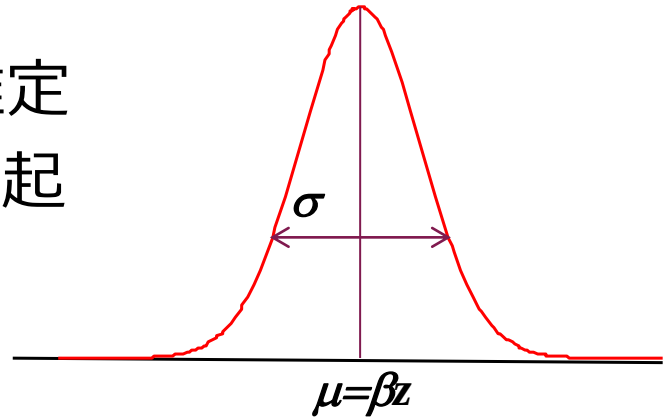
パラメータ推定の基礎

早稲田大学　佐々木邦明

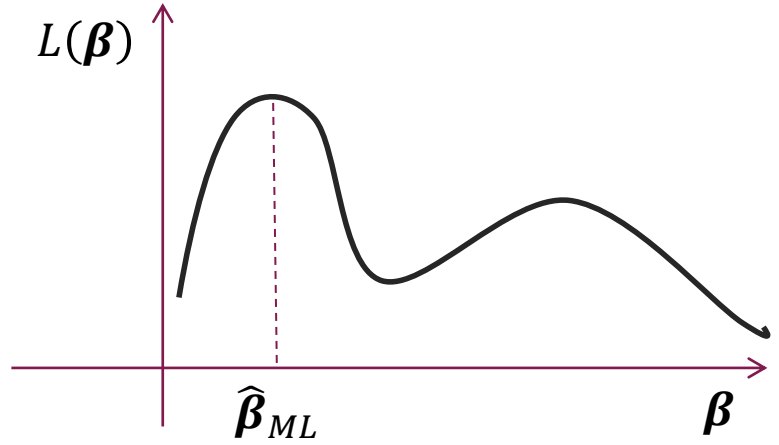
最尤推定

行動モデルの推定と最尤推定

- 有限個のパラメータで記述される確率密度関数の推定
- パラメータベクトル$\boldsymbol{\beta}$，モデル　f　による標本の生起確率を尤度とする
  - $L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(\boldsymbol{y}_i|\boldsymbol{\beta})$



- (対数)尤度関数が最大になるθ を最尤推定値とする
  - $\widehat{\boldsymbol{\beta}}_{ML} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \log L(\boldsymbol{\beta})$



最尤推定法

- 点推定量を求める一般的な方法
- 右上の式をθの関数とみなしたものが尤度関数
- 尤度関数を最大化する θの値を最尤推定量とするのが最尤推定法

$$L_n(\boldsymbol{\beta}|x) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\beta})$$

平均値の推定を例にすると

データ(**x：3,5,4**)が得られたとき,
平均をいくつとするのがよいか？

⇒平均がいくつの分布だったら

データ(**x：3,5,4**)がもっとも得られやすいか？

ロジットモデルの最尤推定

- $L(\mu\boldsymbol{\beta}) = \prod_{i=1}^{n} f(\boldsymbol{y}_i|\mu\boldsymbol{\beta})$

選ばれた選択肢の選択確率
y_{ij}=1: if j=選択, y_{ij}= 0:それ以外

- $f(\boldsymbol{y}_i|\mu\boldsymbol{\beta}) = \prod_{j=1}^{J} \left\{ \frac{exp(\mu\boldsymbol{\beta}\boldsymbol{x}_{ij})}{\sum_{j=1}^{J} exp(\mu\boldsymbol{\beta}\boldsymbol{x}_{ij})} \right\}^{y_{ij}}$

βは未知数， x は観測値

- $V_{ij} = \mu\boldsymbol{\beta}\boldsymbol{x}_{ij} = \mu\beta_1 + \mu\beta_1 x_{1i} + \mu\beta_2 x_{2i} \cdots + \mu\beta_K x_{Ki}$

