機械学習特論演習
補足資料

遠里 由佳子

事後分布の解析的な推定

- 事前分布と尤度から事後分布を積分計算で解析的に算出する。
- 事後分布などを効率的に計算するために、事前分布と事後分布が同じ種類の確率分布を持つよう設計された、共役事前分布を使う。

（教科書 3.1.3節より抜粋）

表 3.1　尤度関数，共役事前分布，予測分布の関係．

尤度関数	パラメータ	共役事前分布	予測分布
ベルヌーイ分布	μ	ベータ分布	ベルヌーイ分布
二項分布	μ	ベータ分布	ベータ・二項分布
カテゴリ分布	$\boldsymbol{\pi}$	ディリクレ分布	カテゴリ分布
多項分布	$\boldsymbol{\pi}$	ディリクレ分布	ディリクレ・多項分布
ポアソン分布	λ	ガンマ分布	負の二項分布
1 次元ガウス分布	μ	1 次元ガウス分布	1 次元ガウス分布
1 次元ガウス分布	λ	ガンマ分布	1 次元スチューデントの t 分布
1 次元ガウス分布	μ，λ	ガウス・ガンマ分布	1 次元スチューデントの t 分布
多次元ガウス分布	$\boldsymbol{\mu}$	多次元ガウス分布	多次元ガウス分布
多次元ガウス分布	$\boldsymbol{\Lambda}$	ウィシャート分布	多次元スチューデントの t 分布
多次元ガウス分布	$\boldsymbol{\mu}$，$\boldsymbol{\Lambda}$	ガウス・ウィシャート分布	多次元スチューデントの t 分布

事後分布の近似的な推定

乱数を発生させサンプリングし、積分を近似する。

- マルコフ連鎖モンテカルロ法 (MCMC)
  - マルコフ過程(定常分布)を用いたサンプリングを行う。解析的に処理できない場合に向いている。
- メトロポリス・ヘイスティング法 (M-H法)
  - MCMCの代表的手法の１つ。
  - 提案した確率分布から生成された乱数を受け入れるか棄却するか判断するメトロポリス法で、分布の対称性を仮定しない場合を指す。
- ハミルトニアン・モンテカルロ法 (HMC法)
  - 確率変数をサンプリングするよう作られたM-H法の実装法を指す。ハイブリット・モンテカルロ法ともいう。
    - Stanで採用。
- ギブス・サンプリング
  - M-H法の応用にあたる。複数の確率変数の同時分布からのサンプリング手法。同時分布を、条件付き分布で近似して推定する。

Stanの特徴

- 推定方法に、HMC法の拡張にあたる No-U-Turn Sampler (NUTS) を採用
  - Hoffman MD. & Gelman A. (2011) https://arxiv.org/abs/1111.4246
- 自動変分ベイズでベイズ推定
  - 単純なMCMCより不正確だが高速
  - Automatic Differentiation Variational Inference (ADVI)
  - Kucukelbir A et al. (2015) https://arxiv.org/abs/1506.03431
- 基本的な手順
  - モデルを書いてコンパイル
  - データを読み込む
  - MCMCでサンプリング
- 親戚：Rstan, MatlabStan, Stan.jl(Julia)

Stan のコード例

```
stan_code = """
data {
    int N;
    array[N] int<lower=0, upper=1> y;
}
parameters {
    real<lower=0,upper=1> p;
}
model {
    p ~ beta(1, 1);
    for (i in 1:N) {
        y[i] ~ bernoulli(p);
    }
}"""
```

変数stan_dataの例:

```
{'y': [1, 0, 0, 1, 0, 1,
0, 1, 1, 1, 0, 0, 1, 0, 0,
0, 0, 1, 0, 0, 1, 0, 0, 1,
1, 1, 1, 0, 0, 1, 0, 0, 1,
0, 0, 1, 0, 1, 0, 0, 0, 0,
0, 0, 1, 0, 1, 0, 0, 0],
'N': 50}
```

```
model = stan.build(stan_code, data=stan_data, random_seed=1)
fit = model.sampling( # MCMCの実行
    num_warmup = 1000, # バーンイン期間
    num_chains = 4, #チェーン数
    num_samples = 2000, # 繰り返し
    num_thin = 1, # 間引き数
)
arvz.plot_trace(fit) # 結果の簡易表示
plt.show()
result = fit.to_frame() # データフレーム型での保存
```

