Pythonで体験するベイズ推論 ~ PyMCによるMCMC入門 ~

第6.5 ~ 6.9章　事前分布をはっきりさせよう

2017/12/19 (火) 教科書輪講

石田研究室 修士一年 大石 智博

目次

目次

主観的な事前分布を使う (ドメイン知識を使う) 利点

1. MCMCの収束スピード向上

   未知パラメータが正であると事前に知っていれば負値を探索せずにすみ計算時間減少

2. 推論の精度向上

   真のパラメータ値付近の重みを上げれば、推論結果はその付近に集まる

3. 不確実さをよく表現できる

   予測の確実性もパラメータで表現可能

分野の専門家から事前分布を引き出す

ベイズ手法を適用するエンジニアは、ドメイン毎に専門家から話を聞き、事前分布をつくる必要有

著者曰く、以下を注意すべき（よくないこと）

- ベイズを知らない人にベータ、ガンマとか言う

統計が専門でない人は連続確率密度関数が１を超えることがあると聞くとひっくり返る(笑)

- ロングテールの稀なイベントを無視して、分布の平均付近に大きすぎる重みを置く
- 推論結果の不確実さを常に過小評価する

どうやって技術者ではないエキスパートから事前知識を引き出すかについては慎重になるべき

ルーレット法 (Trial roulette method)