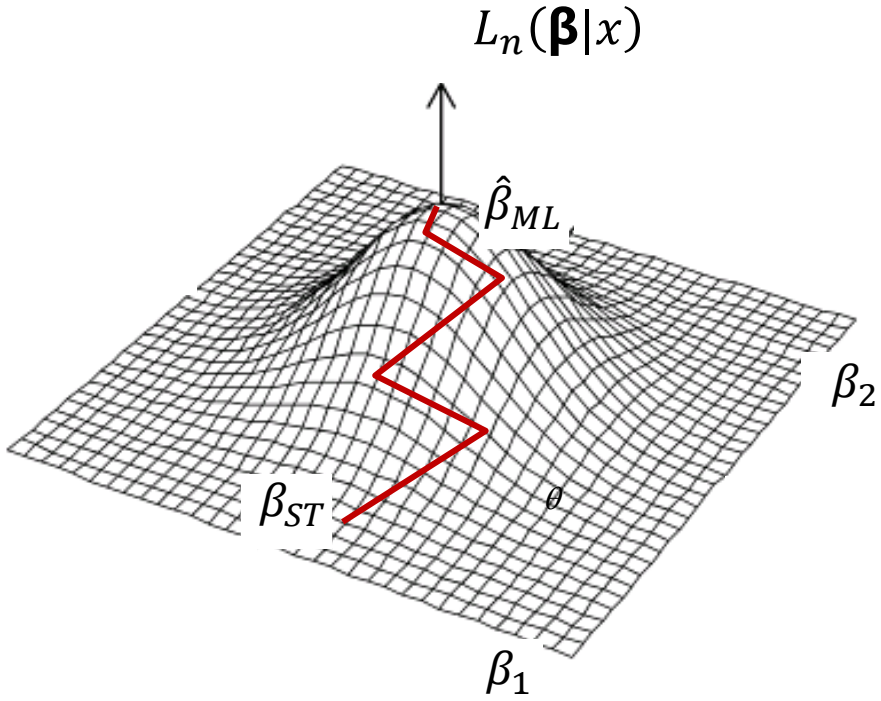
選択結果($\boldsymbol{y}_i$ ： y_1=車, y_2=車, y_3=鉄道, y_4=鉄道, y_5=鉄道, y_6=車, ・・)が得られたとき，$\mu\boldsymbol{\beta}$をいくつとすると，データへの適合がよいのか？

⇒μ$\boldsymbol{\beta}$がいくつだったらデータ(**y**)が得られやすいのか？
μ$\boldsymbol{\beta}$を色々と変えてみて一番**L**が高くなるμ$\boldsymbol{\beta}$を探す

最大化アルゴリズムの考え方

周りがあまり見えない中で，近傍の情報から頂点を目指す

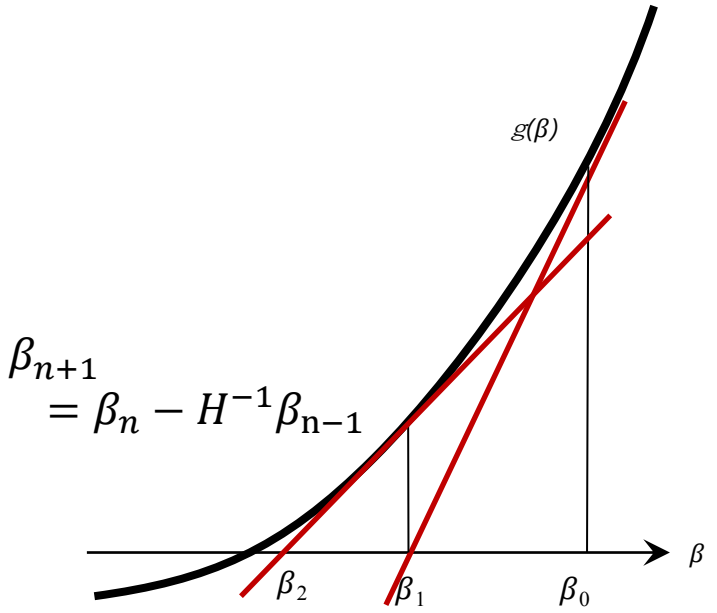
- 対数尤度関数の段階的な最大化
  - 初期値を与える
  - 初期値周りで勾配(1次微分）等を用いて次の推定値の方向を決める
  - 初期値付近で１次微分，２次微分を用いて適切に次の点を決めて推定値を得る
  - 収束基準(一次微分ベクトル)で判定し，収束していない場合は，現在の値から次の推定値に移る



代表的な繰り返し計算法

尤度関数を最大化　尤度関数の一階微分＝０を解く

- Newton-Raphson法
  - テイラー展開の１次近似を利用して進める
- 準Newton法（BFGS，L-BFGS法）
  - ヘッセ行列を，パラメータの差分と一階微分の差分を用いて逐次近似する.
  - L-BFGSはヘッセ行列の更新式を展開して，初期値と差分の関数和で表す.