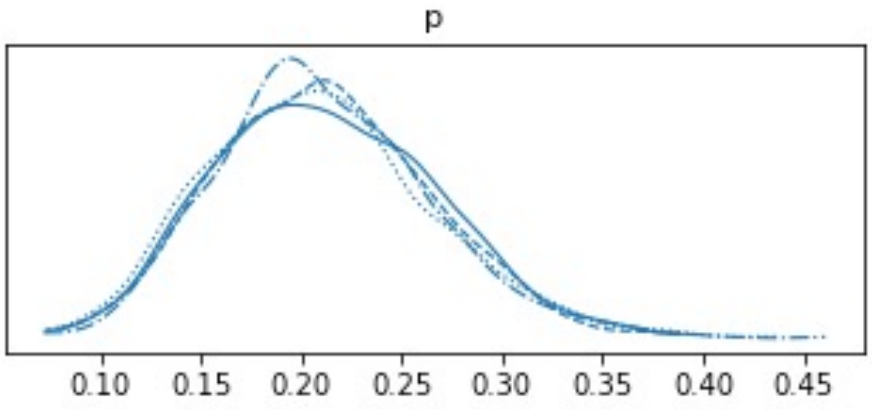
MCMC (Stan) の実行条件の補足

- バーン・イン期間(num_warmup: burn-in)：初期値に依存していると思われる捨てる部分
- MCMCの繰り返し回数（num_samples）
- チェーン数(num_chains)：
  - Iter回の試行を同時に実行するMCMCの数
  - ランダムサンプリングが原因で極端な結果を出す場合があり、それを確認するために、独立したMCMCを複数走らせる。
  - チェーン内の分散とチェーン間の分散の比率に基づくPSRF統計量、通称 $\hat{R}$ の計算（$\hat{R} < 1.1$ ならば収束判定）などに用いられる。
- 何個おきに結果を採用するか(num_thin)：乱数の自己相関をやわらげることが目的。num_samples=6でthin=2なら、1, 3, 5 の結果が用いられる
- (save_warmup: bool)

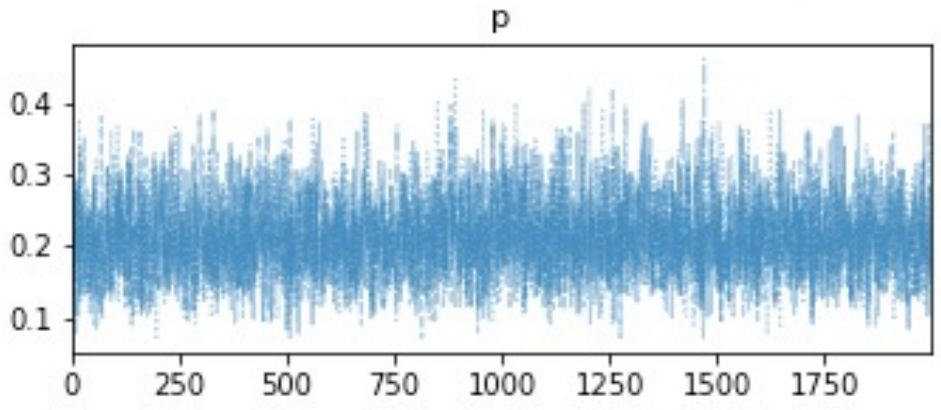
MCMC (Stan) の実行結果の見方

- MCMCの繰り返し回数：iterやstep、drawなどという
- 事後平均　事後標準偏差　信用区間　　有効標本サイズ　ESS　$\hat{R}$

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
p	0.211	0.054	0.112	0.31	0.001	0.001	2985.0	3914.0	1.0



Chainごとのpの分布



Chainごとのトレースプレット

$\hat{R}$ = Gelman & Rubin の指標：1.0はピッタリー致を意味する。1.1より小さければ十分に収束していると判断できる。ただし、Chainsごとの分散が高いときや、分布が裾野が広いときは$\hat{R}$は当てにならない。
ess = 有効標本サイズ。bulk は分布の大部分、tail は分布の両橋でのサンプリング効果

results にある lp__ は対数事後分布

Bayesian Data Analysis P.281 ～、11.4節 – 11.5節参照
http://www.stat.columbia.edu/~gelman/book/BDA3.pdf

Stanのブロックの種類

1-7の順で書く必要あり。

1. function {}:
   - 独自関数を定義するときに使う。
2. data {}
   - モデルに必要なデータやハイパーパラメータを宣言
3. transformed data{}
   - データの中で宣言したもの以外の処理をしたい場合に使う。
4. parameters {}
   - 指定するパラメータを宣言する。
     - 変数の型
       - int, real, vector, matrix, simplex, cov_matrix, …
     - 上限と下限を指定可能
       - lower = 下限、upperで上限
5. transformed parameters{}
   - パラメータの中で宣言以外の処理をしたい場合に使う。
6. model{} – 必須
   - 確率モデルを指定
     - normal, binominal, poison, gamma, beta, lognormal…
7. generated quantities{}
   - 各サンプリングで得られたパラメータ毎の計算をしたい場合に使う。

