区間打切りデータに基づく情報量規準と感染症の潜伏期間推定への応用

神戸薬科大学 阿部興

2020 年 5 月 31 日

背景

- 感染症の潜伏期間の分布は, 防疫上の様々な施策の基礎となる情報
- 発症した時点が特定される場合でも, 感染した日時が観測されることはまれ
- Backer *et al.* (2020) はこの問題に対する 1 つのアプローチを提示
- 新型コロナウイルス（COVID-19）感染症は最初, 中国の武漢で発生
- Backer *et al.* (2020) は武漢への旅行者らが武漢に滞在した期間と, 発症した日のデータを用いて, 感染した時点をある種の潜在変数として扱い, 潜伏期間の分布を推定する方法を提案
- その上で, leave-one-out (loo) 情報量規準を用いて, ワイブル分布, ガンマ分布, 対数正規分布を比較し, COVID-19 感染症の潜伏期間に対してワイブル分布がよく適合すると論じた

本報告の内容

- Backer *et al.* (2020) の方法論は生存時間分析の分野で区間打切り（interval censored）とよばれる観測を扱う問題と等価
- Backer *et al.* (2020) の用いた方法では, loo 情報量がモデルの汎化誤差の近似としてバイアスを持ち, モデル選択の方法として妥当性が十分でないことをシミュレーションを用いて示す
- COVID-19 感染症の潜伏期間は, Backer *et al.*(2020) が推定したものよりも長い可能性

データ

- 2020 年 1 月 20 日から 28 日の報告
- 88 症例
- COVID-19 感染症の患者の武漢への滞在履歴と, 発症した日が記録
- 感染した日は未知
- 武漢への滞在をリスク因子への暴露とみなす

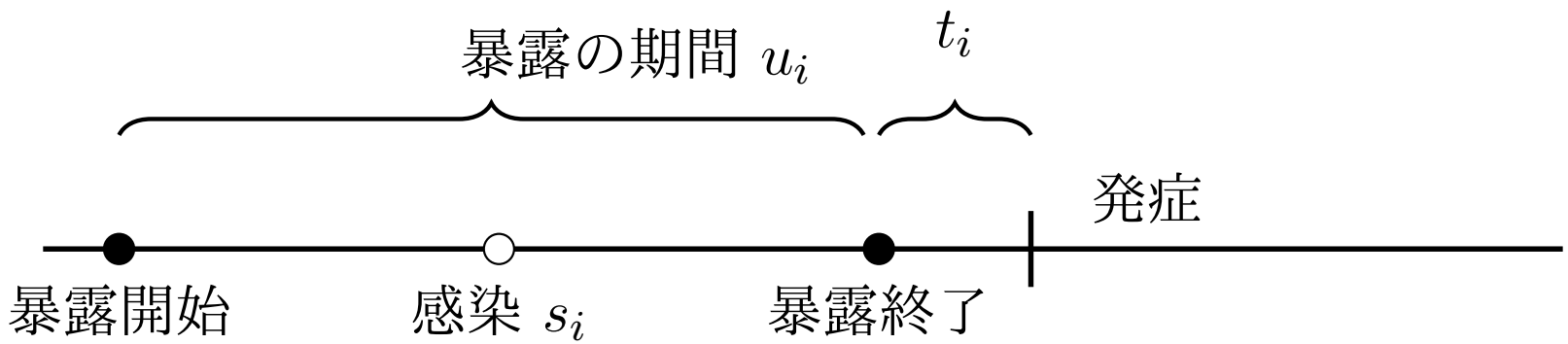

Figure: 本報告で扱う観測

記法

- 患者 i が発症した日: 原点 0
- 原点からさかのぼり, 暴露が終了した日を t_i 日前とする.
- 暴露の終了から s_i 日前を感染した日とする.
- 暴露していた期間を u_i とする.

