経路選択問題におけるモデルとサンプリング

Specification of the cross-nested logit model with sampling of alternatives for route choice models

Xinjun Lai, Michel Bierlaire

Transportation Research Part B 80 (2015) 220-234

夏学期ゼミ #11

交通研B4 村橋拓真

目次

経路選択問題におけるモデルとサンプリング

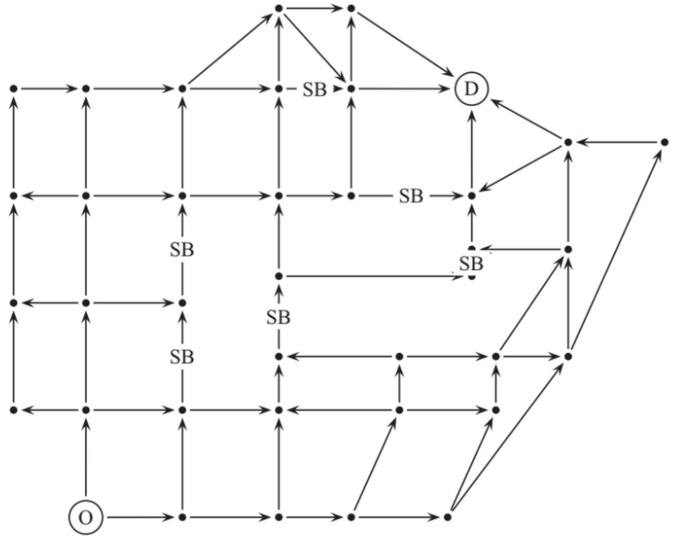
今回はGEVモデルとサンプリング手法の背景的知識はだいぶ縮小して説明しています.
詳しくは補足資料にまとめておくので参考に…
過去のbinの資料にわかりやすいものが多かったので僕が全てまとめ直すことはしていません

あと，用語に関しては調べれる範囲で調べましたが訳し方が間違っていたらすみません…

経路選択問題とは何か

経路選択問題とは

経路選択問題とは，道路ネットワークの中でODが与えられた時に「どのようにリンクを辿るか」という問題である．DijkstraやA*アルゴリズムといった，最短経路探索問題とは異なる．



道路ネットワークの例

1. 経路選択問題のアプローチ

下に示したような効用関数を経路ごとに与え，簡単に言えば，「交通手段選択に関してMNLモデルを適用する」感覚で，ODに対してどの経路が選択されるのかを考える．

$$V_i = \beta_L L_i + \beta_{SB} SB_i$$

β_L：経路の長さ（$=L_i$）に関するパラメータ
β_{SB}：スピードバンプが経路にいくつ含まれているか（$=SB_i$）に関するパラメータ

2. 最短経路探索問題のアプローチ

左のような道路ネットワークの「リンク」一つ一つに”リンクコスト”を想定し，ODを繋ぐ経路で総リンクコストが最小になる「リンクの組み合わせ」を考える．

経路選択問題の特徴

経路選択問題とは

経路選択問題を考えるにあたり，共通のリンクを有する経路間での誤差項の相関に着目する必要がある．

経路選択問題における誤差項の相関

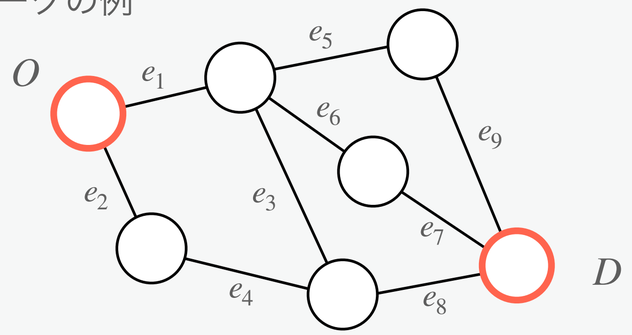
経路における共通部分をどのように考えるか?
従来はPass-Size-Logit（PSL）やLogitといったモデルで，誤差項の相関を考慮する項を組み入れることで対処していた．（ただし経験則に過ぎず，また選択肢に対して非常に敏感で扱いが難しかった．）

一方，Cross-Nested-Logit（CNL）モデルは，**リンクごとにネストを作り，リンクを有する経路を対応するネストに紐づける**ようなモデル
　→根幹から誤差項の相関を捉えることが可能！

一方，Cross-Nested-Logit（CNL）モデルは，**パラメータの数が多く推定が難しい．**
　→**ネットワークの形状から，経験則的に値を設定してしまうことも…**

近年だと経路の列挙を必要としないRLモデルも（今回は扱わない）

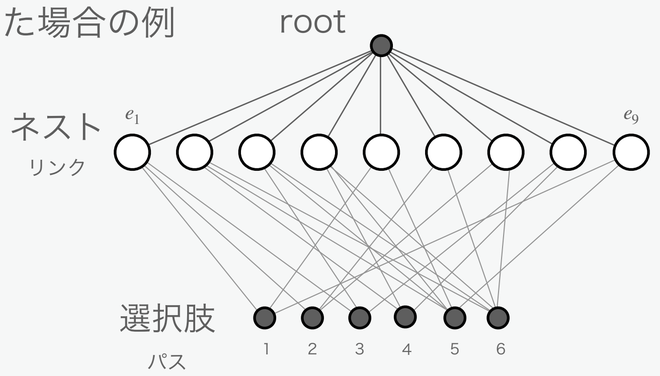
ネットワークの例



経路 { $e_1e_5e_9$, $e_1e_6e_7$, $e_1e_3e_8$, $e_2e_4e_8$, $e_2e_4e_3e_5e_9$, $e_2e_4e_3e_6e_7$ }

左からパス1，2，3，4，5，6とする．

CNLにした場合の例



サンプリングの必要性

経路選択問題とは

実道路ネットワークの場合膨大になる経路をどのようにコンパクトにするか？

①どれくらい経路というものは多いのか？

例えばノードが80あったとして，密度が0.4とした時，リンクの数は
$(80 \times (80-1)/2) \times 0.4$であり，経路は$10^{84}$程度と膨大になってしまう．
（ループ経路を除いても）

↓

サンプリングを行う必要性！

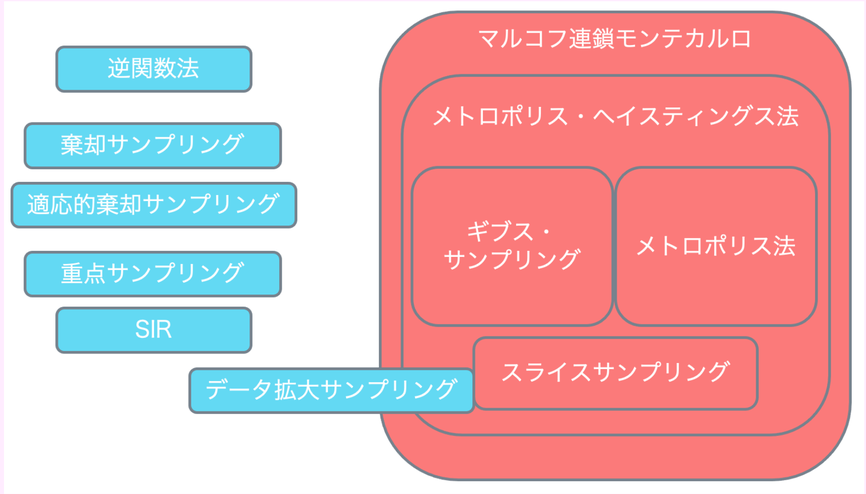
②サンプリングの際に満たすべき条件

1. 選ばれた選択肢がサンプリングされた選択肢群に含まれている
2. 被サンプリング確率が算出可能（サンプリング補正のため必要）

↕

ただ，そのサンプリングの元となる分布は何か？
その分布から正しくサンプリングができるのだろうか？
（実際問題サンプリングの使い方はとても難しい）

③サンプリングとはいっても色々ある



PRML第11章　原さんの資料より

今回扱う手法

サンプリング手法：M-H (Metroporis-Hasting) アルゴリズム

マルコフ連鎖モルテカルロ法の一種.
それぞれの経路の被サンプル確率を事前に決定し，サンプリングしていく．確立質量関数が必要ないのも良い.

GEVモデル①
CNLモデルとデータサンプリングの手法

まずは，GEVモデルとは何か，基本的な周辺知識を確認する.

①GEVモデルとは

- G関数から全て導出が可能.
- MNLのIIA特性を緩和し，選択肢間の誤差相関を柔軟に表すことが可能．全ての誤差相関には対応しないので，その都度G関数を書き換える必要がある.
- 選択確率がクローズドフォームで（積分系が残らない）計算負荷が軽い.

②G関数とは

G関数は4つの性質を満たす.

- $G \geqq 0 \ for\ all\ positive\ values\ of\ y_i\ \forall j$
- G関数は以下の式を満たす

$$G(\rho y_1, \rho y_2, \cdots, \rho y_J) = \rho^{\mu} G(y_1, y_2, \cdots, y_J)$$
$$for\ \rho > 0$$

- $G \to \infty\ as\ y_i \to \infty\ for\ any\ j$
- G関数のk次導関数を考えた時，kが奇数なら導関数は正，偶数なら負.