Stan のデータ型

- int: 整数型
  - int N;
  - array[N] int y;
    - 変数aに実数の要素数がNの配列を宣言
- real: 実数型
  - real<lower=0,upper=1> p;
    - 最小値0、最大値1の実数(他の型でも制約はつけることができる)
  - array[N] real<lower=0,upper=1> p;
- vector[N]
  - N次元ベクトル(要素は実数)
- simplex[N]
  - N次元ベクトルで総和が1
- matrix[N,M]
  - N行M列の行列(要素は実数)
- cov_matrix[M]
  - M行M列の分散共分散行列

事後分布を解析的に求めるとは

- ベイズの定理

尤度関数　事前確率：パラメータがとるであろう確率密度関数

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \qquad p(D) = \int p(D|\theta)p(\theta)d\theta$$

事後確率

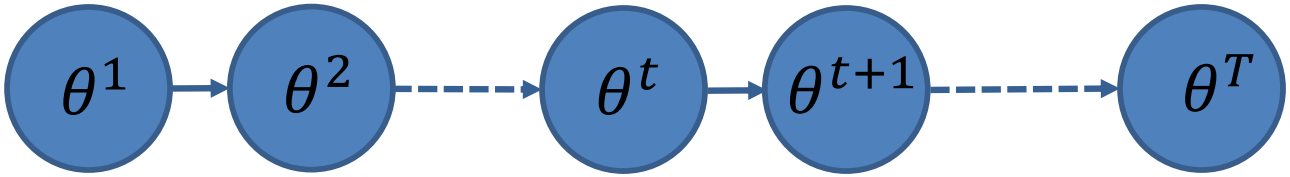
事後確率密度分布　周辺尤度：データが得られる確率（規格化定数）

$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

- ベイズ推定：パラメータθを確率変数とみなし、パラメータの値の確信度を事後確率密度分布を用いて表現する。
- 事後分布の積分を解析的に得くことは難しい。
  - 事後分布の計算の式展開や、積分計算が大変。
  - 共役事前分布を用いると、事前分布と事後分布が関数形にできる。
    - 利点：事後分布の計算が簡単に。
    - 欠点：決まった分布の形にしかならない。
  - 複雑な形の事前分布や、多くのパラメータを持つモデルを設定する場合は、解析的に求めることはできなくなる。

確率的サンプリングとマルコフ連鎖

- パラメータ θ の定常分布 $\pi(\theta)$ から、独立にサンプル $\theta^1, \theta^2, \ldots, \theta^T$ を生成(取り出)し、積分や期待値の計算を近似したい。
- しかし、事後確率分布から独立にサンプルを取り出すことはほとんど不可能である。
- そこで、 θ^t 自体は独立である必要がない、ことに着目した。
- 時刻 t $(1 \leq t \leq T-1)$のサンプル θ^tに依存する分布$\mathrm{P}(\theta^{t+1}|\theta^t)$から、次のサンプル θ^{t+1} を生成し、パラメータの列(マルコフ連鎖) $\theta^1, \theta^2, \ldots, \theta^T$を得ることを考える。
  - t以前のサンプリング結果には依存しない。