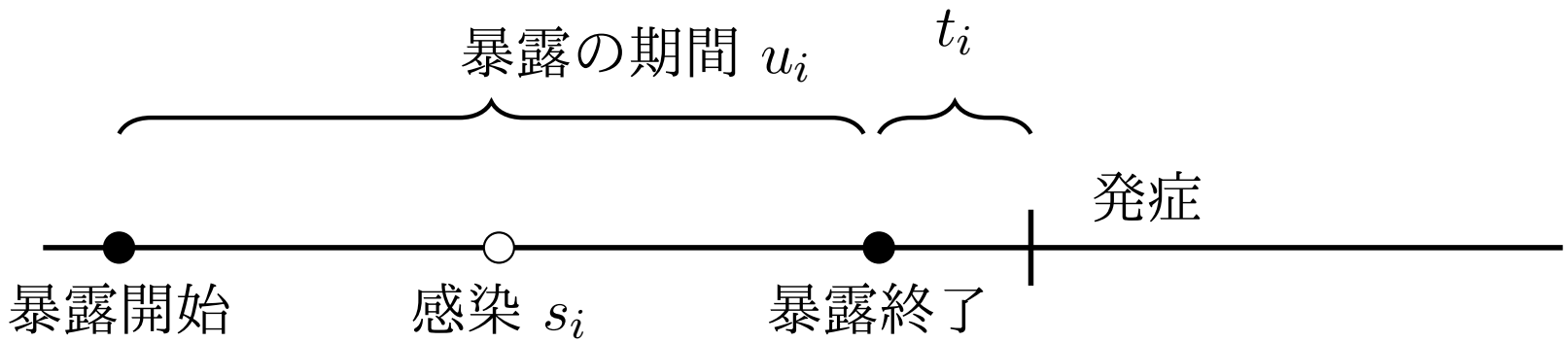

Figure: 本報告で扱う観測

Backer *et al.* (2020) の用いた方法

仮に感染した時点 s_i についての完全な観測が得られた場合, 発症までの待ち時間の確率密度は $f(s_i + t_i)$. ここで $f(x)$ は確率密度関数.

- 尤度

$$L = \prod_{i=1}^{n} f(s_i + t_i)$$

- 事前分布

$$s_i \sim \mathrm{Uniform}(0, u_i)$$

Backer *et al.* (2020) は, 確率密度関数 $f(x)$ にワイブル分布, ガンマ分布, および対数正規分布を選び, s_i に区間 $[0, u_i]$ の一様事前分布を採用することで, $f(x)$ のパラメータと s_i をあわせて推定.

これを Backer 型推定と呼ぶことにする.

s_i を積分消去する場合

感染した時点 s_i を尤度から積分消去する方法も考えられる（この方法を Turnbull 型推定と呼ぶことにする）.

s_i を積分消去する場合, 尤度を構成する因子の患者 i に関する部分は:

$$\int_0^{u_i} f(s_i + t_i)\, ds_i.$$

$x_i = s_i + t_i$ と改めておくと:

$$\int_{t_i}^{u_i+t_i} f(x_i)\, dx_i$$

これは Turnbull (1976) が論じた区間打切りデータに基づく尤度と同じ.

3種類の異なる観測

(1)-(3) 式をすべてかけあわせたものが尤度関数.

1. $u_i = 0$ のとき, 尤度を構成する因子 L_i は:

$$L_i = f(t_i). \tag{1}$$

2. u_i が有限の正の実数のとき, 尤度を構成する因子 L_i は:

$$L_i = F(u_i + t_i) - F(t_i). \tag{2}$$

3. u_i が不明のとき, 尤度を構成する因子 L_i は:

$$L_i = 1 - F(t_i). \tag{3}$$

ここで $F(x)$ は $f(x)$ と対応する分布関数.