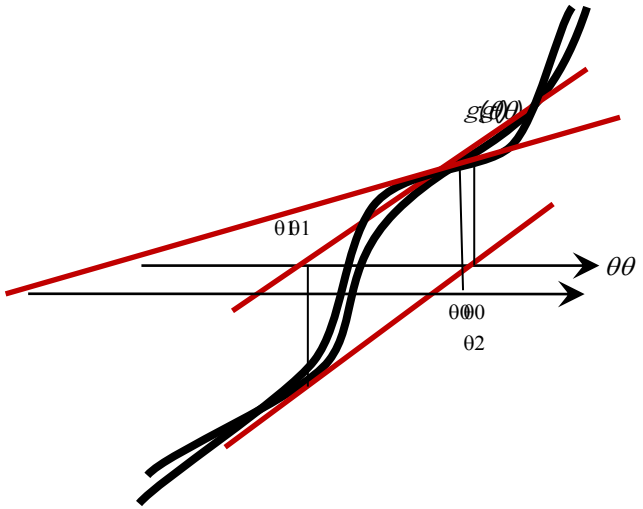


H:尤度関数の二階微分　ヘッセ行列
g: 尤度関数の一階微分

パラメータ推定がうまくいかない

- 収束するとはβ_{n+1}とβ_nが同じになる
  - g′が0になる
- 収束しない
  - 無限に繰り返す
  - β_2が計算不能
- 局所最適解
  - 見かけ上の最大化

- H^{-1}ヘッセ行列の逆行列が早々に死亡
  - **変数が完全相関**
  - **変数が効用関数に影響しないモデル**
- 関数の近似状況
  - 初期値の問題
- モデルに誤り
  - 意思決定者間で異なるが，選択肢間では異ならない変数
  - 選択肢間では異なるが，意思決定者間で異ならない変数



最尤推定法におけるモデル選択

- 真の確率密度関数を近似するものが含まれる必要がある
- ⇒フレキシブルなモデルを選ぶ

- 最尤推定は自由度の高さ前提
- ⇒自由度が低すぎるモデルは不適切（過適合）

$$\begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_{nn} \end{bmatrix}$$

- 平均対数尤度の比較（KL情報量）
  - 例：共分散行列を考える
    - (非)制約モデル（A対称行列，B対角行列，C対角行列で分散同一）を考えるとCはBに含まれ，BはAに含まれるので，平均対数尤度L*は必ず
    - L*(A)≧L*(B)≧L*(C)になる.

$$\mathrm{AIC} = -\sum_{i=1}^{n} \log f\left(x_i, \hat{\theta}_{ML}\right) + t$$

シミュレーションによる推定

シミュレーションによる尤度計算

手順

1. 誤差項の密度関数から選択肢の数の次元の(準)乱数を発生させる
2. この乱数を誤差の値として，各代替案の効用値を計算（積分）する
3. 代替案iの効用値とその他の代替案の効用との値を比較し，それらの大小関係を1-0の変数Gで記述する.
4. 1～3のステップを繰り返す．その反復回数をRとする.
5. シミュレーションされた確率は
となり，この値は不偏推定量である.

効用を確定値にする

確定的に選択を決定

$$P_i = \frac{1}{R}\sum_{r=1}^{R} G^r$$

比率を確率に置き換える

これを尤度として最大になるようにパラメータをアップデートする

準乱数の例

- 準乱数の例としてHalton数列がある．その計算方法は素数pに対して

$$s_{t+1} = \left\{ s_t, s_t + 1/p^t, s_t + 2/p^t \cdots, s_t + (p-1)/p^t \right\}$$

例えばp=3ならば，初期値0として 1 /3, 2/3, 1/9, 4/9, 7/9, 2/9, 5/9, 8/9・・・

- 多次元化
  - 数列の異なる素数pを決めて，それぞれに応じて数列を作り多次元化する．
- 正規分布化
  - 数列を制約つきの乱数発生と同様の変換をして正規分布化

シミュレーションベースのパラメータ推定法

- シミュレーション尤度最大化（MSL）
  - シミュレーションによって計算された確率を尤度として，最大化を行う.
- 特性
  - サンプル数と乱数発生回数に依存する.
  - 乱数発生回数が十分大きいと一致性や漸近的有効性を持ち解析積分と一緒の特性を持つ.
  - 乱数発生回数がサンプル数に対して小さく固定されると一致性もない.