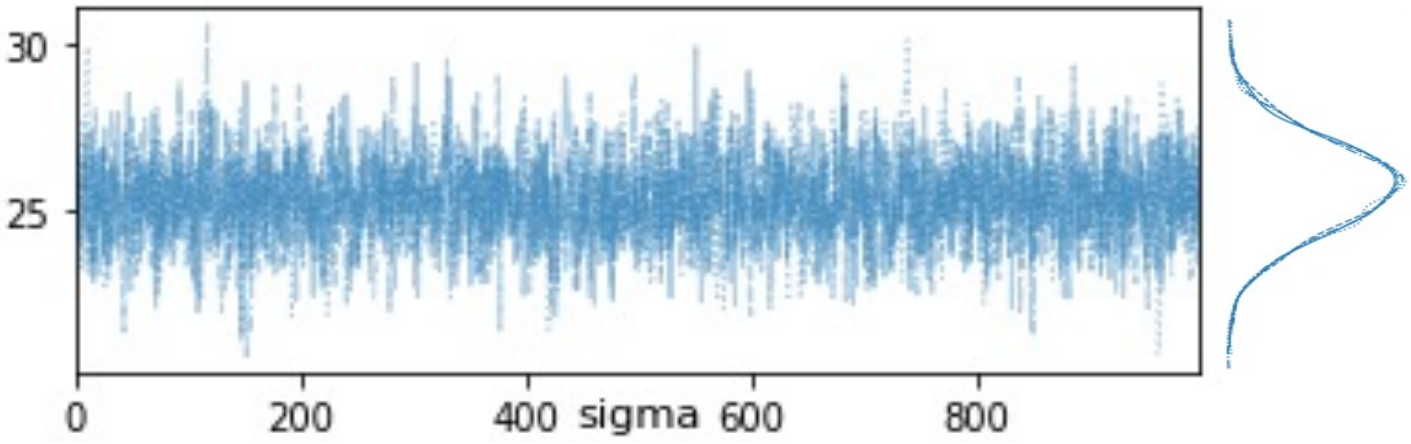


$$\mathrm{P}(\theta^T|\theta^{T-1}, \ldots, \theta^{t+1}, \theta^t, \ldots, \theta^1) = \mathrm{P}(\theta^T|\theta^{T-1}) \cdots \mathrm{P}(\theta^{t+1}|\theta^t) \cdots \mathrm{P}(\theta^2|\theta^1) \cdot P(\theta^1)$$

- マルコフ連鎖は単一の定常分布 $\pi(\theta)$ に収束することが証明されている（証明略）。
- 時刻 t が増えるにつれて定常分布からサンプリングされるようになり、初期状態には依存しない。

マルコフ連鎖とMCMC

- MCMC=マルコフ連鎖モンテカルロ法
  （Markov chain Monte Carlo methods：MCMC）
  - データに合った事後分布の近似確率分布をマルコフ連鎖を用いて生成する(サンプリングする)シミュレーション
  - マルコフ連鎖で得られる定常分布 $\pi(\theta)$ が事後確率分布になる操作
  - サンプリングしたい事後分布に対し、それを定常分布 $\pi(\theta)$ とするマルコフ連鎖を構成する方法



MCMCの試行回数

尤度に比例する確率分布のセットを得る。
平均や95%区間がわかる。

メトロポリス法によるベイズ推定の概要

- 初期パラメータ θ^0 を与える, $t = 0$
- 指定された回数Tだけ以下を繰り返す $t = 1, 2, \ldots, T$:
  - 提案分布と呼ばれる確率密度関数 $J_t(\theta^*|\theta^{t-1})$から新しいパラメータ θ^* を得る
  - 採択確率を計算 $r = P(\theta^*|D)/P(\theta^{t-1}|D)$
    - 比をとるので、規格化定数が不要
  - 採択確率 r が1以下ならば、$\min(r, 1)$ の確率で $\theta^t = \theta^*$とし、そうでないならば、 $\theta^t = \theta^{t-1}$とする。
- バーン・イン期間の計算をすて、パラメータ分布を得る。

- 提案分布は「詳細釣り合い条件」を満たせばどんな形でもOK (証明略)
  - 詳細釣り合い条件：分布の対称性、例えば、$\Delta\theta$と$-\Delta\theta$が同じ確率で現れる条件を意味する。
  - 例えば：適当なcを選び、一様乱数$U[c*-1, c*1]$からランダムに選択した $\Delta\theta \sim U[c*-1, c*1]$を使い、$\theta^* = \theta^{t-1} + \Delta\theta$とする。

MCMCによる正規分布のベイズ推定のコード例
http://web.sfc.keio.ac.jp/~maunz/BS19/BS19_R04.html

Stan における事前分布の設定について

- 選択肢1:何も指定しない→**無情報一様分布**
  - −∞～∞の一様分布
  - サンプルサイズが小さい場合は不適切になることも。
- 選択肢2:パラメータに合った弱情報事前分布
  - 確率，比率：ベータ分布か一様分布
    - ~ beta(1, 1); = ~ uniform(0,1);
  - 平均値，回帰係数: 正規分布(orコーシー分布)
    - ~ normal(0, 100); SDはパラメータのスケールに合わせる
  - 標準偏差，分散: 半コーシー分布
    - ~ cauchy(0, 5); 加えて，パラメータの下限を0にしておく

ベイズ推論の分布

- 事前分布 prior distribution p(θ)
- 事前予測分布 prior predictive distribution
  - 事前分布からどんな観測データが得られると予測できるか。
- 事後分布 posterior distribution p(θ|X)
- 事後予測分布 posterior predictive distribution
  - 事後分布から次にどんな観測データが得られると予測できるか。

- 信用区間(CI: Credible Interval)
  - 100(1-α)%信用区間：$P(a \leq \theta \leq b) = 1 - \alpha$を満たす区間 $[a, b]$
  - 95%信用区間:信用区間に95%の確率でパラメータθが存在する区間