GEVモデルでは，各モデルの選択確率P_iをG関数から導出することが可能.
効用確定項をV_iとすると，$y_i = \exp(V_i)$として

$$P_i = \frac{y_i}{\mu G}\frac{\partial G}{\partial y_i}$$

2016スタートアップゼミ#3 前田さんの資料より

③例えばMNLならば…

一般的なMNLのG関数 $G = \sum_{j=1}^{J} {y_j}^{\mu}$

① $G = \sum_{j=1}^{J} {y_j}^{\mu} = \sum_{j=1}^{J} exp(\mu V_j) \geq 0$

② $G(\rho y_j) = \sum_{j=1}^{J} \rho^{\mu} {y_j}^{\mu} = \rho^{\mu} \sum_{j=1}^{J} Y_j = \rho^{\mu} G$

③ $G \to \infty\ as\ y_j \to \infty\ for\ any\ j$

④ $\frac{\partial G}{\partial y_i} = \frac{\partial}{\partial y_i}\left({y_1}^{\mu} + {y_2}^{\mu} + \cdots + {y_i}^{\mu} + \cdots + {y_J}^{\mu}\right)$
$= \mu {y_i}^{\mu-1} \geq 0$

$\frac{\partial^2 G}{\partial y_i y_j} = 0 \leq 0 \quad \frac{\partial^2 G}{\partial y_i \partial y_i} = \mu(\mu-1){y_i}^{\mu-2} \leq 0$

$G = \sum_{j=1}^{J} Y_j$はG関数

一般的なMNLのP_i導出

$$P_i = \frac{y_i}{\mu G}\frac{\partial G}{\partial y_i} = \frac{y_i}{\mu \sum_{j=1}^{J} {y_j}^{\mu}} \mu {y_i}^{\mu-1}$$
$$= \frac{{y_j}^{\mu}}{\sum_{j=1}^{J} {y_j}^{\mu}} = \frac{\exp(\mu V_i)}{\sum_{j=1}^{J} \exp(\mu V_j)}$$

GEVモデル②
CNLモデルとデータサンプリングの手法

続いてCNLに関して考える.

そもそも経路選択問題におけるCNLモデルとは？（確認）

- 各リンクはネストに紐づけられている
- 各経路は，経路の中でリンクが占める割合に関連した強度で各ネストに紐づけられる.

→経路間での誤差項の相関をネストへの結合で表現.

G関数が満たすべき性質（確認）

- $G \geqq 0 \ for\ all\ positive\ values\ of\ y_i\ \forall j$
- G関数は以下の式を満たす

$$G(\rho y_1, \rho y_2, \cdots, \rho y_J) = \rho^{\mu} G(y_1, y_2, \cdots, y_J)$$
$$for\ \rho > 0$$

- $G \to \infty\ as\ y_i \to \infty\ for\ any\ j$
- G関数のk次導関数を考えた時，kが奇数なら導関数は正，偶数なら負.

G関数と選択確率

GEVモデルでは，各モデルの選択確率P_iをG関数から導出することが可能.
効用確定項をV_iとすると，$y_i = \exp(V_i)$として

$$P_i = \frac{y_i}{\mu G}\frac{\partial G}{\partial y_i}$$

α_{jl}：アロケーションパラメータ（経路 j のネスト l に対する結合度）
μ：モデル全体のスケールパラメータ（1に標準化されることが多い）
μ_l：ネスト l 内のスケールパラメータ
　ネスト l 内の選択肢間の誤差相関の強さ.

CNLモデル

$$y_i = \exp(V_i)$$

$$G = \sum_{l=1}^{k} \left(\sum_{j \in B_l} (\alpha_{jl}^{\frac{1}{\mu}}) \right)^{\frac{\mu}{\mu l}}$$

ネスト内で足し合わせた上で，経路が含まれるネスト全てで和をとる

G関数が満たすべき性質の4つ目に関しては，補足に入れた前田さんの資料を参考に

$$\frac{\partial G}{\partial y_i} = \mu \sum_{l=1}^{k} \{ \alpha_i^{\frac{\mu_l}{\mu}} y_i^{\mu_l - 1} (\sum_{j \in B_l} (\alpha_{jl}^{\frac{1}{\mu}} y_i)^{\mu_l})^{\frac{\mu}{\mu_l} - 1} \}$$

$$P_i = \frac{y_i G_i}{\mu G}$$

各経路の選択確率
複雑で見にくいので次スライド以降は論文での表記を使用

$$= \frac{\sum_{l=1}^{k} (\alpha_{il}^{\frac{1}{\mu}} y_i)^{\mu_l} (\sum_{j \in B_l} (\alpha_{jl}^{\frac{1}{\mu}} y_i)_l^{\mu})^{\frac{\mu}{\mu_l - 1}}}{\sum_{l=1}^{k} (\sum_{j \in B_l} (\alpha_{jl}^{\frac{1}{\mu}} y_i)^{\mu_l})^{\frac{\mu}{\mu_l}}}$$

CNLモデル（GEVモデル③）
CNLモデルとデータサンプリングの手法

論文中での基本的なCNLモデルの書き方と，図の中でそれぞれの項のイメージを確認.

ある個人が経路の選択肢集合 C から経路 i を選択する確率は,

$$G_i = \sum_{m=1}^{M} \left[\mu \alpha_{im} e^{V_i(\mu_m - 1)} \left(\sum_{j \in C} \alpha_{jm} e^{\mu_m V_j} \right)^{\frac{\mu - \mu_m}{\mu_m}} \right]$$

のもとで（G_iはCNLモデルのG関数を$\exp V_i$で偏微分したもの）

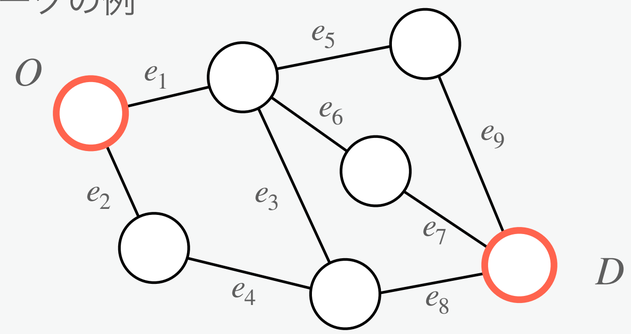
$$\Pr(i \mid C) = \frac{\exp(V_i + \ln G_i(C))}{\sum_{j \in C} \exp(V_j + \ln G_j(C))}$$

で表すことができる.

- V_i：経路 i に関する効用の確定項
- α_{im}：アロケーションパラメータ（経路 i のネスト m に対する結合度）
- μ：モデル全体のスケールパラメータ（0以上で1に標準化されることが多い）
- μ_m：ネスト m 内のスケールパラメータ

（ネスト l 内の選択肢間の誤差相関の強さ．0から1で，1だと相関なし）

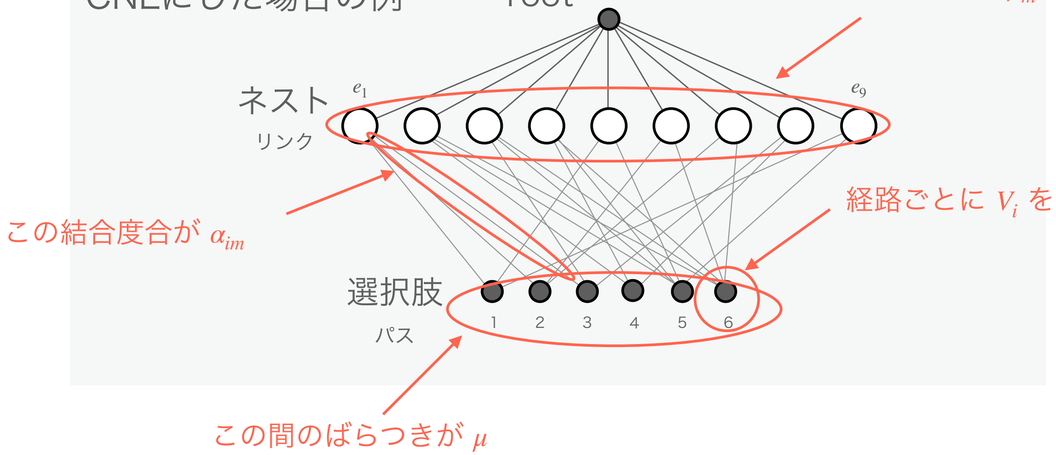
ネットワークの例



経路 { $e_1e_5e_9$, $e_1e_6e_7$, $e_1e_3e_8$, $e_2e_4e_8$, $e_2e_4e_3e_5e_9$, $e_2e_4e_3e_6e_7$ }

左からパス1，2，3，4，5，6とする.

CNLにした場合の例



サンプリング手法①
CNLモデルとデータサンプリングの手法

なぜサンプリングを行うのか，MCMC以外のサンプリング手法（色々あるが今回は本筋ではないので一例）

なぜ経路選択問題でサンプリングは必要か？（確認）

例えばノードが80あったとして，密度が0.4とした時，リンクの数は
$(80 \times (80-1)/2) \times 0.4$であり，経路は$10^{84}$程度と膨大になってしまう．
（ループ経路を除いても）

↓

サンプリングを行う必要性！

サンプリングの際に満たすべき条件（確認）

1. 選ばれた選択肢がサンプリングされた選択肢群に含まれている
2. 被サンプリング確率が算出可能（サンプリング補正のため必要）

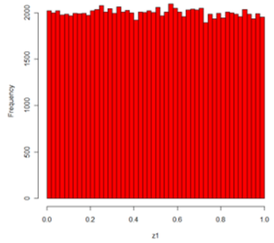
サンプリングとは，「想定した分布」に従う乱数を発生させ，
その分布から抽出を行うこと！