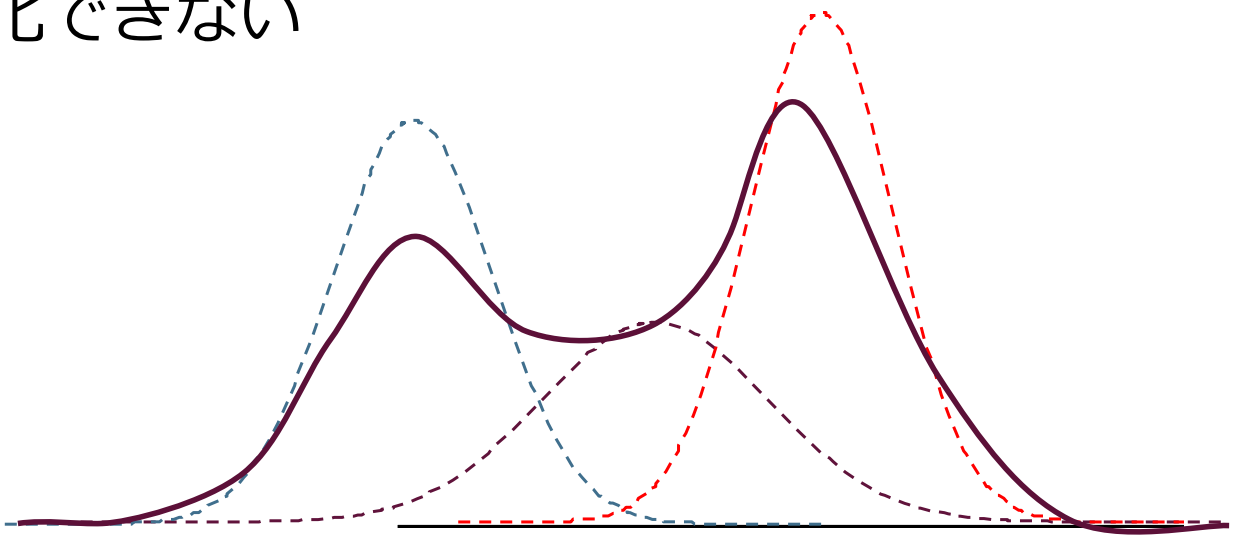
EM-アルゴリズム

E-Mアルゴリズムの適用事例

- 混合モデル
  - $f(\boldsymbol{x}_i|\theta) = \sum_{i=1}^{m} w_i \emptyset(\mu, \sigma^2)$ $L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{x}_i|\boldsymbol{\theta})$
  - $\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}} \log L(\theta)$ $subject\ to \begin{cases} w_{1,} \cdots w_m \geq 0 \\ \prod_{i=1}^{m} w_i = 1 \end{cases}$

媒介変数

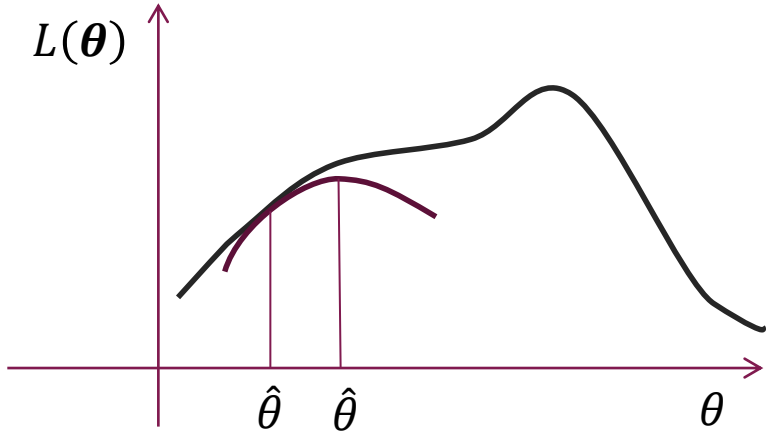
- 媒介変数を用いて尤度関数を表現できる．$w_l = \frac{exp(\gamma_l)}{\sum_{l'=1}^{L} exp(\gamma_{l'})}$
- ただし簡単に最大化できない