カウンタ※を置くだけで、どんな値を取りやすいと専門家が考えているかを引き出し、事前分布を構築できる手法

※カウンタとは、カジノのチップみたいなもの

1. 専門家はN個のカウンタを持っている
2. グリッドのマス目にカウンタを置いていく
3. マス目に積まれたカウンタの数で事前分布がつくれる

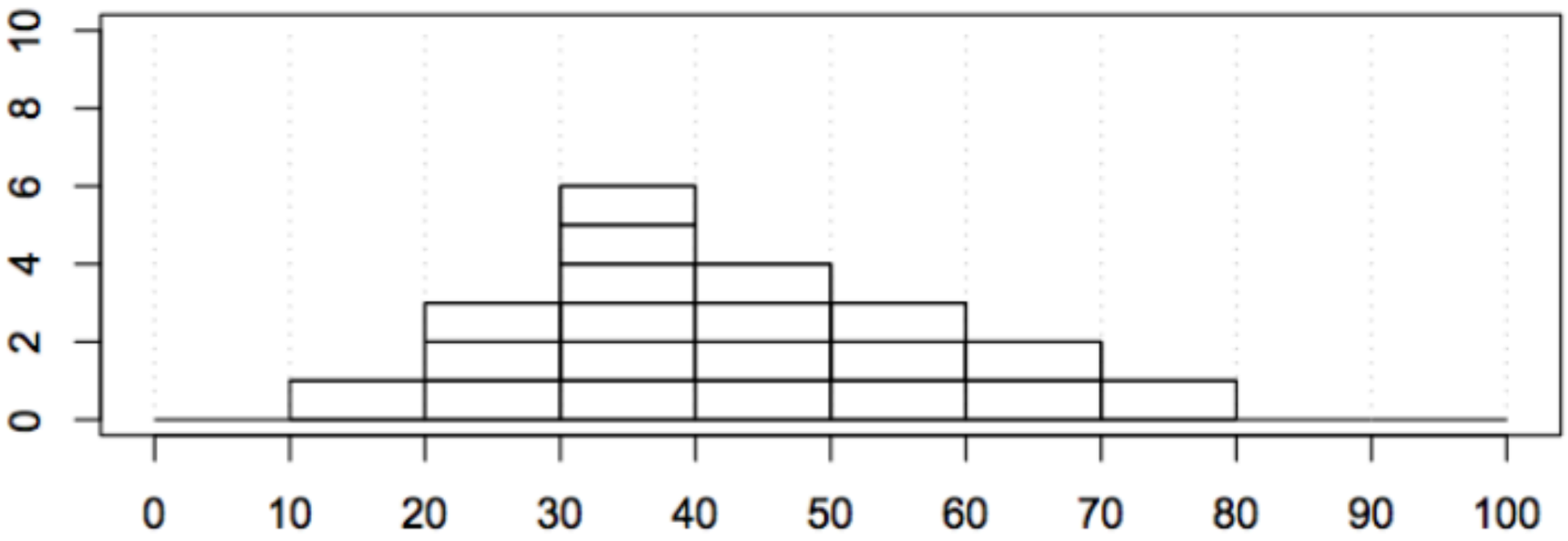
ルーレットみたくどこにベットするかで事前確率を生成

ルーレット法の適用例

学生が将来のテストの点数を予測

カウンタの数 N = 20として、各グリッドにカウンタを置いていく

例えば、60点以上の確率は 3 / 20 = 0.15である



学生がグリッド毎にカウンタを置いた結果[1]

[1] Oakley, Jeremy E. "Eliciting Probability Distributions." (2010).

ルーレット法のメリット

- 長々質問しなくても、統計学者はカウンタの数を数えるだけで、専門家の主観的な事前分布の形状を理解できる

- 事前分布をつくる過程で、専門家は最初においたカウンタを移動させても良いため、専門家が満足した結果にたどり着ける

- 得られるものが確率分布になることが保証される
  - カウンタが全て使われれば、確率は足して１になる

- 相手の統計知識があまりない場合、視覚的な方法のため結果が正確になる

例題：株売買の収益

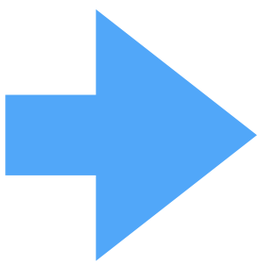
株を買う時、アナリストは株の日次リターン(日次収益率)に注目することが多い

S(t)をt日目の株価とすると、t日目の日次リターンr(t)は以下で表される

$$r(t) = \frac{S(t) - S(t-1)}{S(t)}$$

期待日次リターン、 $\mu = E[r(t)]$ が大きい株を買いたい

- 日次リターンはノイズだらけでパラメータの推定は困難
- 時間経過によってパラメータは変化するため膨大なデータを使うのはあまり賢くない
- 小さいデータセットのサンプル平均は間違ってる可能性が高い