（例）逆関数法　グラフはPRML第11章　原さんの資料より

一様分布からサンプリングすることが可能であることが前提．
求めたい分布の確率密度関数の逆関数が簡単に表記できる場合に使う．
下では例として，指数分布からサンプリングしている．

z → $h^{-1}(z)$ → $p(y)$

一様乱数 → 求めたい分布の不定積分の逆関数 → 求めたい分布からサンプリングした乱数



[0, 1]の一様乱数zを
100000個生成

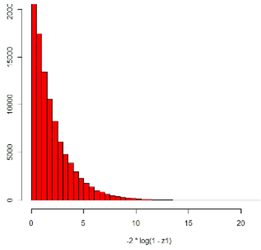
$$p(y) = \lambda \exp(-\lambda y)$$

$$z = h(y) \equiv \int_{-\infty}^{y} p(y')dy'$$

$$z = h(y) = 1 - \exp(-\lambda y)$$

$$h^{-1}(z) = -\lambda^{-1} \ln(1-z)$$

発生させた一様乱数zを
$h^{-1}(z) = -2\ln(1-z)$に代入



指数分布が得られた

- 解析的に正しく，無駄がない．
- 複雑な分布だと逆関数を得ることはほぼ不可能．

分布によって本当にさまざまなサンプリング手法がある．参考資料の資料を参考に．

サンプリング手法②
CNLモデルとデータサンプリングの手法

MCMCの有用性と，その概論

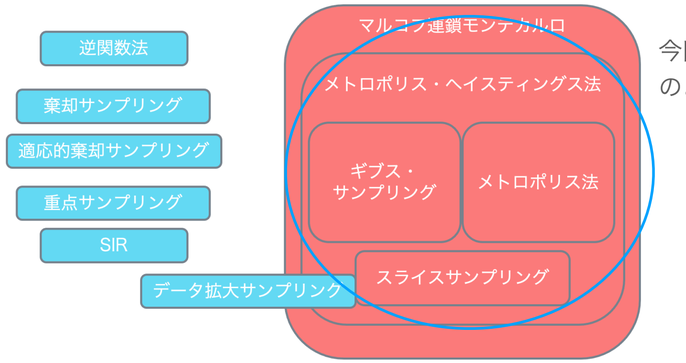
MCMC（マルコフ連鎖モンテカルロ法）の有用性

非MCMCのサンプリング手法では，分布が複雑になったとき直接事後分布からサンプリングするのが難しいので，もっと単純な分布（これを**提案分布**という）**をうまく用意する必要がある**.

ただし，次元が多かったり，分布が複雑なら適切に提案分布を用意することが難しい．（時に不可能）

MCMCなら**定常分布に収束する**という性質を持ち，**初期値に依存しない**
→いい感じのサンプリングが楽に行える！！！

MCMCの概観　PRML第11章　原さんの資料より

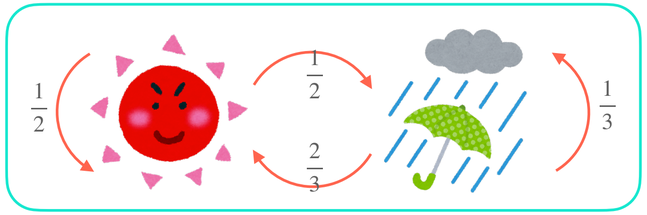


今回扱うM-H法はMCMCの標準系のようなもの

MCMCの基本メカニズム

任意の分布$p(z)$**を不変分布に持つマルコフ連鎖を生成することで**，望む分布に従った乱数を生成するモンテカルロ法のこと.

不変分布とは？



$$\begin{pmatrix} F_{t+1} \\ R_{t+1} \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{2}{3} \\ \frac{1}{2} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} F_t \\ R_t \end{pmatrix}$$

右上の漸化式を解くと

$$\begin{pmatrix} F_t \\ R_t \end{pmatrix} = \begin{pmatrix} \frac{3}{7}(-\frac{1}{6})^t + \frac{4}{7} \\ -\frac{3}{7}(-\frac{1}{6})^t + \frac{3}{7} \end{pmatrix}$$

右上の漸化式を解くと

$$\begin{pmatrix} F_\infty \\ R_\infty \end{pmatrix} = \begin{pmatrix} \frac{4}{7} \\ \frac{3}{7} \end{pmatrix}$$

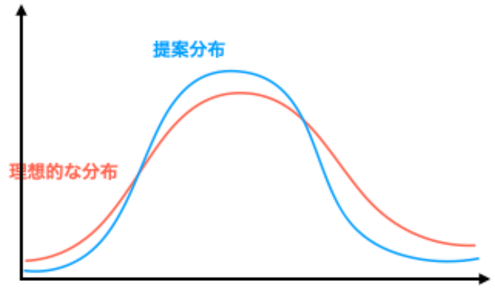
不変分布

マルコフ連鎖では，**初期値によらず最終的にある一定の確率分布に収束**する
→この性質を生かす！

望む分布に向かう推移確率が与えられれば簡単.
例では「推移確率→不変分布」と向かったが，**MCMCは不変分布に到達するような推移確率を与えることで，分布を再現する**.

じゃあ，推移確率をどのようにして与えるか？
　→メトロポリス法，ギブスサンプリング，M-H法

サンプリング手法③
CNLモデルとデータサンプリングの手法



MCMCでの推移確率決定の指針

MCMCでの推移確率決定の基本指針

詳細釣り合いの条件を満たすように推移確率を決めてあげれば良い！

前のスライドの例ならば

$$\frac{4}{7}\times\Pi_{Fine\to Rain}=\frac{3}{7}\times\Pi_{Rain\to Fine}$$

青が不変分布で赤が推移確率
均衡状態のようなイメージ

上の式を満たす推移確率が望む不変分布を与えるなら，
（推移確率）×（不変分布）＝（不変分布）が成立するはず！

$$\begin{pmatrix}1-\Pi_{Fine\to Rain} & \Pi_{Rain\to Fine}\\ Pi_{Fine\to Rain} & 1-\Pi_{Rain\to Fine}\end{pmatrix}\begin{pmatrix}\frac{4}{7}\\ \frac{3}{7}\end{pmatrix}=\begin{pmatrix}\frac{4}{7}-\frac{4}{7}\Pi_{Fine\to Rain}+\frac{3}{7}\Pi_{Rain\to Fine}\\ \frac{4}{7}\Pi_{Fine\to Rain}+\frac{3}{7}-\frac{3}{7}\Pi_{Rain\to Fine}\end{pmatrix}=\begin{pmatrix}\frac{4}{7}\\ \frac{3}{7}\end{pmatrix}$$

より一般的な証明をするなら

詳細釣り合いの条件は

$$Pr\{x\}\times\Pi_{x\to x'}=Pr\{x'\}\times\Pi_{x'\to x}$$

（不変確率）＝（推移確率）×（不変分布）

$$Pr\{x\}=\sum_{x'}\Pi_{x'\to x}\times Pr\{x'\}$$ 推移モデルの一般式

$$=\sum_{x'}\Pi_{x\to x'}\times Pr\{x\}$$ 詳細釣り合いの条件

$$=Pr\{x\}\sum_{x'}\Pi_{x\to x'}=Pr\{x\}$$ 推移確率の和=1

証明できた！
分布では，一点とその他の点での均衡に詳細釣り合い条件が成立します！

メトロポリス・ヘイスティング法と棄却サンプリング

MCMC系を理解するためにはまず棄却サンプリングを知る必要がある！

目標分布：既知，遷移確率：未知
であるので，いきなり遷移確率を得ることは難しい．
→遷移確率の代わりに**適当な遷移確率を”提案分布”として利用**する！

提案分布とは？

これが棄却サンプリングの基本の考え方

目標分布に従ったものではないが，乱数生成が容易な条件付き確率分布から選ぶもの．提案分布から生成された乱数を棄却したり受け入れたりし，目標分布に近い乱数の生成を目指す．

提案分布は基本的には左で示した詳細釣り合い条件を満たさない…
→確率補正を行う！！

確率補正とは？

右の例では下のようになり均衡していないが

$$Pr\{x\}\times\Pi_{x\to x'}<Pr\{x'\}\times\Pi_{x'\to x}$$

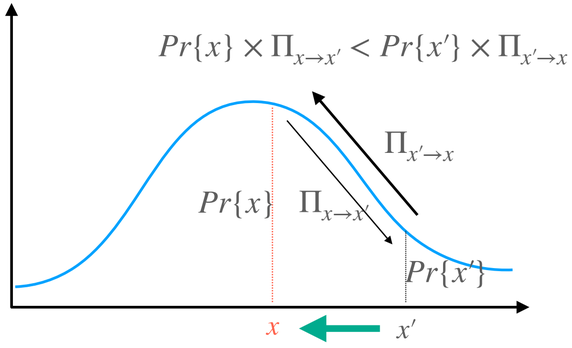
都合よくcとc'符号が正の未知数を与えることで

$$c\ Pr\{x\}\times\Pi_{x\to x'}=c'\ Pr\{x'\}\times\Pi_{x'\to x}$$

とすることができる．

しかしここで二つの問題
- **遷移確率は1以下**
- **2変数で無駄が多い**

$$Pr\{x\}\times\Pi_{x\to x'}<Pr\{x'\}\times\Pi_{x'\to x}$$



$x'\to x$の遷移が卓越

提案分布で詳細釣り合い条件を満たしていない例

サンプリング手法④

CNLモデルとデータサンプリングの手法

メトロポリス・ヘイスティング法の考え方，詳細

$c\ Pr\{x\} \times \Pi_{x \to x'} = c'\ Pr\{x'\} \times \Pi_{x' \to x}$

先スライドで出てきた右の問題を

両辺を c' で割り，$\frac{c}{c'}$で方程式を解く！

$r = \frac{c}{c'}$，　$r' = \frac{c'}{c'} = 1$として考えると，

$r\ Pr\{x\} \times \Pi_{x \to x'} = r'\ Pr\{x'\} \times \Pi_{x' \to x}$　詳細釣り合い条件を満たす推移確率ができた！

しかしここで二つの問題
- **cもc'も1以下である必要**
- **2変数で無駄が多い**

まとめると，M-H法は提案された候補点 a を，確率 min(1, r) で受容し，さもなくばその場に留まることを繰り返すアルゴリズム！

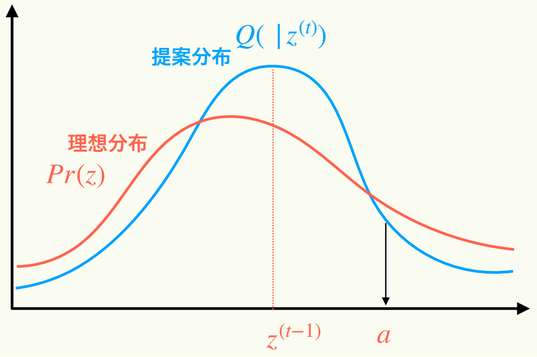
M-H法のアルゴリズム

Step1

まず初期値 $z^{(0)}$ を決め，t=1とする.