混合モデルの推定技法

- EM法
  - 不完全データの最適化法
  - 混合モデルは不完全データからの学習法

- 適当な初期値を定める
- 初期値に応じて媒介変数を求める（E）
- 求めた媒介変数から解を計算する（M）
- 対数尤度関数は減少せず，局所最適解に収束する

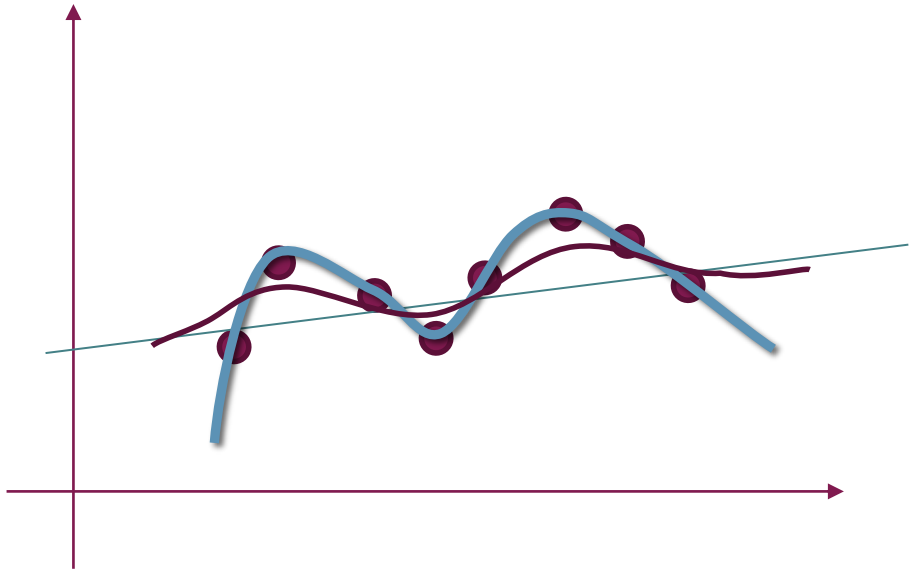


学習（LEARNING）

パラメータ推定と学習

- 機械学習における学習
  - 判断の根拠となるための統計的なモデルを作る過程
  - 統計的機械学習
- 機械学習の主な目的は「予測」
  - ある移動手段がどの程度選ばれそうか
  - ある個人が車を購入しそうか

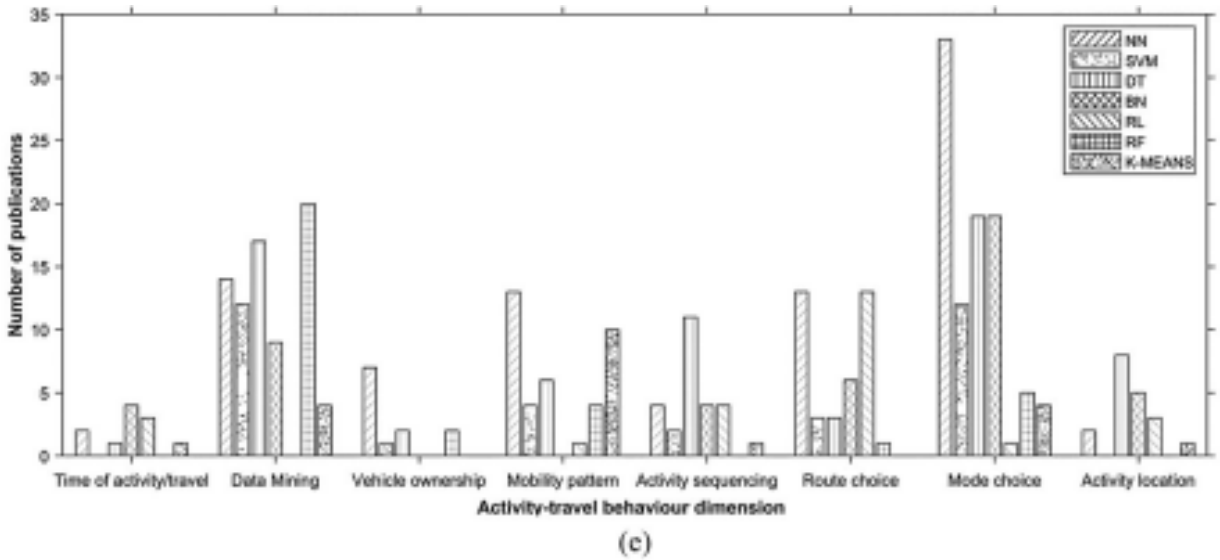
- フィッティング



  - 仮説に基づく制約をモデルとせず，予測精度が上がるようにモデルを自由に作る
  - ハイパーパラメータ

機械学習と交通行動分析

- 機械学習の利点
  - 行動原理の仮定が不要
  - 非線形関係をモデル化して予測精度向上
  - ノイズの多いデータを扱うことが可能
  - カテゴリデータ，順序変数なども効率的に扱い計算時間も短い
  - 外れ値に対して頑健



Koushik, Manoj & Nezamuddin (2020)

各手法と課題

- NN
  - 三層NNでおおむね近似できる(Cybenko, 1989)
  - 離散・連続・カテゴリーデータを柔軟に扱えて，多重共線性を気にしなくていい（Henshcer and Ton, 2000）
  - たいていの場合，予測性能がMNLよりも優れている（Hussain et al., 2017, Assi et al. 2018）
  - 時間・空間的移転性は弱い（Henshcer and Ton, 2000, Mozolin et al., 2000, Tang et al., 2018)
    - 過剰適合が主な原因
- SVM
  - NNと比較して，高速でオーバーフィッティングも少ない.
  - モード選択，GPSデータマイニング，ライフスタイル分類に適用
  - データの量によっては過剰適合する (Allahviranloo & Recker, 2003)
  - 基本バイナリ分類機なのでマルチクラスの問題には適用が難しい

Cont.