推定された値と不確実さを見ることができる
ベイズ推論が手法として適切

例題：株売買の収益

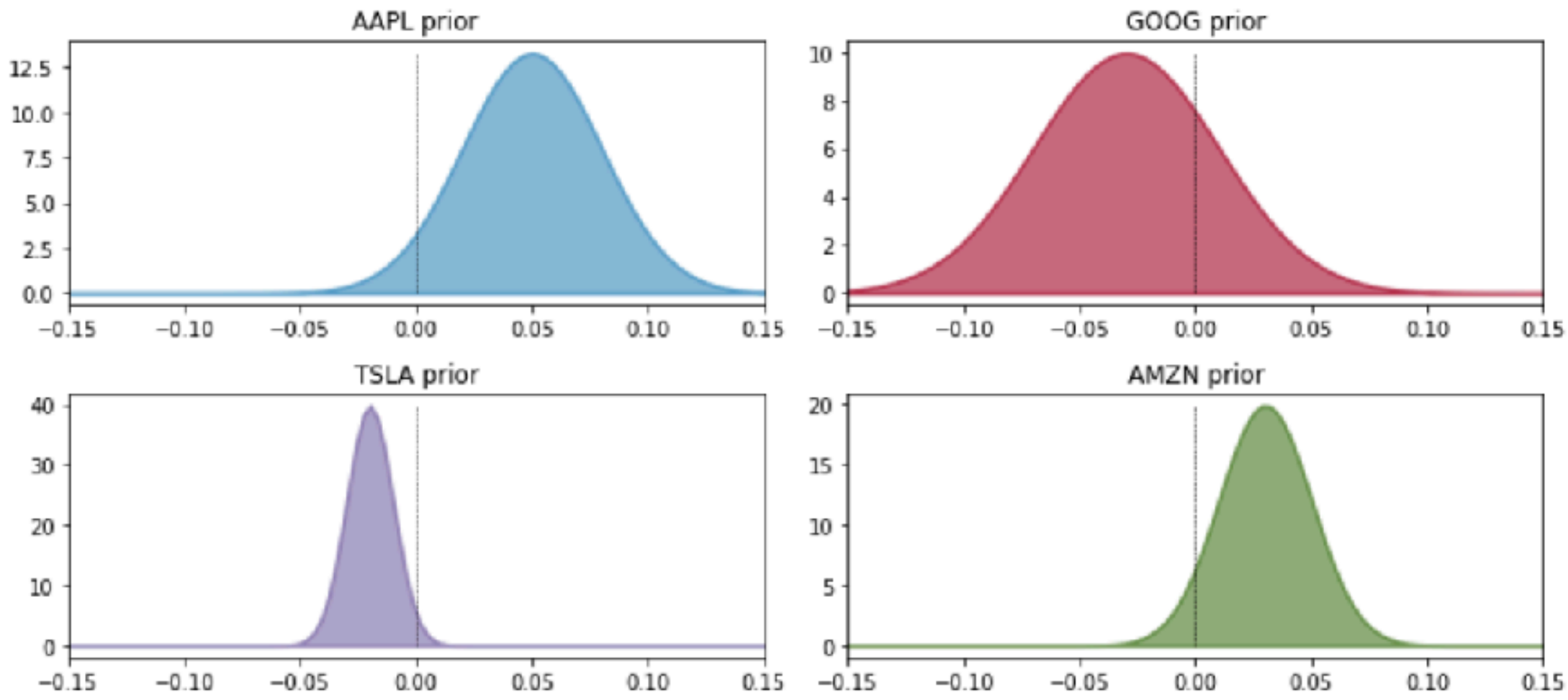
アップル、グーグル、テスラモーター、アマゾンの４銘柄の日次リターンの特性について考える

1. ルーレット法を適用して、ファンドマネジャー (金融のプロ)の考えるリターン特性を調べる
2. それが以下のような値を持つ正規分布に見えて、正規分布を当てはめることにする

	AAPL	AMZN	GOOG	TSLA
μ	0.05	0.03	-0.03	-0.02
σ	0.03	0.02	0.04	0.01

3. それぞれの銘柄の事前分布をつくろう

銘柄ごとのリターンの事前分布取得



これらはファンドマネージャーによる主観的な事前分布

リターンのモデリングの精度を上げよう

共分散行列を考えてみる

非常に相関の高い２つの銘柄はおそらく一緒に値下がりするので、同時に投資するのはあまり賢くない（分散投資すべき）

そこで共分散を考えて相関を調べる

これをモデリングするためにウィシャート分布を使う

銘柄毎の過去のデータを扱いたい

Yahoo Finance URL not working

42

I have been using the following URL to fetch historical data from yahoo finance for quite some time now but it stopped working as of yesterday.

https://ichart.finance.yahoo.com/table.csv?s=SPY

28

When browsing to this site it says:

> Will be right back...
>
> Thank you for your patience.
>
> Our engineers are working quickly to resolve the issue.

However, since this issue is still existing since yesterday I am starting to think that they discontinued this service?

My SO search only pointed me to this topic, which was related to https though...