3 種類の異なる観測

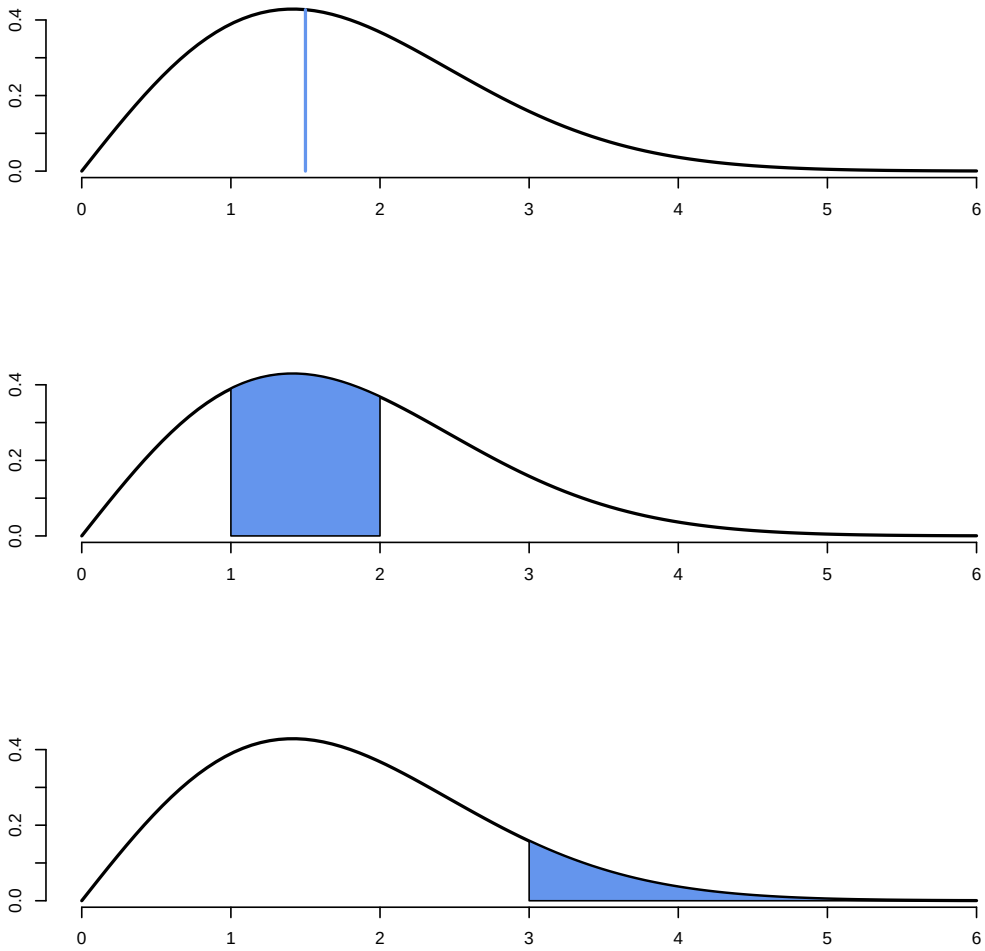

Figure: 3 種類の観測

注意

第 3 の観測（u_i が不明）の場合を Backer 型推定では扱うことができない.

Backer *et al.* (2020) では u_i に適当な大きな数を与えることで推定を行っている.

loo 情報量

- θ: すべての未知パラメータ
- $\phi(\theta)$: 事前分布の密度関数
- $p(x|\theta)$: 評価の対象となる確率モデルの密度関数
- $\phi^*(\theta)$: 事後分布の密度関数
- $\phi_k^*(\theta)$: 得られたサンプル $x_1, x_2, \ldots, x_n$ から 1 つのサンプル x_k を除いてできるデータから実現された事後分布の確率密度

loo 情報量:

$$\mathrm{LOOIC} = -\sum_{k=1}^{n} \log\left(\int p(x_k|\theta)\phi_k^*(\theta)\, d\theta.\right)$$

loo 情報量

$$
\begin{aligned}
\mathrm{LOOIC} &= -\sum_{k=1}^{n} \log \left(\frac{\int \phi(\theta) p(x_k|\theta) \prod_{i \neq k} p(x_k|\theta)\, d\theta}{\int \phi(\theta) \prod_{i \neq k} p(x_i|\theta)\, d\theta} \right) \\
&= -\sum_{k=1}^{n} \log \left(\frac{\int \phi(\theta) \prod_{i=1}^{n} p(x_i|\theta)\, d\theta}{\int \phi(\theta) p(x_i|\theta)^{-1} \prod_{i=1}^{n} p(x_i|\theta)\, d\theta} \right) \\
&= \sum_{k=1}^{n} \log \left(\frac{\int \phi(\theta) p(x_k|\theta)^{-1} \prod_{i=1}^{n} p(x_i|\theta)\, d\theta}{\int \phi(\theta) \prod_{i=1}^{n} p(x_i|\theta)\, d\theta} \right) \\
&= \sum_{k=1}^{n} \log \left(\int p(x_k|\theta)^{-1} \phi^*(\theta)\, d\theta \right)
\end{aligned}
$$

サンプル 1 つあたりの尤度の逆数を事後分布により平均したもので, loo 情報量が計算できる.