- デシジョンツリー
  - ALBATROSSで使用実績あり
  - 意思決定のプロセスではないが，意思決定に関与する変数の理解に使うことが可能（Beckman & Goulias, 2008, Hafezi et al., 2017)
  - DTとMNLの比較で一致する（Yamamoto et al.，2002)
  - DTは頑健性が弱く，データの変化に対して木の構造が変わってしまう可能性ががある（Witten et al., 2011)
- EL (Ensemble Learners)
  - 頑健性が高く，ノイズの影響が小さい
  - アンサンブルの木を増やしても過剰適合しない
  - GPSからのデータ抽出
    - モード検出，目的予測，トリップ構成
  - アンサンブルなだけに，モデルの解釈性が劣る．変数の重要度を計算したものもあるが，意思決定プロセスが不明確

Explainable AI　(XAI)

- ランダムフォレストを用いたアクティブトラベルの選択問題

- SHAPバリュー
  - Shapley値（のようなもの）を各要素について推定
  - 判別に大きな影響をもたらす変数とその割合を示す
  - 右図では，アクティブトラベルの選択に最も大きな影響を持つ要因は自動車保有

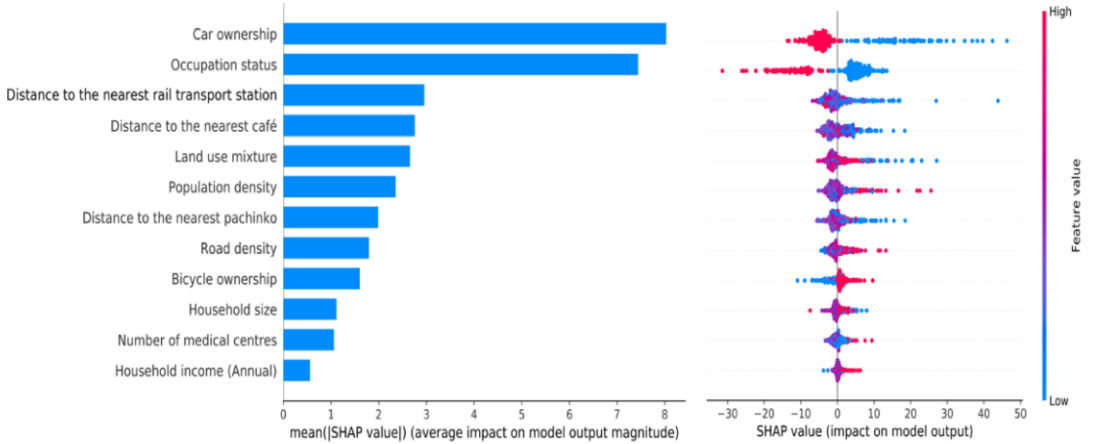


Fig. 6.2. Relative importance of independent variables and a summary of local explanations (Male)

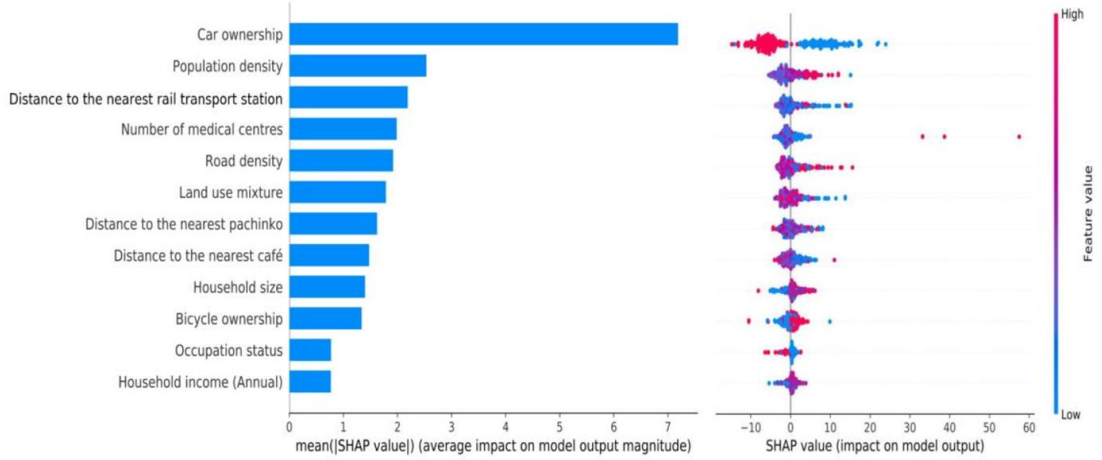


Fig. 6.3. Relative importance of independent variables and a summary of local explanations (Female)