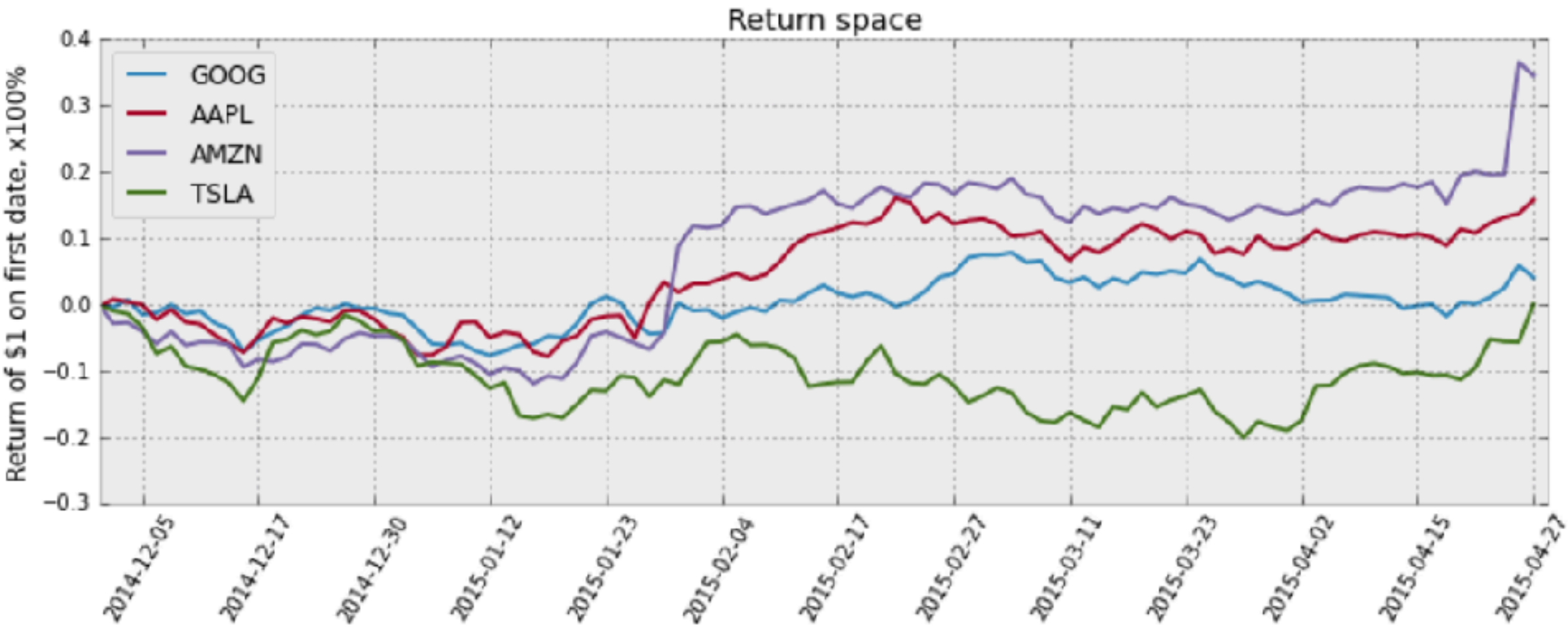
Is anyone else experiencing this issue? How can I resolve this problem? Do they offer a different access to their historical data?