Step2

提案分布 $Q(\ |z^{(t-1)})$ を利用し乱数 a を得る.



Step3

以下の命題を判定する.

$Q(a|z^{(t-1)})\ Pr(z^{(t-1)}) > Q(z^{(t-1)}|a)\ Pr(a)$

命題が真ならば，遷移先より遷移前の方が事後確率上確率密度が高い！

→補正が必要！

命題が偽ならば，遷移前より遷移後の方が事後確率上確率密度が高い！

→補正は不必要！（遷移を無条件に受容）

遷移が受容された場合，　$z^{(t)} = a$

遷移が棄却された場合，　$z^{(t)} = z^{(t-1)}$

命題が真の場合の補正方法

$$r = \frac{Q(z^{(t-1)}|a)\ Pr(a)}{Q(a|z^{(t-1)})\ Pr(z^{(t-1)})}$$

$Q(z^{(t-1)}|a)\ Pr(a)$：$a \to z^{(t-1)}$ の条件付き確率密度

$Q(a|z^{(t-1)})\ Pr(z^{(t-1)})$：$z^{(t-1)} \to a$ の条件付き確率密度

rは0から1であるため，[0, 1]の一様乱数を生成し，rと比べる.
- 乱数 ＞ r ならば遷移を受容
- 乱数 ＜ r ならば遷移を棄却

Step4

t=t+1として，ステップ2に戻る.

CNLモデルとM-Hサンプリングの融合・バリエーション①

CNLモデルとデータサンプリングの手法

論文のモデルの解説

CNLの基本式は以下の式で与えられる.

$$\Pr(i \mid C) = \frac{\exp(V_i + \ln G_i(C))}{\sum_{j \in C} \exp(V_j + \ln G_j(C))}$$

$$G_i = \sum_{m=1}^{M} \left[\mu \alpha_{im} e^{V_i(\mu_m - 1)} \left(\sum_{j \in C} \alpha_{jm} e^{\mu_m V_j} \right)^{\frac{\mu - \mu_m}{\mu_m}} \right]$$

ここで，経路の選択肢集合を C から D に限定する.

ただし，$\Pr(j \mid D) > 0, \forall j \in D$

この場合，CNLモデルに**サンプリングを行なったことによって生じた誤差相関を加味する項を加える**．結果は以下の通り.

$$\Pr(i \mid D) = \frac{\exp(V_i + \ln G_i(C)) + \ln \Pr(D \mid i)}{\sum_{j \in D} \exp(V_j + \ln G_j(C)) + \ln \Pr(D \mid j)}$$

ここで，選択肢集合が限定されたことにより G_i や $\Pr(i \mid D)$ を再計算する.

この論文内ではGuevara and Ben-Akiba（2013）の方法を採用.

論文内では，M-Hサンプリングで

- $\Pr(i \mid D)$の分母部分の計算で考える選択肢集合 D
- G_i を近似して計算するための選択肢集合 D'

の**二つの選択肢群をサンプリングする必要がある**ことが指摘されている.
ただし，D のみは選択確率を計算したい経路を必ず含む必要がある.

サンプリング手法でM-Hサンプリングを採用している理由は「**パスのサンプリング確率をこちらで決めることができる**から」．パス i のサンプリング確率 $q(i)$ の算出方法は以下の式で与える.

$$q(i) = \frac{b_i}{B \sum_{j \in C} b(i) = b(j)}$$

- b_i：経路の重み
  （例えば経路 i の長さ L_i を利用して $\exp(-\theta L_i)$など）
- B：経路の重みの和（選択肢集合 C 全て）
- $\sum_{j \in C} b(i) = b(j)$：同じ重みの経路が何本あるか

ここで B が計算できなくても良いのは，
M-Hでどうせ比を取り，B を約分できるから

B，$|b|$の計算方法（近似）は以下の通り．$|C|$の計算は補足資料に.

$$B = \sum_{j \in C} b(j) = |C| \bar{b}, \quad |b| = \frac{\sum_{i \in C} b(i)}{|C|} \approx \frac{\sum_{i \in D} b(i)}{|D|}$$

経路の重みの平均

CNLモデルとM-Hサンプリングの融合・バリエーション②

CNLモデルとデータサンプリングの手法

論文のモデルの解説

サンプリングは重複ありで行われる．パスiの被サンプル回数をk_iとして

$$\Pr(i\,|\,D) = \frac{\exp(V_i + \ln G_i(C)) + \ln \Pr(D\,|\,i))}{\sum_{j\in D} \exp(V_j + \ln G_j(C)) + \ln \Pr(D\,|\,j))} \quad \cdots ①$$

の中の$\Pr(D\,|\,i)$を

$$\Pr(D\,|\,i) = K_D \frac{k_i}{q(i)} = K_D B \frac{k_i}{b(i)}$$

とかける．ただしK_Dはiから独立な係数．これを①に代入すると

$$\Pr(i\,|\,D) = \frac{\exp(V_i + \ln G_i(C)) + \ln \frac{k_i}{b(i)})}{\sum_{j\in D} \exp(V_j + \ln G_i(C)) + \ln \frac{k_j}{b(j)})} \quad \cdots ②$$

続いて，$G_i(C)$ も選択肢群 D から計算したもので近似する．前スライドと同じ論文で提案された方法を用いる．

$$G_i(C) \approx G_i(\hat{D'},\, w) = \sum_{m=1}^{M} \left[\mu \alpha_{im} e^{V_i(\mu_m - 1)} \left(\sum_{j\in D'} w_j \alpha_{jm} e^{\mu_m V_j} \right)^{\frac{\mu-\mu_m}{\mu_m}} \right] \quad \cdots ③$$

元々の式との違いは w_j が入っていることで，これは**「ネストの中でサンプルされなかった経路」を補う拡大係数**である．

この研究ではこの**拡大係数をどのように表現するかに関してバリエーション**を持たせ，適切な手法を探している．

拡大係数のバリエーション

(1)Guevara and Ben-Akiba (2013)

$$w_j^G = \frac{k_j}{E[k_j]} = \frac{k_j}{q(j)R} = \frac{k_j B}{b(j)R}$$

R は D' を生成する際にM-H法を回したステップの回数．Bに関しては，前のスライドで示した通り $B = \bar{b}\,|C| \approx \frac{\sum_{i\in D} b(i)}{|D|}|C|$ と表すことができる．

(2)Frejinger et al. (2009)

$$w_j^F = \begin{cases} \frac{B}{b(j)R} & if\ b(j)R > B \\ 1 & otherwise \end{cases}$$

Pass Size Logitモデルと同じような形で近似した拡大係数の置き方．各値は上のものと同じ．

(3)Guevara and Ben-Akiba (2013)

$$w_j^L = \frac{k_j}{q(j)R} \approx \frac{k_j}{q(j)R} \frac{q(s)R}{k_s} = \frac{k_j}{k_s} \frac{b(s)}{b(j)}$$

大数の法則が成り立てば=1

経路の重みの和 B を計算しなくて良い．選択肢群 D を生成する際に最もサンプルされた数が多い経路を s としている．
$k_s \approx q(j)R$ は詰まるところ大数の法則を用いた近似であるのでサンプル数が多い時に使う．

拡大係数の意味とは？

実情としては，①の式に③を代入し，拡大係数をG関数に組み込まなくても推定料の一貫性は保たれる．拡大係数の意味とは**サンプル数が少なかった時に推定量を補正するもの**．
特に(1)の拡大係数は③式内の $\sum_{j\in D'} w_j \alpha_{jm} \exp(\mu V_j)$の項の普遍性を保障している．

$$\Pr(i\,|\,D,\, D',\, w) = \frac{\exp(V_i + \ln G_i(\hat{D'},\, w) + \ln \frac{k_i}{b(i)})}{\sum_{j\in D} \exp(V_j + \ln G_i(\hat{D'},\, w) + \ln \frac{k_i}{b(i)})}$$

モデル検証のフロー

モデルの検証

合成データ，実データを用いてモデルの検証を行っているが，そのフローを確認する．

合成データ

合成データで検証を行う目的は

- サンプリングしたデータの特徴
- 拡大係数
- M-Hアルゴリズムの構成

に関して，大枠のイメージを掴むこと．

まずは相応しいサンプリングを目指し，その後より良いモデルを目指す

実データ

まずは簡単なLogitとかで大体のパラメータを調べておく．
調べたパラメータをもとに，経路サンプリング確率における重みづけをいろいろ試す
場合に応じて使用可能な拡大係数を使う
実データに対して正しいパラメータと比較することは無理だからデータを二つに分けておいて検定

実データを使ったモデルの検証①

モデルの検証

合成データを用いてモデルの精度の検証をする．その設定を確認する．

設定

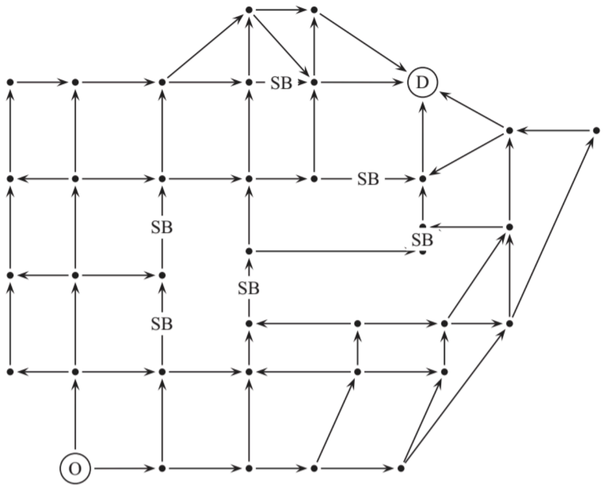


Fig. 1. Network with 170 paths from O to D.

