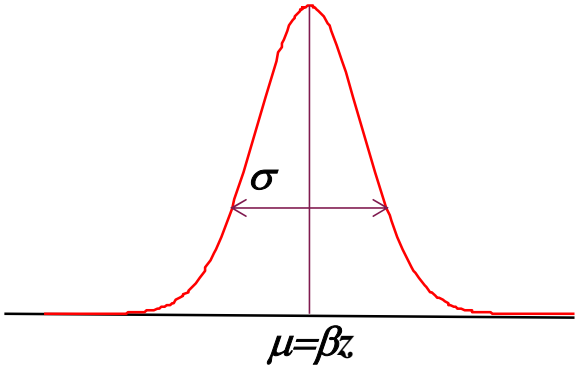
パラメータ推定の基礎と学習

早稲田大学　佐々木邦明

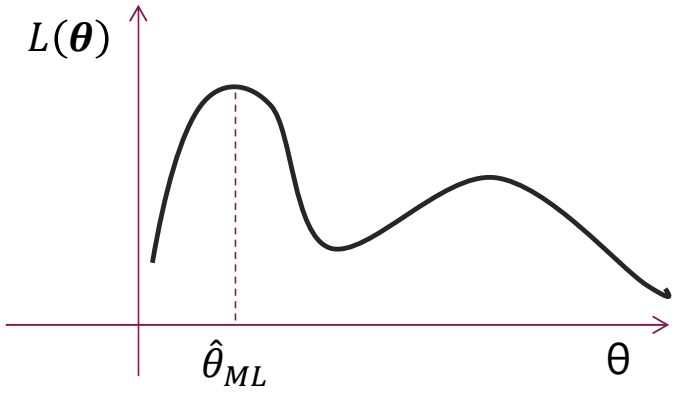
最尤推定

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θの下で，モデルfによる標本の生起確率を尤度とする
  - $L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\boldsymbol{y}_i|\boldsymbol{\theta})$



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



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

$$L_n(\theta \mid x) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

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

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

選ばれた選択肢の選択確率

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

βは未知数，x は観測値

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

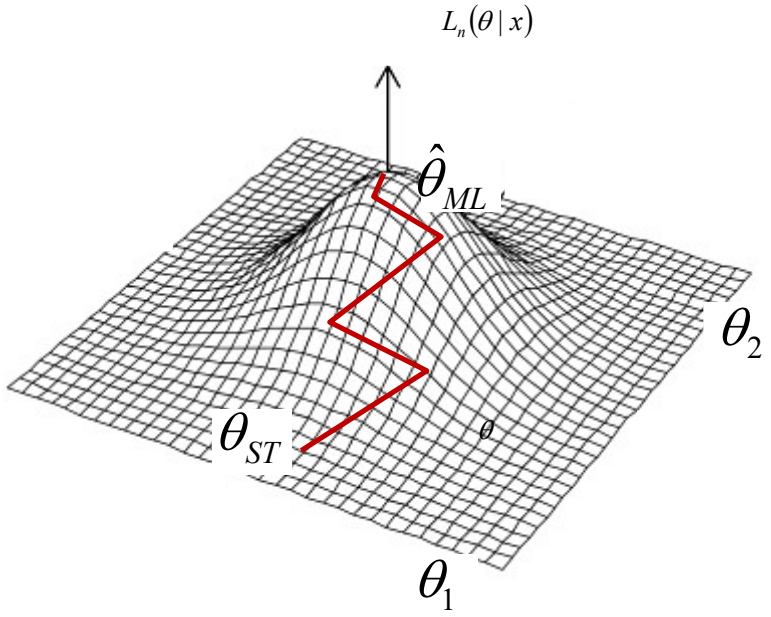
データ(**y**：車, 車, 鉄道, 鉄道, 鉄道, 車, 鉄道, ・・)が
得られたとき，$\boldsymbol{\beta}$をいくつとすると，再現性が高いのか？

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

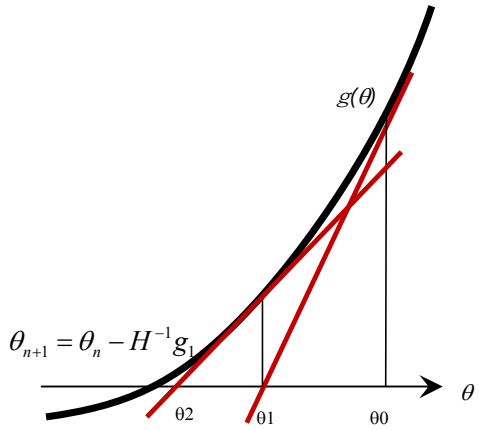
- 対数尤度関数の段階的な最大化
  - 初期値を与える
  - 初期値周りで勾配(1次微分）等を用いて次の推定値の方向を決める
  - 初期値付近で１次微分，２次微分を用いて適切に次の点を決めて推定値を得る
  - 収束基準(一次微分ベクトル)で判定し，収束していない場合は，現在の値から次の推定値に移る



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.