教科書通りだとichartsのサービスが終了していて、取得できない

以降、取得した結果のみ示す

ソースコードはCh6_Priors_PyMC3.ipynb参照

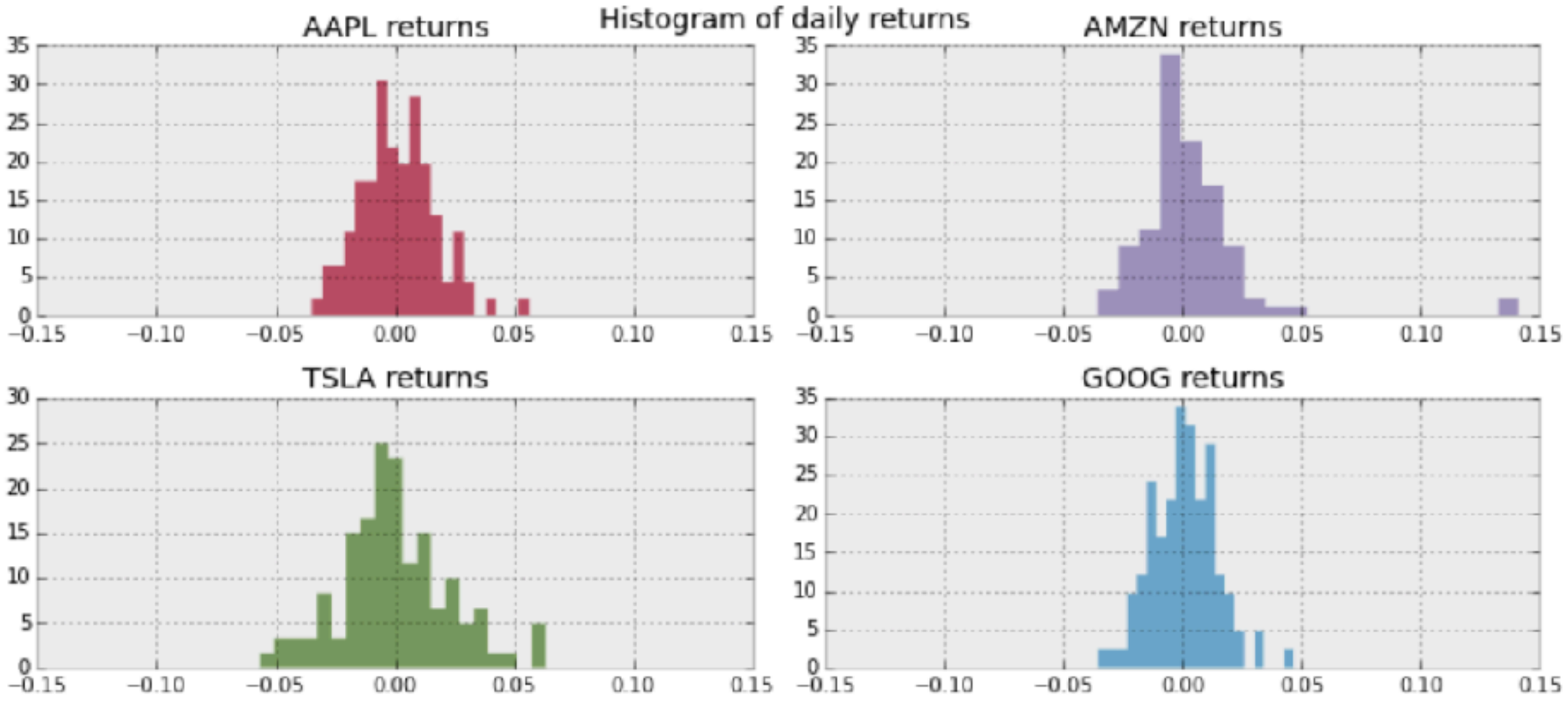
各銘柄の日次リターン（収益率）



2012/9/1 ~ 2015/4/27 までの銘柄ごとの日次リターン

各銘柄の日次リターンの履歴

縦軸：密度

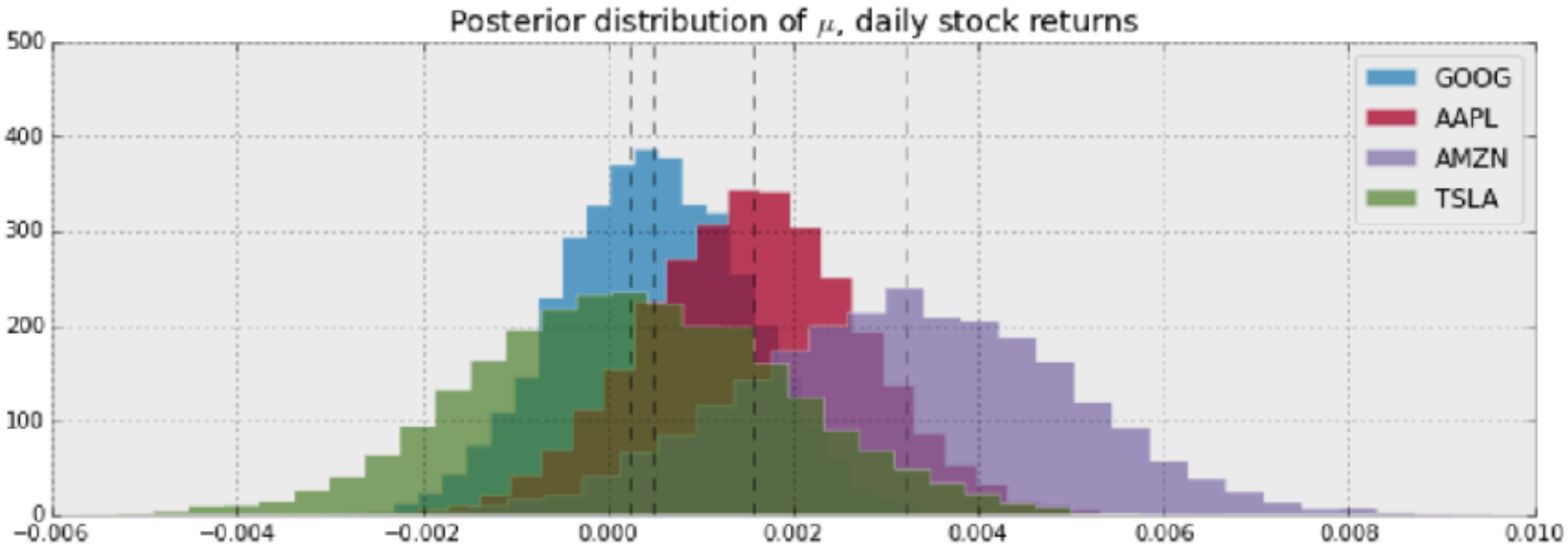