道路ネットワークはスウェーデンのボルレンゲのネットワーク形状を参考に簡略化したもの．図の中のODを結ぶ経路は高々170．図の中にあるようにSpeed Bump（SBと書く）がある．経路 i の長さを L_i ，経路上のSBの数をSB_iとして

$$V_i = \beta_L L_i + \beta_{SB} SB_i$$

と経路の効用の確定項を表す．サンプリングなしのCNLモデルの推定結果からわかっている正しいパラメータ値$\beta_L = -0.5, \beta_{SB} = -0.1$と比較を行う．モデルはCNLで，それぞれのリンクを一つずつネストとする．

G関数のパラメータの設定は以下の通り．

- $\alpha_{im} = l_m / L_i$
- $\mu_m = 1.5$（β_Lなどと一緒にサンプリングなしのCNLで得た）
- $\mu = 1$

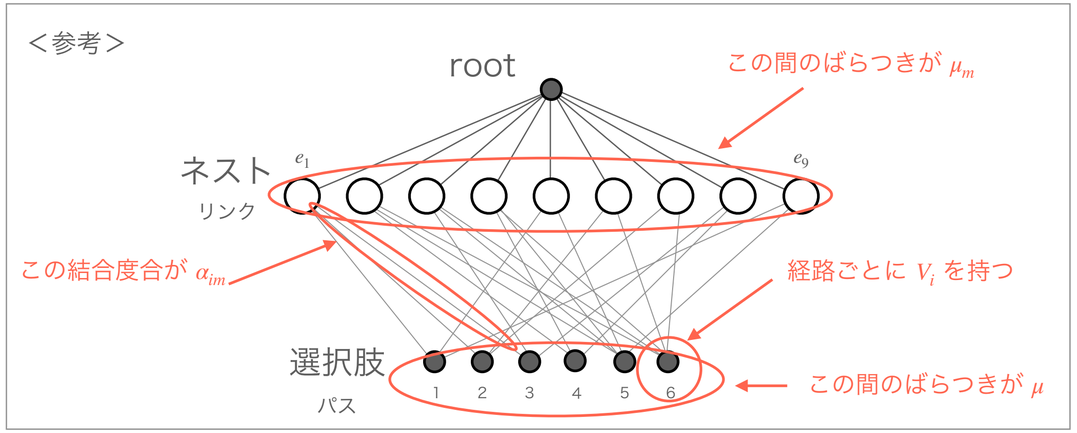


Table 1
Estimations for the CNL model with the full choice set and synthetic data.

Parameters	Estimated value	Standard error	t-Test (0)	t-Test (true value)
β_L	−0.501	0.0118	43.1	0.678
β_{SB}	−0.0910	0.0240	3.19	0.375
μ_m	1.49	0.0269	55.2	0.0535

合成データを使ったモデルの検証②

モデルの検証

実際の値がわかっている合成データを用いてモデルの精度の検証をする．その設定を確認する．

設定

使用するCNLモデルは，

$$\Pr(i\,|\,D) = \frac{\exp(V_i + \ln G_i(C)) + \ln \frac{k_i}{b(i)})}{\sum_{j\in D} \exp(V_j + \ln G_i(C)) + \ln \frac{k_j}{b(j)})} \quad \cdots ①$$

$$\Pr(i\,|\,D) = \frac{\exp(V_i + \ln G_i(\hat{D}',\, w) + \ln \frac{k_i}{b(i)})}{\sum_{j\in D} \exp(V_j + \ln G_i(\hat{D}',\, w) + \ln \frac{k_i}{b(i)})} \quad \cdots ②$$

$$\Pr(i\,|\,D) = \frac{\exp(V_i + \ln G_i(\hat{D}',\, 1))}{\sum_{j\in D} \exp(V_j + \ln G_i(\hat{D}',\, 1))} \quad \cdots ③$$

の三つである．③は拡大係数，サンプリングを補正する項の$\ln \frac{k_i}{b(i)}$を省略している．

それぞれ

① D のみサンプリングすれば良い．MHの構成による変化が見やすい．

② 拡大係数と推定結果の関係性．D' の果たしている役割を確認する．

③ ②のモデルとの比較用．

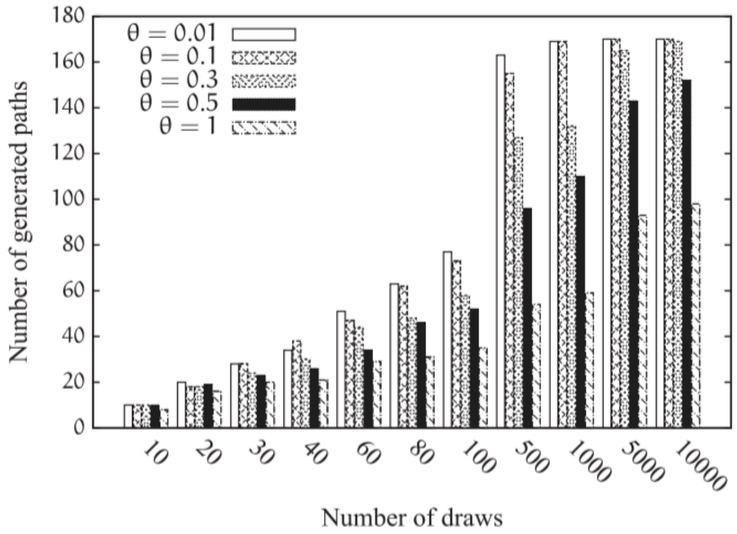
といった役割を持たせている．

M-Hサンプリング

M-Hサンプリングでの経路 i のサンプリング確率は，経路 i の経路長を L_i として

$$q(i) = \frac{b_i}{B\sum_{j\in C} b(i) = b(j)},\quad b(i) = \exp(-\theta L_i) \qquad ※\theta \text{ は正の数}$$

とした．短い経路ほど選択されやすいという考え方．$\theta \approx 0$ ならば均等にサンプリングすることになり，θ が大きければ偏ってサンプリングすることになる．



全経路は170（上限）

ig. 2. Size of the choice set generated by the MH algorithms for various values of θ.

サンプリングした回数（x軸）と，サンプリングされた経路の種類（y軸）

合成データを使ったモデルの検証③

モデルの検証

実際の値がわかっている合成データを用いてモデルの精度の検証をする.

モデル①の検証

M-Hアルゴリズムで，θ を0.5にした場合と0.01にした場合，サンプリングする選択肢を10にする場合と40にする場合の，**2×2通りに対して100回実験**を行った（サンプリングとパラメータ推定をそれぞれ100回）モデルは下の通り.

$$\Pr(i \mid D) = \frac{\exp(V_i + \ln G_i(C)) + \ln \frac{k_i}{b(i)})}{\sum_{j \in D} \exp(V_j + \ln G_i(C)) + \ln \frac{k_j}{b(j)})}$$

$\theta = 0.5$の場合

常に良いt値を得た．サンプリング回数が40でもサンプリングされた経路の数は25程度であった.

$\theta = 0.01$の場合

サンプリングされた経路の数は $\theta = 0.5$ の場合より大きく増えたものの，パラメータ値に対して良いt値は得られなかった.

つまり，**多様な経路がサンプリングされるべきだが，サンプリング回数が少ないのならば重み付けを強めに行わなければならない**.

サンプリングなしで求めた真の値

75%信頼区間に入っていた数

Sampling protocol	Draws	Param.	True	Mean	Stdev	t-Test (0)	t-Test true	Min.	Max.	LowBound	UpBound	Count
$\theta = 0.5$	10	β_L	−0.5	−0.491	0.0127	38.6	0.669	−0.513	−0.471	−0.516	−0.483	76
		β_{SB}	−0.1	−0.0771	0.0263	2.93	0.867	−0.0994	−0.0564	−0.133	−0.0665	84
		μ_m	1.5	1.51	0.0275	54.9	0.308	1.42	1.58	1.46	1.53	70
	40	β_L	−0.5	−0.497	0.0136	36.5	0.217	−0.506	−0.475	−0.517	−0.482	88
		β_{SB}	−0.1	−0.0779	0.0312	2.49	0.706	−0.0865	−0.0681	−0.139	−0.0603	100
		μ_m	1.5	1.49	0.0288	51.7	0.302	1.47	1.53	1.46	1.53	99
$\theta = 0.01$	10	β_L	−0.5	−0.531	0.0141	37.6	2.21	−0.550	−0.512	−0.517	−0.482	2
		β_{SB}	−0.1	−0.129	0.0370	3.48	0.790	−0.159	−0.101	−0.146	−0.0530	92
		μ_m	1.5	1.41	0.0290	48.6	2.80	1.34	1.47	1.46	1.53	1
	40	β_L	−0.5	−0.536	0.0130	41.2	2.78	−0.548	−0.523	−0.516	−0.483	0
		β_{SB}	−0.1	−0.130	0.0293	4.43	1.05	−0.147	−0.115	−0.137	−0.0627	67
		μ_m	1.5	1.39	0.0262	53.0	3.99	1.35	1.44	1.46	1.53	0