(Yang et. al, 2022)

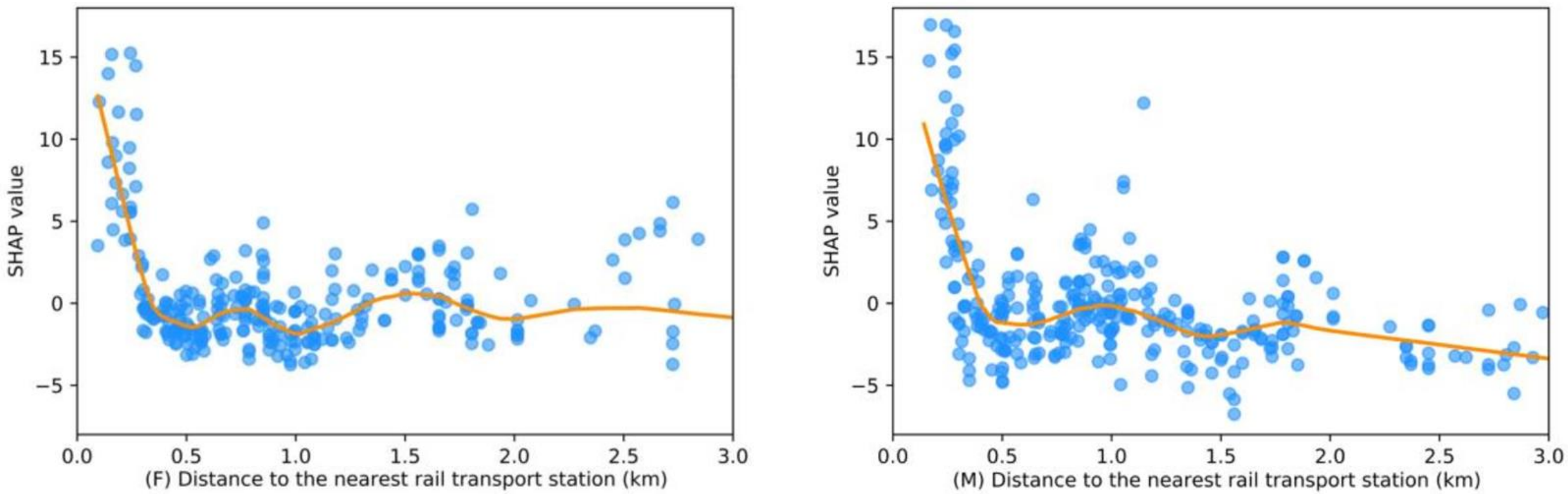


Fig. 6.7. The SHAP value plot of distance to the nearest rail transport station on active travel time