横軸：値

銘柄ごとの日次リターンのヒストグラム

日次リターンの事後分布

縦軸：
密度

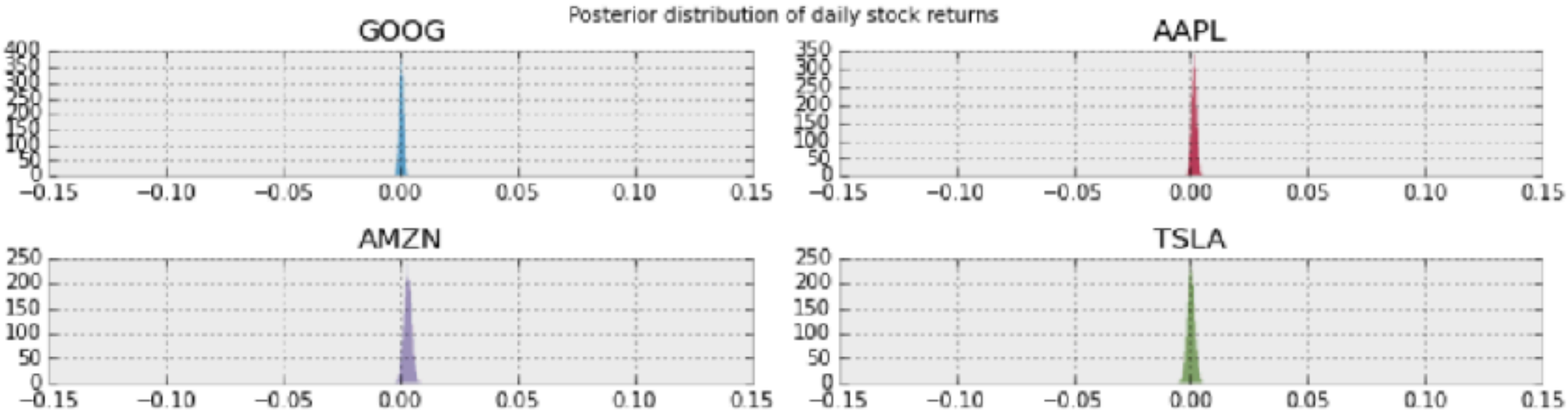


横軸：値

アマゾンが日次リターンの平均が高い

日次リターンの事後分布

縦軸：
密度

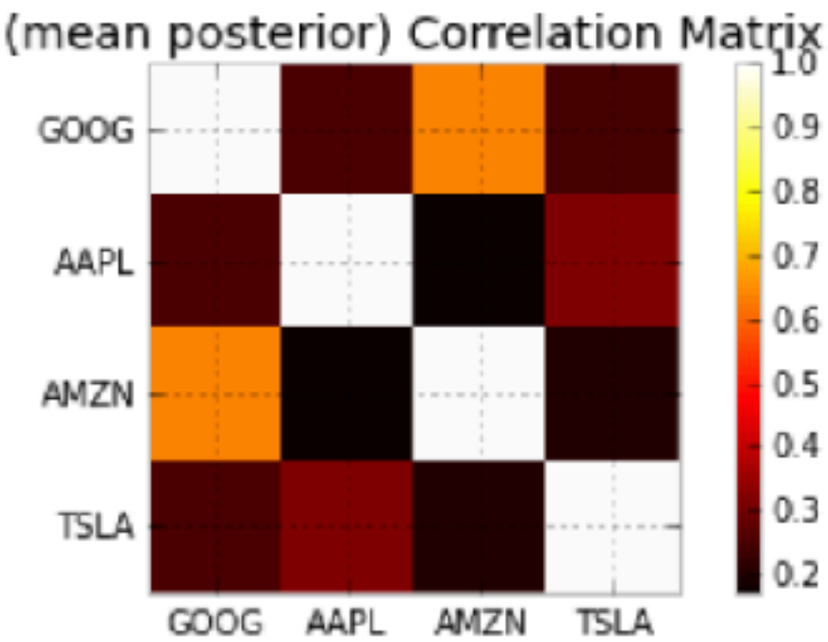


横軸：値

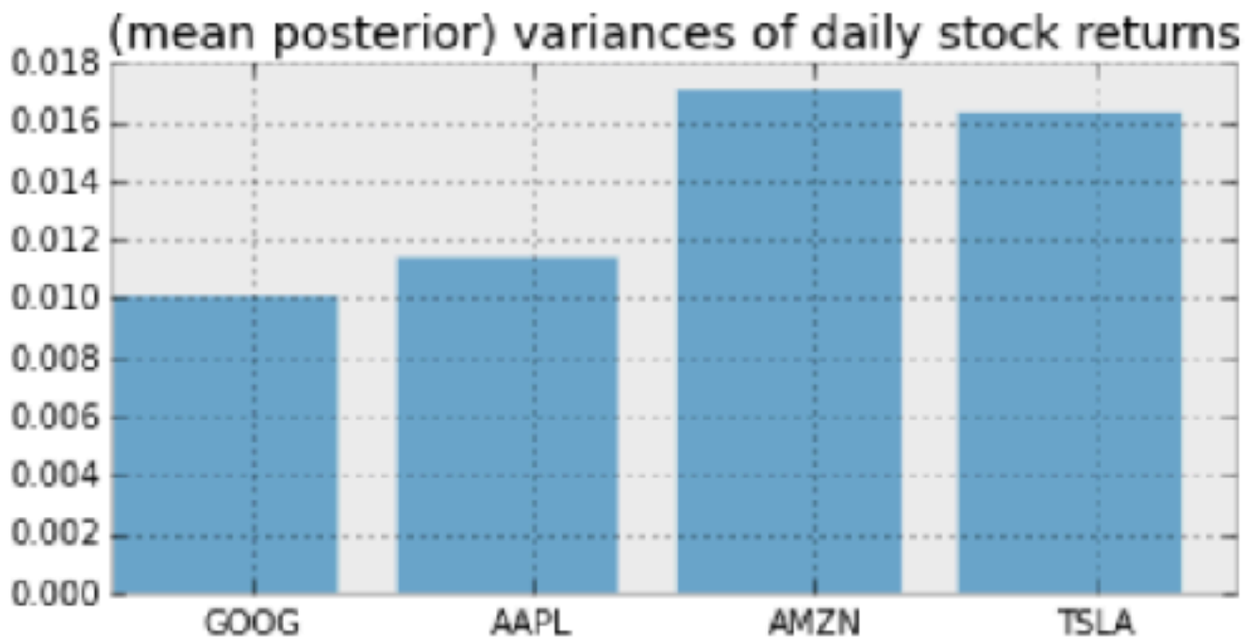