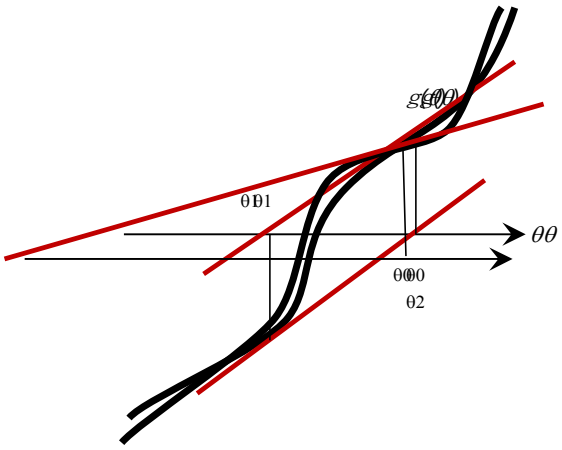


H:尤度関数の二階微分　ヘッセ行列
g: 尤度関数の一階微分

パラメータ推定がうまくいかない

- 収束するとはθ_{n+1}とθ_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
  - θ_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

- 平均対数尤度の比較（KL情報量）

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

  - 例えば，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.

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

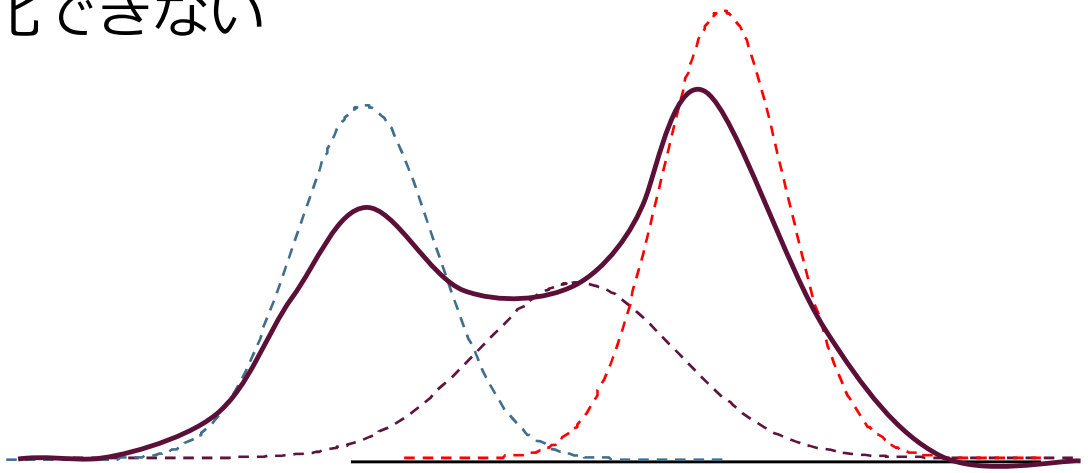
EM-アルゴリズム

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

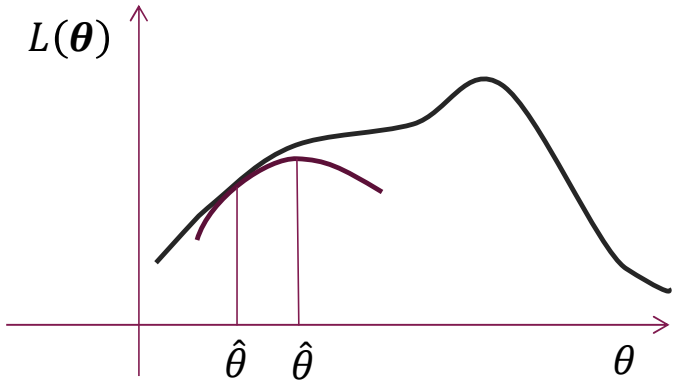
- 媒介変数を用いて尤度関数を表現できる．$w_l = \frac{exp(\gamma_l)}{\sum_{l'=1}^{L} exp(\gamma_{l'})}$
- ただし簡単に最大化できない



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



EMアルゴリズムとK-Means法

- クラスターへの帰属は未知数
  - クラスターの中心をランダムに設定
  - 上記で決まったクラスターを用いてクラスター中心を再計算
  - 新しいクラスター中心からクラスターの帰属を割り付けなおす