汎化損失

汎化損失:

$$\mathrm{GE} = -\int q(x) \log p(x)\, dx.$$

ここで $q(x)$ はデータを生成した分布, $p(x)$ は予測分布の密度関数である.

$$\mathrm{GE} = \int q(x) \log \frac{q(x)}{p(x)}\, dx - \int q(x) \log q(x)\, dx.$$

- 第 1 項: $q(x)$ と $p(x)$ のカルバック・ライブラ情報量
- 第 2 項: データを生成した分布 $q(x)$ のみによって定まる. 予測分布 $p(x)$ の選び方に依存しない.

汎化損失が小さいほどデータを生成した分布に近い予測分布が得られている.

注意

- loo 情報量は汎化損失の近似となる（渡辺, 2012）
- 一般にデータを生成した分布 $q(x)$ は未知
- 汎化損失は直接計算することができない
- そのため, loo 情報量規準が汎化損失の近似となることが重要

ここで生じる疑問

サンプルごとにパラメータ（潜在変数）を持つようなモデルの場合, loo 情報量は汎化誤差の近似になるのか？

Table: leave-one-out

train	train	test
train	test	train
test	train	train

サンプル 1 つあたりの尤度

Backer 型推定と Turnbull 型推定は本質的には同じと考えられる.
しかし「サンプル 1 つあたりの尤度」が異なる.

- Backer 型推定のサンプル 1 つあたりの尤度:

$$f(s_i + t_i)$$

- Turnbull 型推定のサンプル 1 つあたりの尤度:

$$\int_{t_i}^{u_i+t_i} f(x_i)\, dx_i.$$

シミュレーションの設定

Turnbull 型推定と Backer 型推定について, loo 情報量をシミュレーション.

- Backer *et al.* (2020) に習い, すべてのパラメータの事前分布に一様分布（フラットプライヤ）を採用
- 事後分布の実現には確率的プログラミング言語 Stan を用いた
- データを生成する分布: ワイブル分布（形状パラメータ, 尺度パラメータともに 2）
- 評価の対象となるモデルの確率分布: ワイブル分布
- シミュレーションの試行回数: 100 回
- サンプルサイズ: $n = 25, 50, 100$
- 区間打切りは生成したデータの小数点以下を切り落とすことにより *, 人工的に生成

