初学者のためのRチュートリアル

ベイズ推定の基礎 & 構造化プロビットモデル

ベイズの定理の導出

周辺分布　$f(a) = \int f(a,b)db,$　条件付き分布　$f(b|a) = \frac{f(a,b)}{f(a)}$　より

$$f(\theta|y) = \frac{f(\theta)f(y|\theta)}{f(y)} = \frac{f(\theta)f(y|\theta)}{\int f(\theta,y)d\theta}$$

θ　：パラメータ

$f(y|\theta)$：モデルの尤度 (確率密度)

周辺尤度 (モデルエビデンス)

→ モデルからデータが出現する尤もらしさ

多くの場合　$\int f(\theta,y)d\theta$　は解析的に計算不可なため、サンプリングによる近似が必要

$$f(\theta|y) \propto f(\theta)f(y|\theta)$$

事後分布　　事前分布

メトロポリス-ヘイスティングス (MH) アルゴリズム

- ランダムウォークMH

$\sigma' = \sigma + \varepsilon,\qquad \varepsilon$: 平均0のランダム項, e.g., $\varepsilon \sim N(0,1)$

$$\alpha(\sigma,\sigma') = \min\left\{\frac{f(\sigma')f(y|\beta,\sigma')}{f(\sigma)f(y|\beta,\sigma)}, 1\right\},$$

$\alpha(\sigma,\sigma')$ の確率で σ' へ推移

$1-\alpha(\sigma,\sigma')$ の確率で σ' への推移を棄却し σ に留まる

◆ 演習1: ランダムウォークMHによる相関係数の推定

$\sigma \sim N(g_0, G_0)$, σの事前分布

$y_i \sim N(X_i\beta, \Sigma)$, $y_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4})'$,

$$\Sigma = \begin{pmatrix} 1 & \sigma & 0 & 0 \\ \sigma & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

ギブスサンプリング

- 正規分布に従う線形回帰モデル　$y_i \sim N(x_i'\beta, \sigma^2),\ i \in n$　を例に

事前分布　$\beta \sim \underline{N(b_0, B_0)},$　共役事前分布

$$f(\beta|y, \sigma^2) \propto f(\beta) f(y|\beta, \sigma^2),$$

事後分布　$\beta \,\big|\, y, \sigma^2 \sim \underline{N(\bar{b}, \bar{B})},$

$$\bar{b} = \bar{B}\left(B_0^{-1} b_0 + \sigma^{-2} \sum_{i\in n} x_i y_i\right),\quad \bar{B} = \left(B_0^{-1} + \sigma^{-2} \sum_{i\in n} x_i x_i'\right)^{-1}$$

完全条件付き分布の導出 → ギブスサンプリングによる効率的な近似が可能

潜在変数のデータ拡大 (data augmentation)

- 2項プロビットモデルを例に

$$y_i = 1 \text{ if } y_i^* > 0, \quad y_i = 0 \text{ if } y_i^* \leq 0, \qquad y_i^* \sim N(x_i'\beta, 1)$$

ベイズの定理より $f(\beta|y) \propto f(\beta) f(y|\beta)$

β と同時に潜在変数 y_i^* を完全条件付き分布からサンプリング (データ拡大)

$$f(\beta, y^*|y) \propto f(\beta) f(y^*|\beta) f(y|\beta, y^*)$$

$$f(\beta, y^*|y) \propto f(\beta) \prod_{i \in n} f(y_i^*|\beta) \{I(y_i^* > 0) I(y_i = 1) + I(y_i^* \leq 0) I(y_i = 0)\},$$

データ拡大: $y_i^* \sim TN_{[0,\infty]}(x_i'\beta, 1)$ if $y_i = 1$ and $y_i^* \sim TN_{[-\infty,0]}(x_i'\beta, 1)$ if $y_i = 0$

$$\beta | y, y^* \sim N(\bar{b}, \bar{B}),$$

$$\bar{b} = \bar{B}\left(B_0^{-1} b_0 + \sum_{i \in n} x_i y_i^*\right), \quad \bar{B} = \left(B_0^{-1} + \sum_{i \in n} x_i x_i'\right)^{-1}$$

多項選択の場合も同様に潜在変数を拡大することでギブスサンプリングが実装可能

Albert and Chib (1993) Bayesian Analysis of Binary and Polychotomous response Data, Journal of the American Statistical Association, Vol.88, No.422, pp.669-679.

