of nc categories is

$$M_c = \sum_{k=1}^{nc} \pi_k \, k$$

Then the true μ and ϕ of nc categories are

$$\begin{aligned} \mu_c &= M_c \, (1 - \alpha), \\ \phi_c &= \left[\sum_{k=1}^{nc} \pi_k \, (k - M_c)^2 \right] (1 - \alpha^2), \end{aligned}$$

respectively.

When the categorical data are ready, the same four estimation methods as those in simulation study 1 are used to estimate the model. The other factors examined in this study include the initial value, the number of categories, the lengths of series and the number of subjects:

• The initial value y_1 has three cases: (i) a fixed yc_1 based on a fixed $y_1 = 0$; (ii) a random yc_1 based on a random y_1 from $N(0, \phi)$; and (iii) a random yc_1 based on a random y_1 from $N(\frac{\mu}{1-\alpha}, \frac{\phi}{1-\alpha^2})$.

• The number of categories: $nc = (5, 7, 9)$.

• The lengths of series and the number of subjects: $T = (5, 10, 15, 20, 30, 40, 50)$ and $N = (50, 100, 150, 200)$.

The true values are $\mu = \mu_c$, $\alpha = 0.5$, $\phi = \phi_c$, and the replication number is 1000.

Simulation Results.

Again, to save space, we only present several result tables which show the changing trend:

Insert Table 9 - 26 here.

Conclusions.

Through the simulation, we have the following conclusions: (1) The pooled likelihood methods perform better than the pooling data methods; (2) For the case of random $y_1 \sim N(\frac{\mu}{1-\alpha}, \frac{\psi}{1-\alpha^2})$, the pooled likelihood exact MLE are the best. For the other two initial values, the pooled likelihood conditional MLE are the best; (3) The more categories, the more accurate the estimate; (4) The longer the time series, the more accurate the estimate; (5) The more individual participated, the more accurate the estimate; and (6) For categorical data, the coverage probabilities are poor. This is because the categorical data seem more "flat" than the corresponding continuous data, so autoregressive coefficient for categorical time series is smaller than the "true" autoregressive coefficient which is for the continuous time series.

Simulation Study 3: Continuous Random Coefficients Model

Simulation Design.

From the previous two simulation studies, we conclude that the pooled likelihood conditional MLE are the best if the initial values are not from the true distribution of y_t. In other word, the pooled likelihood conditional MLE is not sensitive to initial values. Therefore, in the simulation study of continuous random coefficients model, we mainly focus on the method of pooled likelihood conditional estimation.

In order to catch the change patterns, we investigate and compare the estimates through different lengths of series $T = (10, 15, 20, 30, 40, 50)$ and different number of subjects $N = (50, 100, 150, 200)$. We reduce the influence of the initial values by deleting the first 50 observations when we generate the data.

The true values are $\phi = 25$, $\mu = 16$, $\alpha = 0.4$,$\Sigma = \begin{pmatrix} 30 & -0.5 \\ -0.5 & 0.02 \end{pmatrix}$. And the replication number is still 1000.

Simulation Results.

Several result tables which show the changing trend are presented as follows:

Insert Table 27 - 50 here.

Conclusions.

With the simulation results, we have the following rough conclusions: (1) The conditional MLE method performs well; (2) The longer the time series, the more accurate the estimate; (3) The more individual participated, the more accurate the estimate.

A Real Data Analysis

Data Description

The data were collected in Fall 1991 or Spring 1992 at the University of Illinois. They are daily self-reports of students' emotional experiences for 52 consecutive days.

Totally, there are 153 participants, 52 occasions. There are 2 variables include the pleasant affect (PA) and the unpleasant affect (UA). The variable PA is combined from 8 items: love, affection, caring, fondness, joy, happiness, contentment, and satisfaction. And the variable UA is combined from 8 items: depression, unhappiness, shame, nervousness, loneliness, sadness, anxiety, and irritation. For each item, there is 7-point Likert-type scale, so the total scale for PA or UA is $(1-7)\times 8 = (8-56)$.

Estimation

Considering the sample size is quite large (N=153), we use the conditional random coefficients model to fit this real data set. The results for PA and UA variables are as follows:

	PA		UA	
	Estimate	SE	Estimate	SE
fixed effect:				
intercept μ	16.035	0.640	9.197	0.299
slope α	0.428	0.018	0.388	0.016
random effect:				
$\text{Var}(z_{it}) = \phi$	24.943	1.472	11.771	0.704
$\text{Var}(\mu_i) = \sigma_1^2$	30.175	6.552	7.480	1.767
$\text{Var}(\alpha_i) = \sigma_2^2$	0.021	0.003	0.020	0.004
$\text{Cov}(\mu_i, \alpha_i) = \sigma_{12}$	-0.495	0.119	-0.208	0.078

Conclusions

From the results, we can see that the pleasant affect has a higher intercept than the unpleasant affect which means in the first semester students are happier. The pleasant affect also has a larger variance of the white noise which means it has a bigger fluctuation than the unpleasant affect, so it is easier to be unstable. The two slopes of these two variables are very close to each other which means the changing trends are similar for both pleasant affect and unpleasant affect.

Discussions

We analyzed the simplest AR(1) model. We can extend the analysis to AR(p) model

$$y_{it} = \mu_i + \sum_{l=1}^{p} \alpha_{il} y_{i(t-l)} + z_{it}$$

We can also extent the study to the other dynamic system models, such as the damped oscillation model which is described as follows

$$y''_{it} = a y'_{it} + b y_{it} + z_{it}.$$