*例を上げると 1.5 という乱数が生成された場合, これを区間 $[1, 2]$ の区間打切りデータとして扱う

シミュレーション結果

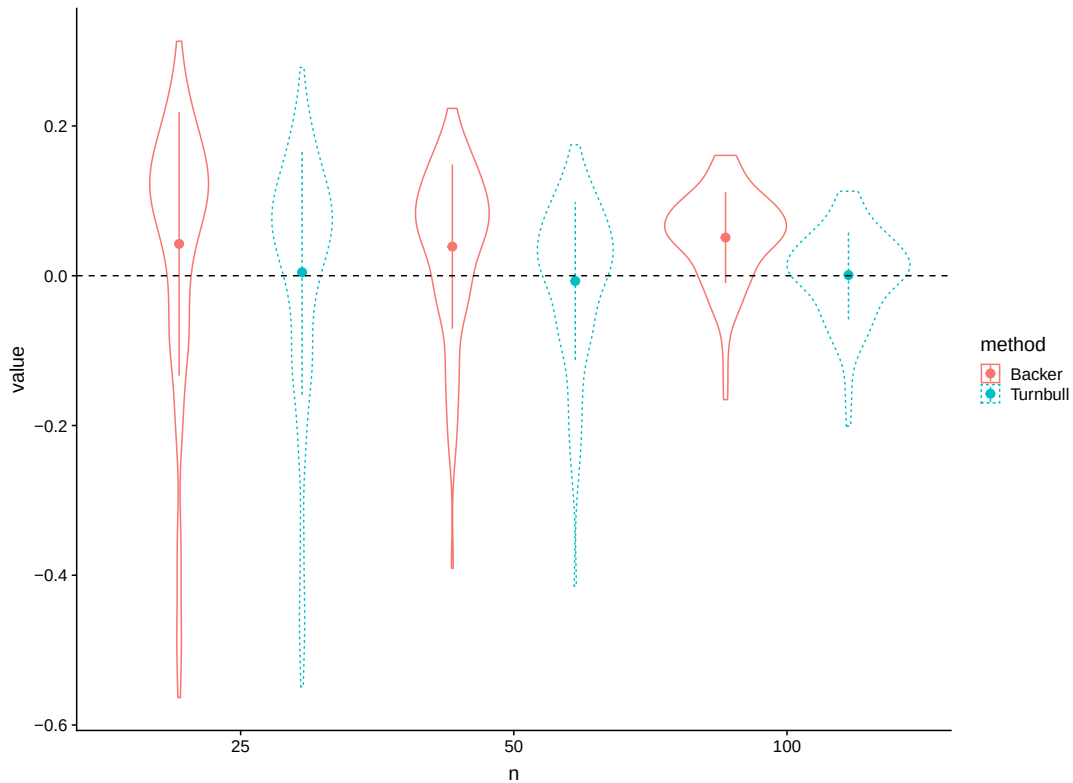

Figure: シミュレーション結果. エラーバーは標準偏差, 点は平均を表す.

図中の value は $\mathrm{GE} - \mathrm{LOOIC}\,/n$.

パラメータの推定量としての違い

Table: 推定量（事後期待値）の平均

	形状パラメータ		尺度パラメータ	
n	Turnbull	Backer	Turnbull	Backer
25	2.18	2.18	2.08	2.08
50	2.09	2.09	2.03	2.03
100	2.03	2.03	2.01	2.01

Table: 推定量（事後期待値）の標準誤差

	形状パラメータ		尺度パラメータ	
n	Turnbull	Backer	Turnbull	Backer
25	0.45	0.45	0.23	0.23
50	0.30	0.30	0.18	0.18
100	0.20	0.20	0.12	0.12

シミュレーション結果に対する考察

- Backer 型推定では loo 情報量は汎化損失の推定量としてバイアスを持つ
- バイアスはサンプルサイズを大きくしても小さくならない
- Turnbull 型推定はより正確に汎化損失を近似

実データでの例

Backer *et al.* (2020) と同じデータを用いて Turnbull 型推定により loo 情報量を計算した.

Table: Turnbull 型推定

分布	loo 情報量
ワイブル	73.89
ガンマ	73.35
対数正規	73.32

Table: Backer *et al.* (2020) より

分布	loo 情報量
ワイブル	486
ガンマ	545
対数正規	592

この表での loo 情報量は LOOIC の 2 倍.

潜伏期間の分布

対数正規分布が最も裾の重い予測を与える.

