R package details for: Toy data generation for Bayesian likelihood regression-based estimation

Dr. Weichang Yu

June 14, 2025

Key words: Dynamic treatment regimes, Parallel computing, Linear regression, Bayesian inference

Preamble notations

Our observed data is $\mathcal{D} = \{y_i, \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, a_{i1}, \ldots, a_{iT}\}_{i=1}^n$, where the final outcome is $y_i \in \mathbb{R}$, the intermediate covariates is $\mathbf{x}_{it} \in \mathbb{R}^{p_t}$ at time $t = 2, \ldots, T$, $\mathbf{x}_{i1} \in \mathbb{R}^{p_1}$ is the baseline covariates, and $a_{it} \in \mathcal{A}_t$ denotes the assigned treatment at time t. For example, in a study on optimal drug treatment assignment for Type II diabetes patients, $\mathbf{x}_{it}$ may denote the blood pressure, HbA1c, BMI, comorbidities index of the i-th patient at follow-up clinic visit t and y_i denote the final HbA1c reading after T clinic follow-ups. We assume that each $\mathcal{A}_t$ is a finite set, i.e., $\mathcal{A}_t = \{1, \ldots, |\mathcal{A}_t|\}$. We denote the standard t-distribution with df degree of freedom as t_{df}. We denote the multivariate t-distribution with location $\boldsymbol{\mu}$, scale matrix $\mathbf{S}$, and degree of freedom ν by $t_\nu(\boldsymbol{\mu}, \mathbf{S})$. We use ":" to denote contiguity, for example, $\mathbf{x}_{i;1:t} = (x_{i1}, \ldots, x_{it})^\top$, $a_{i;1:t} = (a_{i1}, \ldots, a_{it})^\top$, and $\mathbf{x}_{1:n;t} = (x_{1t}, \ldots, x_{nt})$.

Generate univariate test dataset ($p_t = 1$)

Fix $n = 5000$, $T = 5$, $p_t = 1$ and $|\mathcal{A}_t| = 3$ for all $t = 1, \ldots, T$,

1. Generate $\mathbf{x}_{i1} \sim t_{10}$, where t_{df} denote the t-distribution with df degrees of freedom.

2. Generate a_{it} with equal probabilities from $\mathcal{A}_t$.

3. For each $t = 2, \ldots, T$, generate

$$\begin{aligned}\mathbf{x}_{it} &= \mathbb{I}\{a_{i;t-1} = 2\}\{t\mathbf{x}_{i;t-1} - (t-1)\mathbf{x}_{i;t-2} + (t-2)\mathbf{x}_{i;t-3}\} \\ &\quad + \mathbb{I}\{a_{i;t-1} = 3\}\left\{-t\mathbf{x}_{i;t-1} + \sqrt{t-1}\mathbf{x}_{i;t-2} + \sqrt{t-2}\mathbf{x}_{i;t-3}\right\} + \xi_{it}\end{aligned}$$

   where $\xi_{it} \sim \mathrm{N}(0, 0.5^2)$ and $\mathbf{x}_{it} = \mathbf{0}$ if $t < 1$.

4. Generate

$$y_i \sim \mathrm{N}(m_i(\mathbf{x}_{i;1:T}, a_{i;1:T}), 1)$$

where (standardize $\mathbf{x}$'s first)

$$
\begin{aligned}
m_i(\mathbf{x}_{i;1:T}, a_{i;1:T}) = 3 &+ \sum_{t=1}^{T} \mathbb{I}\{a_{it} = 2\} \{\sin(10t)\mathbf{x}_{i;t} - \sin(10t-10)\mathbf{x}_{i;t-1} + \sin(10t-20)\mathbf{x}_{i;t-2}\} \\
&+ \sum_{t=1}^{T} \mathbb{I}\{a_{it} = 3\} \left\{\cos(10t)\mathbf{x}_{i;t} - \cos(10t-10)\mathbf{x}_{i;t-1} + \sqrt{|\cos(10t-20)|}\mathbf{x}_{i;t-2}\right\}
\end{aligned}
$$

and $\mathbf{x}_{it} = \mathbf{0}$ if $t < 1$.

Generate multivariate test dataset ($p_t > 1$)

Obtain user-input for n, T, p_t, and $|\mathcal{A}_t|$ for all $t = 1, \dots, T$.

1. For each $i = 1, \dots, n$, generate $\mathbf{x}_{i1} \sim t_{10}(\mathbf{0}, \mathbf{I})$, where t_{df} denote the multivariate t-distribution with df degrees of freedom.

2. For each $i = 1, \dots, n$ and $t = 1, \dots, T$, generate a_{it} with equal probabilities from $\mathcal{A}_t$.

3. For each $i = 1, \dots, n$ and $t = 2, \dots, T$, generate

$$
\begin{aligned}
\mathbf{x}_{it} = &\mathbb{I}\{a_{i;t-1} = 2\} \left\{t\mathbf{C}_{p_t \times p_{t-1}}\mathbf{x}_{i;t-1} - (t-1)\mathbf{C}_{p_t \times p_{t-2}}\mathbf{x}_{i;t-2} + (t-2)\mathbf{C}_{p_t \times p_{t-3}}\mathbf{x}_{i;t-3}\right\} \\
&+ \mathbb{I}\{a_{i;t-1} = 3\} \left\{-t\mathbf{C}_{p_t \times p_{t-1}}\mathbf{x}_{i;t-1} + \sqrt{t-1}\mathbf{C}_{p_t \times p_{t-2}}\mathbf{x}_{i;t-2} + \sqrt{t-2}\mathbf{C}_{p_t \times p_{t-3}}\mathbf{x}_{i;t-3}\right\} + \xi_{it}
\end{aligned}
$$

   where $\xi_{it} \sim \text{MVN}(\mathbf{0}, 0.5^2\mathbf{I})$, $\mathbf{x}_{it} = \mathbf{0}$ if $t < 1$ and $\mathbf{C}_{a \times b} = \{c_{rs}\}_{1 \le r \le a; 1 \le s \le b}$ is a a by b matrix such that the (r, s) entry is $c_{rs} = (-1)^{r+s}$.

4. Generate

$$
y_i \sim \text{N}(m_i(\mathbf{x}_{i;1:T}, a_{i;1:T}), 1)
$$

   where (standardize $\mathbf{x}$'s first)

$$
\begin{aligned}
m_i(\mathbf{x}_{i;1:T}, a_{i;1:T}) = 3 &+ \sum_{t=1}^{T} \mathbb{I}\{a_{it} = 2\} \left\{\sin(10t)\mathbf{x}_{i;t}^{\top}\mathbf{1} - \sin(10t-10)\mathbf{x}_{i;t-1}^{\top}\mathbf{1} + \sin(10t-20)\mathbf{x}_{i;t-2}^{\top}\mathbf{1}\right\} \\
&+ \sum_{t=1}^{T} \mathbb{I}\{a_{it} = 3\} \left\{\cos(10t)\mathbf{x}_{i;t}^{\top}\mathbf{1} - \cos(10t-10)\mathbf{x}_{i;t-1}^{\top}\mathbf{1} + \sqrt{|\cos(10t-20)|}\mathbf{x}_{i;t-2}^{\top}\mathbf{1}\right\}
\end{aligned}
$$

   and $\mathbf{x}_{it} = \mathbf{0}$ if $t < 1$.

References