モデル①のシミュレーション結果

合成データを使ったモデルの検証④

モデルの検証

モデル②，③の精度の検証をする．

モデルの説明 -モデル②，③

検証するのは「サンプリング補正」「拡大係数」それぞれの効果と「D に加え D' をモデルに組み込んだ意味」である．これらを検証するために以下の条件を変えている．

- D' のサンプリング回数：100，200，300（ D は100）
- D' のM-Hサンプリングの重み付けにおけるパラメータθ：0.5，0.01（ D は0.5）
- 拡大係数：モデル②に関して下の4つ．（モデル③は拡大係数がない）

$$w_j^G = \frac{k_j}{E[k_j]} = \frac{k_j}{q(j)R} = \frac{k_j B}{b(j)R}$$

$$w_j^F = \begin{cases} \frac{B}{b(j)R} & if\ b(j)R > B \\ 1 & otherwise \end{cases}$$

$$w_j^L = \frac{k_j}{q(j)R} \approx \frac{k_j}{q(j)R}\frac{q(s)R}{k_s} = \frac{k_j}{k_s}\frac{b(s)}{b(j)}$$

$$w = 1$$

モデルは下の通り．

$$\Pr(i\,|\,D,\,D',\,w) = \frac{\exp(V_i + \ln G_i(\hat{D'},\,w) + \ln\frac{k_i}{b(i)})}{\sum_{j\in D}\exp(V_j + \ln G_i(\hat{D'},\,w) + \ln\frac{k_i}{b(i)})} \cdots ②$$

$$\Pr(i\,|\,D,\,D') = \frac{\exp(V_i + \ln G_i(\hat{D'},\,1))}{\sum_{j\in D}\exp(V_j + \ln G_i(\hat{D'},\,1))} \cdots ③$$

まず，$\bar{b}$と$|C|$に関して推定を行う．（拡大係数を計算するため）

$\bar{b}$の推定は$|b| = \frac{\sum_{i\in C} b(i)}{|C|} \approx \frac{\sum_{i\in D} b(i)}{|D|}$から行う．（100回実験を実施）

$|C|$はあるOD間の全経路の数だが，その推定は以下のアルゴリズムに従う．

1. 現在ノードを c，パスの確率を l とし，それぞれ $c = O$，$l = 1$ と初期化．またノード $i \to j$ が移動できるかを0_1で表した行列を A，ノードの集合を v とする．
2. $A(\cdot\,|\,c) = 0$として，Oに戻ることがないようにする．
3. A を参照し，c と繋がっているノードの集合を $v'_c = \{k \in v\,|\,A(c,\,k) = \}$とする．
4. v'_c からランダムにノードを一つ選び，それを c' とする．
5. $l = l/|V'_c|$とする．
6. $c = D$ ならアルゴリズムを終了し，違うなら $c = c'$ としステップ2に戻る．
7. ステップ6が終了するまでの繰り返し回数を N として $|C| \approx \frac{1}{N}\sum_{t=1}^{N}\frac{1}{l(i)}$

推定の結果は以下の通り．（ほぼ完璧に推定できている）

Table 3
Normalization factor B: estimations and t-tests w.r.t. the true values.

	True value	Mean	Standard error	t-Test (true value)
$\bar{b}$	0.688	0.684	0.00231	1.62
$\|C\|$	170	169	2.52	0.0722

合成データを使ったモデルの検証⑤

モデルの検証

モデル②，③の精度の検証をする．

モデルの検証

それぞれのモデルに対して100回実験を行った．また選択肢集合に関しては D は $\theta = 0.5$で40サンプリング，D' に関しては$\theta = 0.5, 0.01$，サンプリング回数も100, 200，300としている．（ D' の方が経路選択で選ぶ選択肢群）

結論としては

- $\theta = 0.5$の場合では，サンプリング回数が100回の場合でも w^G と w^L は良い結果を出した．300まで増やすと w^F でも良い結果を出したが，モデル③と $w = 1$ の場合はうまく推定ができなかった．
- 推定時間は $\theta = 0.5$ の方がだいぶ速い．
- $\beta_L, \beta_{SB}, \mu_m$ に関しての推定では，w_L ではサンプリングが40回でも β_L, β_{SB} で十分な結果が得られ，μ_m もサンプリングが60回で良い結果を得た．w_G ではサンプリングが40回でも β_{SB} 以外の β_L, μ_m もではサンプリングが100回でようやく良い結果を得た．

総括すると，G_i の補正 / パラメータ推定の両面で，

拡大係数：w^L，D' に関して $\theta = 0.5$，サンプリング回数：100回

が最も効率的．

ただ，**これらは経験知であるのでさまざまなデータでさらに検証する必要**がある．

② $$\Pr(i|D, D', w) = \frac{\exp(V_i + \ln G_i(\hat{D'}, w) + \ln \frac{k_i}{b(i)}}{\sum_{j \in D} \exp(V_j + \ln G_i(\hat{D'}, w) + \ln \frac{k_i}{b(i)}}$$

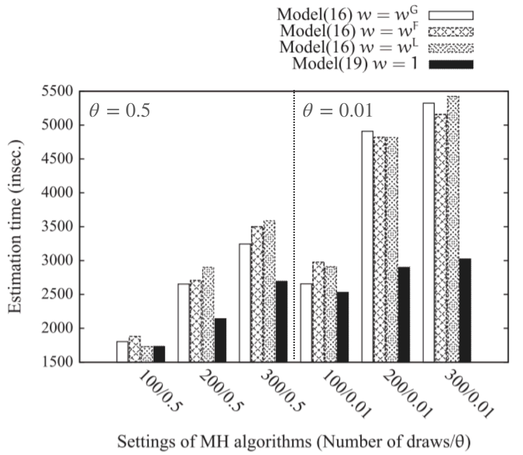
③ $$\Pr(i|D, D') = \frac{\exp(V_i + \ln G_i(\hat{D'}, 1)}{\sum_{j \in D} \exp(V_j + \ln G_i(\hat{D'}, 1)}$$

$$w_j^G = \frac{k_j}{E[k_j]} = \frac{k_j}{q(j)R} = \frac{k_j B}{b(j)R}$$

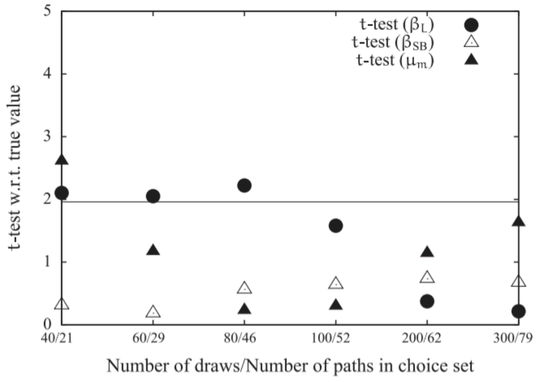
$$w_j^F = \begin{cases} \frac{B}{b(j)R} & if\ b(j)R > B \\ 1 & otherwise \end{cases}$$

$$w_j^L = \frac{k_j}{q(j)R} \approx \frac{k_j}{q(j)R} \frac{q(s)R}{k_s} = \frac{k_j}{k_s} \frac{b(s)}{b(j)}$$

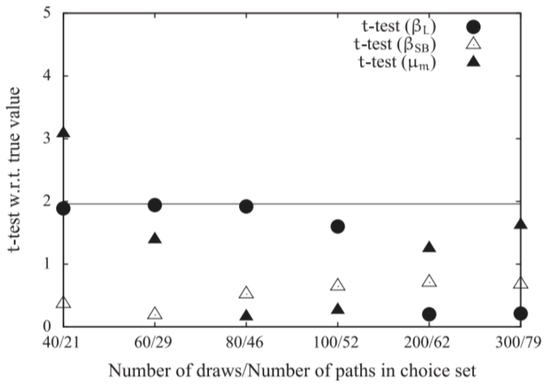
$w = 1$



$\theta = 0.5, \theta = 0.01$ で拡大係数ごとの推定時間



$\theta = 0.5, w = w^G$ での真値とのt検定



$\theta = 0.5, w = w^L$ での真値とのt検定

合成データを使ったモデルの検証⑥（補足）

モデルの検証

モデル②③に関してのパラメータ推定の結果.

Table 4
Estimation results for model (16) and (19).