- ２つのプロセスを繰り返してクラスター中心までの距離の最適化

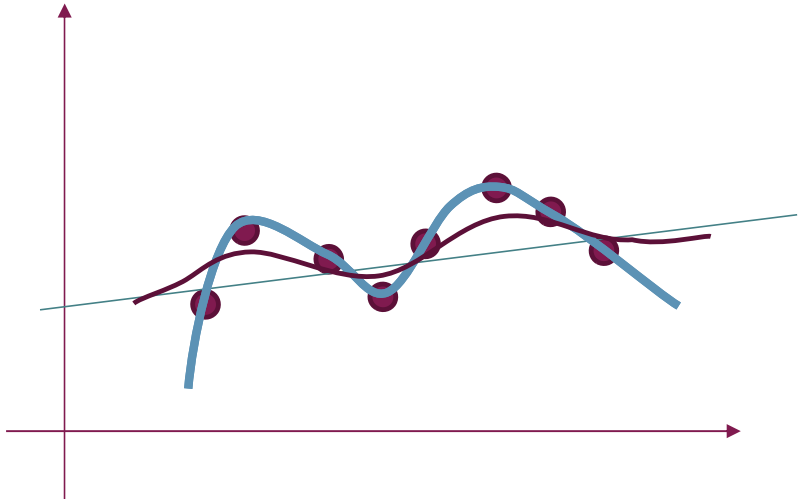
  - モデル推定　→　Learning
  - 仮定　→　ハイパーパラメータ

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

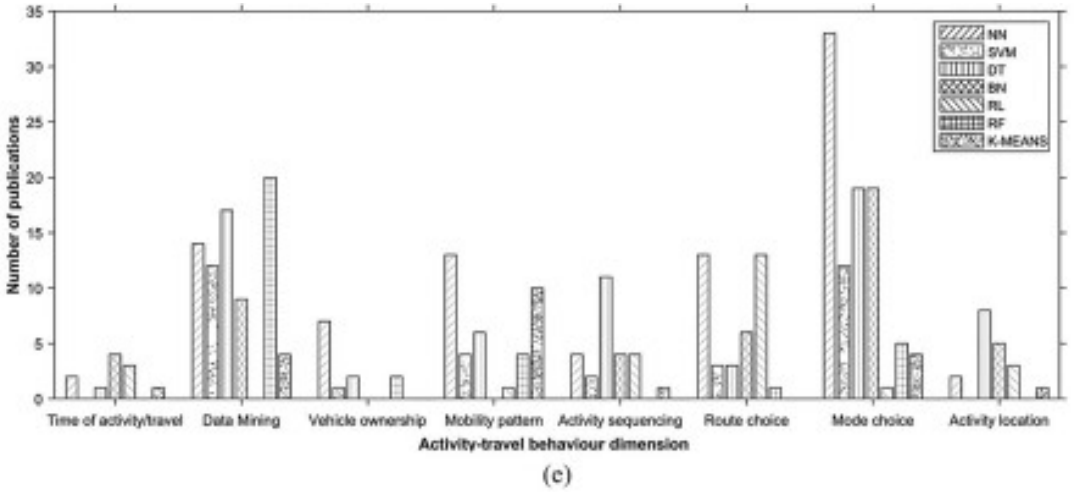
モデルのパラメータ推定と学習

- 1990年代後半～2000年代
  - 行動モデル界における派閥争い
- 事実発見派　　〇大学△先生
- 予測主義派　　△大学〇先生
  - 行動モデル：条件Sの時，（効用最大化した結果として）行動Aが選ばれる
  - 行動計量モデル
    - 選択結果とその時の条件データをもとに，要素のインパクトを，仮定した行動原理にしたがったモデルのパラメータとして求める

- 行動と機械学習
  - 多項ロジットモデルでの分析では，IIA の問題がうるさく言われる
  - SoftMax関数（ロジット関数）を用いた機械学習による分析でIIA の話はない
- 機械学習は，人間が行動に関して知識を得る(Fact Findings)というよりも予測(Forecasting）

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.

- DT（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)
  - 実質的にRandom Forestのこと
  - 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からのデータ抽出に使われる
    - 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

機械学習では変数の効果が不明確

- Connection weightを用いた手段選択モデルのパラメータの感度分析 Golshani et. al(2018)

- 結論
  - 行動に関しての知識が限定される場合，予測精度が重要視される場合には優れる

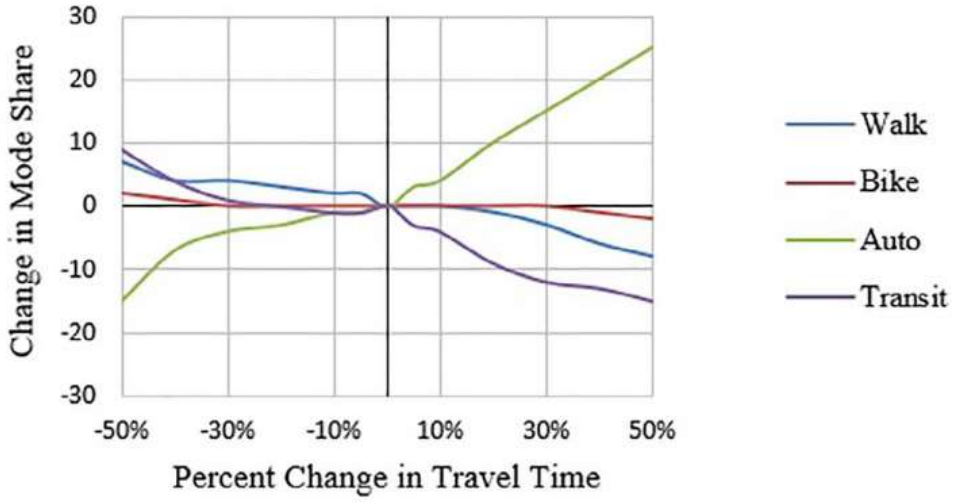


Fig. 5. Sensitivity analysis of important exploratory variables in NN mode choice model.