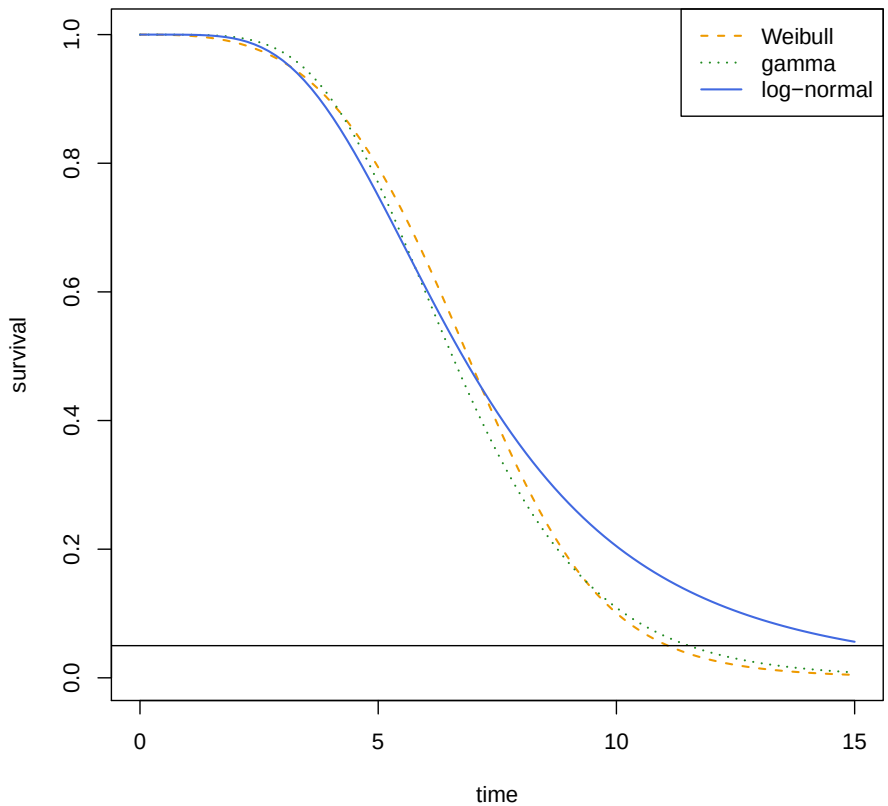

Figure: 潜伏期間の生存関数

予測区間

Table: 潜伏期間の予測区間

	2.5%	50%	97.5%
ワイブル	2.52	6.86	12.11
ガンマ	2.97	6.55	12.89
対数正規	2.66	6.78	18.99

参考: Backer *et al.* (2020) の示した 95%予測区間は $[2.1, 11.1]$

議論

- 潜在変数がある場合の情報量規準の使用には注意が必要
- COVID-19 感染症の潜伏期間は, Backer *et al.*(2020) の推定より長い可能性
- 潜伏期間の情報は確立されたものではなく, さらなる議論が必要なもの

参考文献

- Backer, J. A., Klinkenberg, D., & Wallinga, J. (2020). Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20-28 January 2020. *Eurosurveillance*, 25(5), 2000062.
- Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. Journal of the Royal Statistical Society: Series B (Methodological), 38(3), 290-295.
- 渡辺澄夫 (2012). ベイズ統計の理論と方法. コロナ社.

シミュレーションの設定

Turnbull 型推定と Backer 型推定について, loo 情報量をシミュレーション.

- Backer *et al.* (2020) に習い, すべてのパラメータの事前分布に一様分布（フラットプライヤ）を採用
- 事後分布の実現には確率的プログラミング言語 Stan を用いた
- データを生成する分布: ガンマ分布（形状パラメータ 2, レートパラメータ 0.25）
- 評価の対象となるモデルの確率分布: ガンマ分布
- シミュレーションの試行回数: 100 回
- サンプルサイズ: $n = 25, 50, 100$
- 区間打切りは生成したデータの小数点以下を切り落とすことにより, 人工的に生成

ガンマ分布の場合

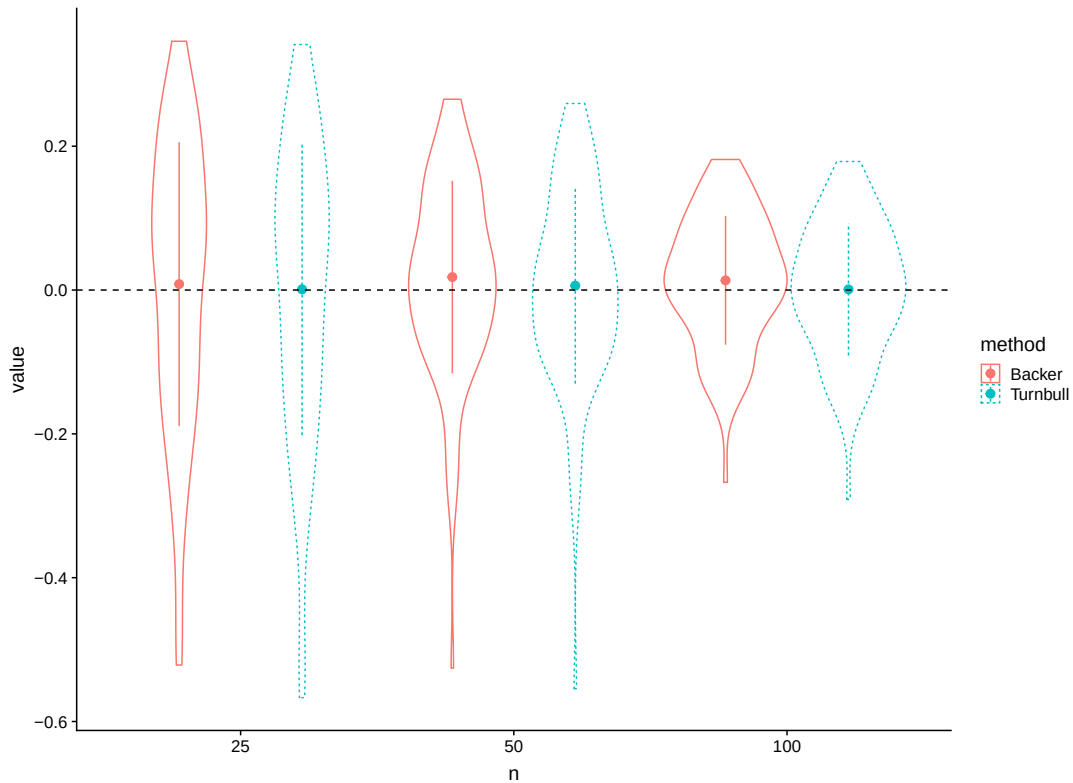

Figure: シミュレーション結果. エラーバーは標準偏差, 点は平均を表す.