時系列・ビッグデータと推定

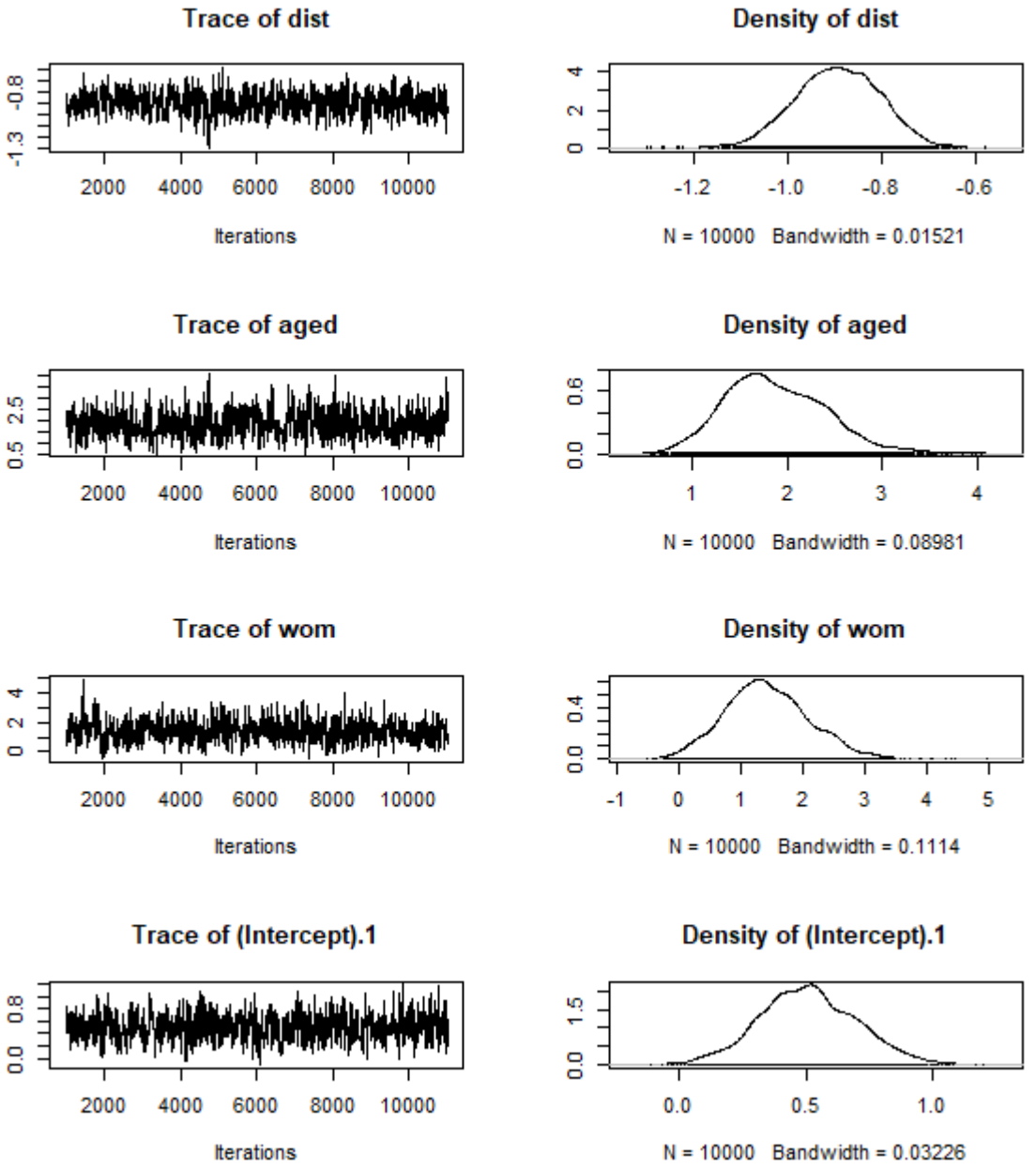
データサイズとビッグデータ

- Gelman，A. (2005) “N is never large”
  - サンプル数が大きい場合，というものは存在しない．もしNが小さすぎて十分な推定値を得ることができないならば，データをもっと増やす(もしくはもっと多くの仮定を用いる)必要がある．しかし，Nが十分に大きいならば，データを分割してもっと多くの情報を得ることができる．
  - Nが十分であることはない．もし「十分」だとしたら，もっと多くのデータを必要とする次の問題に取り組んでいるはずだ．

  - ビッグデータを利用して予測分析する問題は，比較的単純なアルゴリズムが使われる.もっと難しい問題は，スモールデータの場合である．(C. Davidson, 2009)

計算機による疑似乱数

- MCMC法
- 前の状態に基づいて（Markov Chain）新しいパラメータをランダムにサンプリング（Monte Carlo）する
- 最尤推定と違う
  - 最尤推定⇒最適化＝どっかで止まる
  - MCMC⇒分布を再現＝いつまでも動く
- 乱数を使った単純操作の繰り返しで事後確率の大きいところを探しながらサンプルを生成する
- カーネルの比較　$\frac{p(X|\theta')p(\theta')}{p(X)}\frac{p(X)}{p(X|\theta)p(\theta)}$



行動モデルを用いた政策分析

- **行動モデルを推定し，そのパラメータを用いて，変数の変化による選択の変化を見る.**

- 政策：変数の変化
  - 例えば：**所要時間**を短縮，**駐車場の料金**を割り引く等，政策に対応した変数が必要

- 政策評価
  - 数え上げ法：個人の選択確率を予測して，積み上げる　$S(j) = \frac{1}{N}\sum_{i=1}^{N} P_i(j)$
    - 最大効用の選択肢をカウント
    - 確率の平均値を求める

シミュレーションによる政策分析

- 行動モデルを推定し，そのパラメータを用いて，変数の変化による選択の変化を見る.
- 回遊行動の分析
  - マイクロシミュレーションを用いることで，複雑なモデルの組み合わせを政策評価可能
  - シミュレーションなので，複数回実施して平均的な評価を行う

Step 1	サンプルのデータ，個人属性や発ゾーン，LOSデータなどを各段階のモデルに個人nの個人属性やLOS，各種ダミー等といった説明変数データを代入し，全ての選択肢ごとの選択確率P_{in}を求める．求めた選択確率から確率分布F_{in}を作成する.
Step 2	$[0,1]$の一様乱数γ_nを発生させ，γ_nの値が，$F_{(i-1)n} \leq \gamma_n < F_{in}$を満たす選択肢$i$を選択するものとする.