Draws				100				200				300			
Model	Param.	True	True model[a]	Mean	Stdev	*t*-Test true	Count	Mean	Stdev	*t*-Test true	Count	Mean	Stdev	*t*-Test true	Count
$\theta = 0.5$															
Mod. (16) w^G	β_L	−0.5	−0.501	−0.480	0.0120	**1.58**	31	−0.495	0.0122	**0.374**	93	−0.502	0.0122	**0.214**	92
	β_{SB}	−0.1	−0.0910	−0.0849	0.0236	**0.639**	100	−0.0821	0.0243	**0.733**	93	−0.083	0.0247	**0.671**	92
	μ_m	1.5	1.49	1.49	0.0247	**0.298**	70	1.47	0.0255	**1.14**	55	1.45	0.0259	**1.63**	97
Mod. (16) w^F	β_L	−0.5	−0.501	−0.423	0.0108	7.05	0	−0.459	0.0115	3.45	0	−0.481	0.0120	**1.50**	36
	β_{SB}	−0.1	−0.0910	−0.0796	0.0214	**0.948**	96	−0.0750	0.0224	**1.11**	76	−0.0784	0.0233	**0.923**	94
	μ_m	1.5	1.49	1.65	0.0228	6.91	0	1.57	0.0243	3.21	4	1.51	0.0252	**0.716**	60
Mod. (16) w^L	β_L	−0.5	−0.501	−0.480	0.0120	**1.60**	32	−0.497	0.0122	**0.201**	78	−0.502	0.0123	**0.213**	98
	β_{SB}	−0.1	−0.0910	−0.0847	0.0235	**0.645**	100	−0.0827	0.0244	**0.707**	100	−0.0832	0.0247	**0.677**	100
	μ_m	1.5	1.49	1.49	0.0247	**0.265**	68	1.46	0.0256	**1.25**	40	1.45	0.0259	**1.62**	62
Mod. (16) $w = 1$	β_L	−0.5	−0.501	−0.540	0.0153	2.60	0	−0.543	0.0136	3.18	0	−0.540	0.0125	3.20	0
	β_{SB}	−0.1	−0.0910	−0.124	0.0292	**0.836**	62	−0.0913	0.0269	**0.320**	100	−0.0919	0.0273	**0.295**	100
	μ_m	1.5	1.49	1.25	0.0273	9.06	0	1.35	0.0289	5.05	0	1.36	0.0289	4.69	0
Mod. (19)	β_L	−0.5	−0.501	−0.176	0.0102	31.5	0	−0.171	0.0100	32.7	0	−0.166	0.00974	34.1	0
	β_{SB}	−0.1	−0.0910	−0.0687	0.0236	**1.32**	5	−0.0671	0.0236	**1.39**	24	−0.0658	0.0230	**1.48**	8
	μ_m	1.5	1.49	1.62	0.0423	2.85	0	1.65	0.0410	3.77	0	1.70	0.0398	5.02	0
$\theta = 0.01$															
Mod. (16) w^G	β_L	−0.5	−0.501	−0.568	0.0146	4.68	0	−0.532	0.0131	2.49	0	−0.514	0.0124	**1.14**	62
	β_{SB}	−0.1	−0.0910	−0.112	0.0295	**0.413**	100	−0.0939	0.0269	**0.224**	100	−0.0843	0.0254	**0.614**	100
	μ_m	1.5	1.49	1.21	0.0320	8.76	0	1.34	0.0266	5.96	0	1.42	0.0259	2.94	4
Mod. (16) w^F	β_L	−0.5	−0.501	−0.567	0.0146	4.62	0	−0.530	0.0129	2.36	2	−0.511	0.0123	**0.957**	85
	β_{SB}	−0.1	−0.0910	−0.112	0.0294	**0.437**	100	−0.0939	0.0267	**0.228**	100	−0.0852	0.0251	**0.585**	100
	μ_m	1.5	1.49	1.21	0.0315	8.90	0	1.35	0.0266	5.50	0	1.43	0.0259	2.33	2
Mod. (16) w^L	β_L	−0.5	−0.501	−0.568	0.0146	4.69	0	−0.532	0.0131	2.46	0	−0.516	0.0125	**1.31**	44
	β_{SB}	−0.1	−0.0910	−0.112	0.0295	**0.421**	100	−0.0941	0.0269	**0.215**	100	−0.0857	0.0256	**0.553**	100
	μ_m	1.5	1.49	1.21	0.0320	8.78	0	1.34	0.0266	5.86	0	1.41	0.0259	3.41	2
Mod. (16) $w = 1$	β_L	−0.5	−0.501	−0.569	0.0146	4.72	0	−0.530	0.0129	2.36	2	−0.512	0.0126	**0.965**	72
	β_{SB}	−0.1	−0.0910	−0.112	0.0295	**0.435**	90	−0.0939	0.0267	**0.228**	100	−0.0852	0.0250	**0.588**	100
	μ_m	1.5	1.49	1.21	0.0319	8.85	0	1.35	0.0266	5.50	0	1.43	0.0261	2.39	2
Mod. (19)	β_L	−0.5	−0.501	−0.140	0.00918	39.1	0	−0.147	0.00966	36.4	0	−0.162	0.0106	31.7	0
	β_{SB}	−0.1	−0.0910	−0.0652	0.0200	**1.72**	0	−0.0720	0.0216	**1.29**	42	−0.0844	0.0248	**0.625**	98
	μ_m	1.5	1.49	1.84	0.0353	9.77	0	1.71	0.0346	6.15	0	1.50	0.0334	**0.140**	62

Bold number means its value (*t*-test true) is smaller than 1.96, indicating the estimate is not significantly different from the true value.
[a] The estimates obtained by the model with the full choice set, as shown in Table 1.

② $$\Pr(i \mid D, D', w) = \frac{\exp(V_i + \ln G_i(\hat{D}', w) + \ln \frac{k_i}{b(i)})}{\sum_{j \in D} \exp(V_j + \ln G_i(\hat{D}', w) + \ln \frac{k_i}{b(i)})}$$

③ $$\Pr(i \mid D, D') = \frac{\exp(V_i + \ln G_i(\hat{D}', 1)}{\sum_{j \in D} \exp(V_j + \ln G_i(\hat{D}', 1)}$$

$$w_j^G = \frac{k_j}{E[k_j]} = \frac{k_j}{q(j)R} = \frac{k_j B}{b(j)R}$$

$$w_j^F = \begin{cases} \frac{B}{b(j)R} & if\ b(j)R > B \\ 1 & otherwise \end{cases}$$

$$w_j^L = \frac{k_j}{q(j)R} \approx \frac{k_j}{q(j)R} \frac{q(s)R}{k_s} = \frac{k_j}{k_s} \frac{b(s)}{b(j)}$$

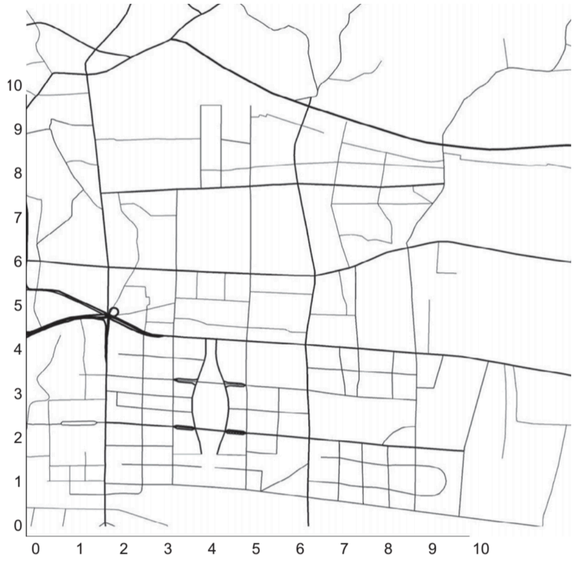
$w = 1$

実データを使ったモデルの検証①

モデルの検証

合成データを用いてモデルの精度の検証をする．その設定を確認する．

設定



道路ネットワークは中国の広州のCBD．ネットワークは208のノード，662の片方向リンクを有し，リンクの内訳は，24の主要道，34の幹線道路，32の補助道路．また，57の信号が含まれている．
データはGPSから得たタクシー運転手の軌跡から読み取ったもの．740のトリップが観測された．$Length_i$ が経路長，$ArteryRoadRatio_i$ が経路長に占める主要道と幹線道路の割合（長さ），$Signal_i$ が経路に含まれる信号の数として，

$$V_i = \beta_L Length_i + \beta_{ARR} ArteryRoadRatio_i + \beta_S Signal_i$$

と経路ごとの効用の確定項は表現される．

G関数のパラメータの設定は以下の通り．

- $\alpha_{im} = l_m / L_i$
- μ_m：補助道路のリンクのネストに関するスケールパラメータ
- μ_{mA}：主要道・幹線道路のリンクのネストに関するスケールパラメータ
- $\mu = 1$

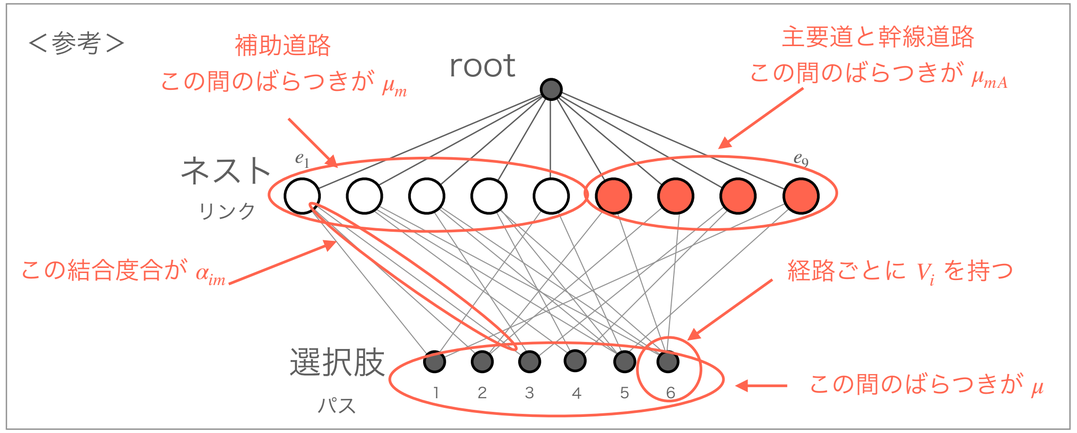


Table 5
Statistics on routes attributes.

Attributes	Min.	Average	Max.
Length (km)	0.447	2.84	8.34
Artery road ratio	0	0.797	1
Number of signal-controlled intersections	0	3.66	14

道路ネットワークの様子

実データを使ったモデルの検証②

モデルの検証

合成データを用いてモデルの精度の検証をする．その設定を確認する．

Table 6
Average size of the generated choice sets (with 10,000 draws).