図中の value は $\mathrm{GE} - \mathrm{LOOIC}\,/n$.

シミュレーションの設定

Turnbull 型推定と Backer 型推定について, loo 情報量をシミュレーション.

- Backer *et al.* (2020) に習い, すべてのパラメータの事前分布に一様分布（フラットプライヤ）を採用
- 事後分布の実現には確率的プログラミング言語 Stan を用いた
- データを生成する分布: 対数正規分布（平均パラメータ 0 , 分散パラメータ 1）
- 評価の対象となるモデルの確率分布: 対数正規分布
- シミュレーションの試行回数: 100 回
- サンプルサイズ: $n = 25, 50, 100$
- 区間打切りは生成したデータの小数点以下を切り落とすことにより, 人工的に生成

対数正規分布の場合

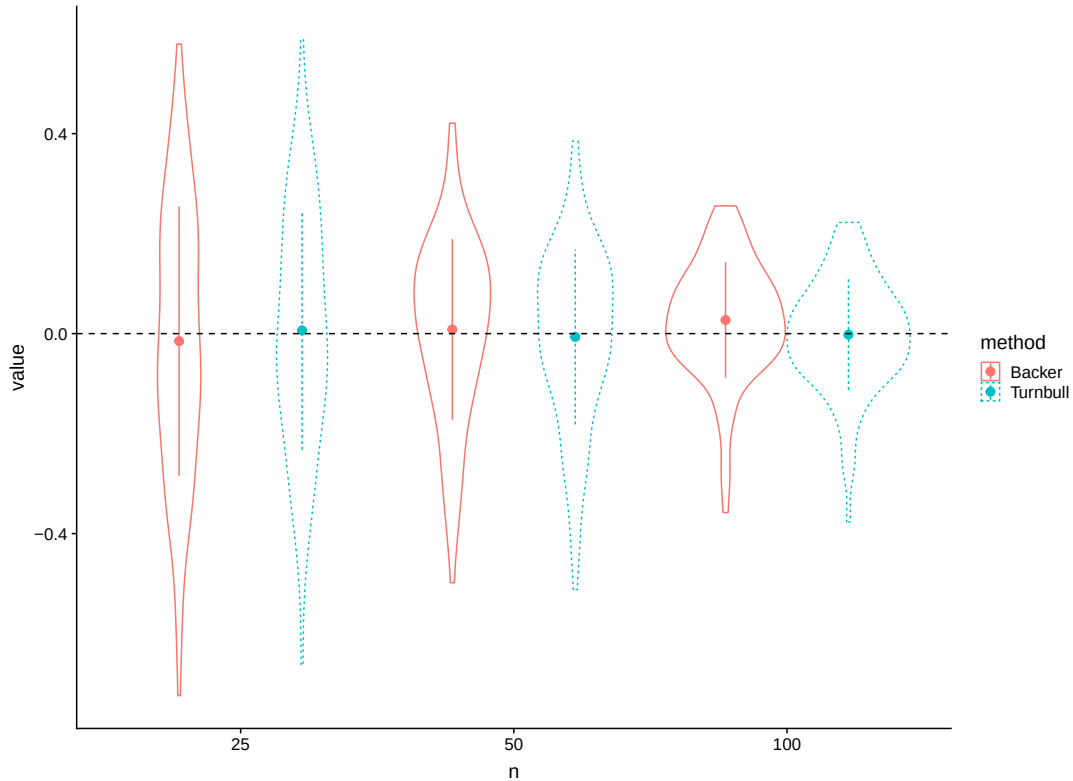

Figure: シミュレーション結果. エラーバーは標準偏差, 点は平均を表す.

図中の value は $\mathrm{GE} - \mathrm{LOOIC}\,/n$.