構造化プロビットモデル (Yai, Iwakura,Morichi; 1997)

経路rを選択する確率 $$P_r = \int_{\varepsilon_1=-\infty}^{\varepsilon_r+V_r-V_1} \dots \int_{\varepsilon_r=-\infty}^{\infty} \dots \int_{\varepsilon_R=-\infty}^{\varepsilon_r+V_r-V_R} \Phi(\varepsilon)\, d\varepsilon_R \dots d\varepsilon_1,$$

$$\Phi(\varepsilon) = (2\pi)^{-\frac{R}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}\varepsilon\Sigma^{-1}\varepsilon^{\mathrm{T}}\right],$$

$$\Sigma = \sigma_1^2 \begin{pmatrix} L_1 & L_{12} & \cdots & L_{1R} \\ L_{12} & L_2 & \cdots & L_{2R} \\ \vdots & \vdots & \ddots & \vdots \\ L_{1R} & L_{2R} & \cdots & L_R \end{pmatrix} + \sigma_2^2 \begin{pmatrix} k_1 & k_{12} & \cdots & k_{1R} \\ k_{12} & k_2 & \cdots & k_{2R} \\ \vdots & \vdots & \ddots & \vdots \\ k_{1R} & k_{2R} & \cdots & k_R \end{pmatrix} + \sigma_3^2 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

鉄道経路選択に適用する場合

L_p：経路の路線長

L_{rq}: 経路間の重複区間長

k_p：乗り換え可能な駅の数

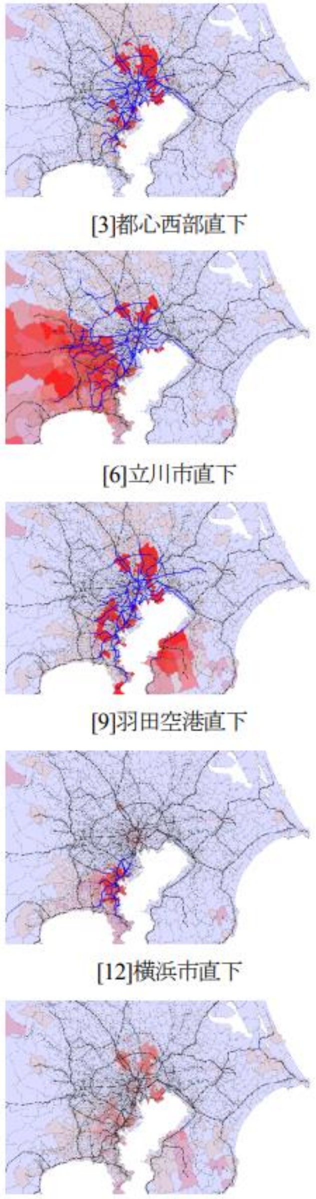
k_{rq}: 経路間で共通する乗り換え可能な駅の数

個人毎に異なる分散・共分散行列を表現

研究紹介:
構造化プロビットモデルを用いた都市鉄道ネットワークの脆弱性評価

- 震度分布データ (14ケース) を基に、東京都市圏においてリンクが途絶した場合の脆弱性評価指標 (OD間の一般化費用を交通量で重み付けしたものの合計値の変化分) を算出
- 一定震度以上の地域に含まれるリンクのうち特定の構造物を含むものを抽出し、それらのリンクが途絶したと仮定
- 途絶した場合利用者が他の経路へ迂回するため、一般化費用は増大

- 東京駅から約80km県内を対象地域とし、構造化プロビットモデルを用いて鉄道経路選択モデルを構築、リンク途絶による交通需要の変化を分析
- 都心直下だけでなく、周辺部でも途絶した場合の影響が大きいことが示唆



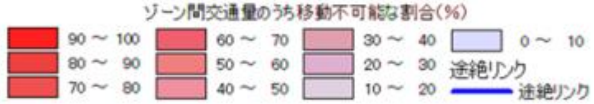


図-9 各ケースのゾーン間の移動不可能な交通量の割合

さらなる応用可能性について (屋井, 中川; 1996)

1. 交通機関の組み合わせ選択モデル

- 空路と新幹線の乗り継ぎなど、複数の交通機関の組み合わせ選択を記述
- 個人毎に組み合わせが異なる形で重複区間を分散共分散行列で表現

2. 目的地選択モデル

- 目的地間の相対的な近接性を考慮したモデリング

3. 出発時刻選択モデル

- 「時間帯」選択モデルを構築し、時間距離の考え方を用いて選択肢間の親近性を表現