θ	$\lvert\mathcal{D}\rvert$	θ	$\lvert\mathcal{D}\rvert$
0.005	29	0.0025	3813
0.004	54	0.0023	5624
0.003	201	0.002	7766
0.0028	2036	0.001	9836

設定

使用するモデルはCNLモデルに加え，サンプリング補正をそれぞれ加えたLogitモデルとPass size logit(PSL)モデルを用いる．

$$\Pr(i\,|\,D)=\frac{\exp(V_i+\ln\frac{k_i}{b(i)})}{\sum_{j\in D}\exp(V_j+\ln\frac{k_j}{b(j)})}\quad\cdots\text{Logit}$$

$$\Pr(i\,|\,D)=\frac{\exp(V_i+\ln EPS_i+\frac{k_i}{b(i)})}{\sum_{j\in D}\exp(V_j+\ln EPS_j+\ln\frac{k_j}{b(j)})}\quad\cdots\text{PSL}$$

$$EPS_i=\sum_{a\in\Gamma}\frac{l_a}{L_i}\frac{1}{M_a^{EPS}},\ M_a^{EPS}=\sum_{j\in C}\delta_a w_j.\ w_j=1,\ \text{if}\ \delta_j=1\ \text{or}\ q(j)R_n\geqslant 1;\ w_j=\frac{1}{q(j)R_n},\ \text{otherwise}$$

$$\Pr(i\,|\,D)=\frac{\exp(V_i+\ln G_i(\hat{D}',w)+\ln\frac{k_i}{b(i)})}{\sum_{j\in D}\exp(V_j+\ln G_i(\hat{D}',w)+\ln\frac{k_i}{b(i)})}\quad\cdots\text{CNL}$$

CNLの場合サンプリングの補正 $\ln\frac{k_i}{b(i)}$ がある場合とない場合と調べる． 今回は $|C|$ が大きくて大変なため，拡大係数の w^G と w^F は使えず，w^L にする．

M-Hサンプリングをする際の θ は大きくても小さくてもダメで，難しい．

→Logitモデルでおおよその $\beta_L, \beta_{ARR}, \beta_S$ の値を調べ，そこから θ を試していく．

→最終的に $\theta=0.003$ に（右上）

推定 / 検定

合成データと違い”正しい”パラメータはわからないので，半分ずつにデータをわけ，片方を”パラメータ推定用”，もう片方を”検定用”とする．組を変え，3回行った．

シミュレーション結果の要点を抜き出すと，以下の通り．

- 全ての場合においてCNLでは$\mu_m > \mu_{mA}$が成立した．ネストの間で，それぞれの内部での”ばらつき”が調べられるのはCNLの大きな長所の一つ．
- 調整済み尤度比に着目すると，サンプリング補正ありのCNLが最もいい．補正がないと，サンプリング数が多い時はPSLの方が結果が良かったりする．
- 計算速度に着目するとCNLの補正の有無で比較すれば大差ないが，例えばPSLと比べるとCNLは30倍以上時間がかかる．
- 推定したパラメータをもとに，検定用のデータを使って検定を行うと，
  - モデルによらずサンプリング数を増やせば検定時の尤度が上がること
  - 尤度が高いのは補正ありのCNL
  - 補正がなくてもCNLの方がPSLやLogitより尤度が高いこと

  が考察できた．

つまり，サンプリングの補正を行ったCNLが最もよく推定できていたのみならず，補正なしでもCNLの方がPSLらよりもよく推定できていたのだから，

CNLは経路選択問題における誤差相関を非常によく捉えたモデル

と言えるだろう．

実データを使ったモデルの検証③

モデルの検証

推定結果

Estimation results of exercise 1: real data case.

	Logit			PSL			CNL with corrections			CNL without corrections		
Draws	10–50	50–250	100–500	10–50	50–250	100–500	10–50	50–250	100–500	10–50	50–250	100–500
Estimation data set 1												
β_L est.	**−3.27**	**−3.50**	**−3.32**	**−2.15**	**−1.11**	**−0.982**	**−2.34**	**−1.90**	**−1.42**	**−1.42**	**−1.41**	**−1.30**
Std. err.	*−3.27*	*0.110*	*0.103*	*0.137*	*0.114*	*0.104*	*0.678*	*0.143*	*0.0686*	*0.174*	*0.113*	*0.112*
t-Test (0)	26.4	31.6	32.1	15.6	9.70	9.431	3.45	13.2	20.8	8.14	12.4	11.6
β_{ARR} est.	**10.6**	**9.67**	**9.45**	**11.1**	**10.8**	**10.7**	**10.6**	**9.55**	**10.0**	**6.30**	**4.80**	**4.86**
Std. err.	*10.6*	*0.845*	*0.686*	*8.76*	*0.830*	*0.671*	*0.850*	*1.25*	*1.09*	*1.95*	*3.65*	*2.05*
t-Test (0)	5.62	11.4	13.7	1.27	13.1	16.0	12.5	7.60	9.22	3.22	1.31	2.36
β_{IS} est.	**−0.0275**	**−0.0463**	**−0.233**	**−0.464**	**−0.834**	**−0.914**	**0.131**	**−0.869**	**−0.857**	**0.558**	**0.647**	**0.464**
Std. err.	*−0.0275*	*0.183*	*0.175*	*0.198*	*0.206*	*0.194*	*0.210*	*0.277*	*0.286*	*0.228*	*0.178*	*0.216*
t-Test (0)	−0.143	0.252	1.33	2.33	4.03	−4.69	0.625	3.13	2.99	2.44	3.62	2.14
β_{EPS} est.				**2.64**	**5.83**	**5.90**						
Std. err.				*0.249*	*0.205*	*0.187*						
t-Test (0)				10.5	28.4	31.5						
μ_m est.							**2.27**	**2.84**	**4.22**	**2.39**	**2.51**	**3.59**
Std. err.							*0.140*	*0.324*	*0.148*	*0.266*	*0.803*	*0.733*
t-Test (0)							16.2	8.77	28.4	8.98	3.12	4.89
μ_{mA} est.							**1.51**	**2.04**	**2.68**	**1.00**	**1.27**	**1.21**
Std. err.							*0.131*	*0.176*	*0.246*	*0.488*	*0.716*	*0.443*
t-Test (0)							11.5	11.5	10.9	2.05	1.77	2.73

各モデルのパラメータ推定値（データセット1）

補助道路の間のばらつきの方が
幹線道路・主要道の間のばらつきよりも大きい.

実データを使ったモデルの検証④

モデルの検証

推定結果

ここだけ比べると若干PSLの方がいい数値

修正済み尤度比が一番高い

段違いに時間がかかっている

Model fit measures and estimation times of three randomly generated data sets.

	Logit			PSL			CNL without corrections			CNL with corrections		
Goodness of fit measure $\bar{\rho}^2$												
Draws	10–50	50–250	100–500	10–50	50–250	100–500	10–50	50–250	100–500	10–50	50–250	100–500
Estimation data set 1	0.729	0.701	0.684	0.731	0.708	0.693	0.737	0.717	0.690	0.807	0.790	0.805
Estimation data set 2	0.723	0.768	0.688	0.725	0.715	0.698	0.725	0.720	0.691	0.799	0.792	0.805
Estimation data set 3	0.727	0.704	0.687	0.729	0.829	0.696	0.729	0.708	0.739	0.806	0.793	0.804
Estimation time t												
Draws	10–50	50–250	100–500	10–50	50–250	100–500	10–50	50–250	100–500	10–50	50–250	100–500
Estimation data set 1	57.8 s	87.4 s	98.7 s	61.2 s	94.8 s	116 s	22.3 min	3.38 h	7.76 h	25.6 min	4.04 h	9.76 h
Estimation data set 2	47.2 s	71.6 s	89.6 s	54.3 s	81.4 s	97.6 s	28.0 min	3.44 h	8.02 h	32.3 min	5.45 h	8.68 h
Estimation data set 3	45.1 s	63.2 s	79.2 s	54.8 s	66.9 s	97.7 s	25.6 min	3.61 h	7.13 h	39.2 min	5.73 s	8.41 h

モデルごとのパラメータ推定における修正済み尤度比

僅差で補正なしCNLの方が数値がいい

段違いにいい数値

Log likelihood values for the validation data sets.

	Draws	Validation data set 1	Validation data set 2	Validation data set 3	Sum
$L(0)$		−1705	−3652	−3606	–
Logit	10–50	−951	−710	−611	−2272
	50–250	−937	−706	−610	−2253
	100–500	−935	−699	−608	−2242
PSL	10–50	−938	−716	−613	−2267
	50–250	−948	−707	−613	−2268
	100–500	−949	−700	−603	−2252
CNL without corrections	10–50	−961	−667	−602	−2230
	50–250	−891	−662	−602	−2155
	100–500	−873	−568	−596	−2037
CNL with corrections	10–50	−826	−466	−428	−1720
	50–250	−789	−461	−425	−1675
	100–500	−778	−426	−408	−1612

モデルごとの検定における初期尤度と最終尤度

まとめ

経路選択問題におけるモデリングとサンプリング

1. 経路選択問題では，（大きなネットワークなら）サンプリングをしよう！

2. CNLはGEVモデルの一種で，「選択肢が複数のネストに所属している」ことがポイント！経路選択問題なら，”リンク”をそのまま”ネスト”にしてしまうことが一般的！

3. 経路選択問題では，CNLを使おう！またサンプリングをしたときは補正をしっかり加えよう！補正の加え方にもいろいろあるので，使える方法の中でいろいろと比較してみよう！

4. M-H法でサンプリングをするときには”分布”が必要”！経路選択問題ならば，「どのような確率頻度でサンプリングしたらいいのか掴むために先に簡単なモデルでパラメータ値を探ろう！

参考

大枠の内容	出典元	リンク
C-Logit model	bin	http://bin.t.u-tokyo.ac.jp/kaken/pdf/2014_oyama1.pdf
model -summer school 2016	bin	http://bin.t.u-tokyo.ac.jp/model16/lecture/Chikaraishi.pdf
CNL(cross-nested-logit) model	bin	http://bin.t.u-tokyo.ac.jp/kaken/pdf/CNL-instruction.pdf
様々な経路選択モデル	bin	http://bin.t.u-tokyo.ac.jp/startup14/file/2-2.pdf
経路行動分析	名大	http://www.trans.civil.nagoya-u.ac.jp/~miwa/doc_paper/Chapter5.pdf
サンプリング（M-H samplingなど）	bin	http://bin.t.u-tokyo.ac.jp/summercamp2015/document/prml11_chika.pdf
MEV generating function	EPFL	https://transp-or.epfl.ch/courses/dca2018/slides/08-mev.pdf
GEVモデル	bin	http://bin.t.u-tokyo.ac.jp/startup16/file/3-2.pdf
“Sampling of alternatives in Multivariate Extreme Value Models”	論文	https://www.sciencedirect.com/science/article/abs/pii/S0191261512001518

あとbinWikiの過去の資料から前田さんの資料（GEVモデルの導出）と原さんの資料（サンプリング）を参考にしました．フォルダに入れておきます．