事前分布と同じ横軸のスケールで表示している

事後分布の分散は事前分布より遥かに小さくなっている

各銘柄の相関と分散の事後平均



相関行列の事後平均



日次リターンの分散の事後平均

共分散行列を正規化し、相関行列をとる

GOOGとAMZNが相関が比較的高い

目次

共役事前分布

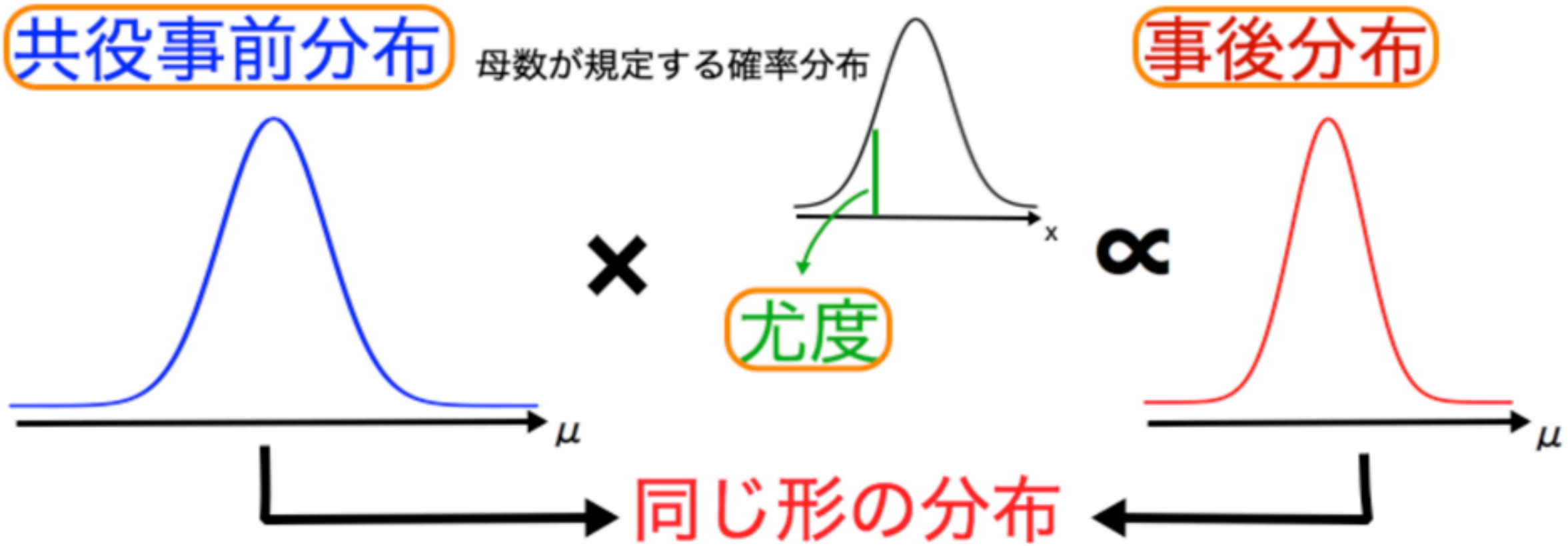
- 共役事前分布とは、尤度をかけて事後分布を求めるとその関数形が同じになるような事前分布のこと
- 事後分布の形が事前分布と同じだとMCMCによる近似推論などを必要とせず事後分布を直接計算可能（計算が楽）

問題点

1. 共役事前分布は客観的でないため、主観的な事前分布が使われるときだけ有用
2. 共役事前分布が存在するのは一般に単純で一次元の問題に限られる。
   - 問題が大きくなり複雑になると共役事前分布は存在しない

（筆者の個人的見解：数学的に便利なだけで問題についての洞察は得られない）

共役事前分布

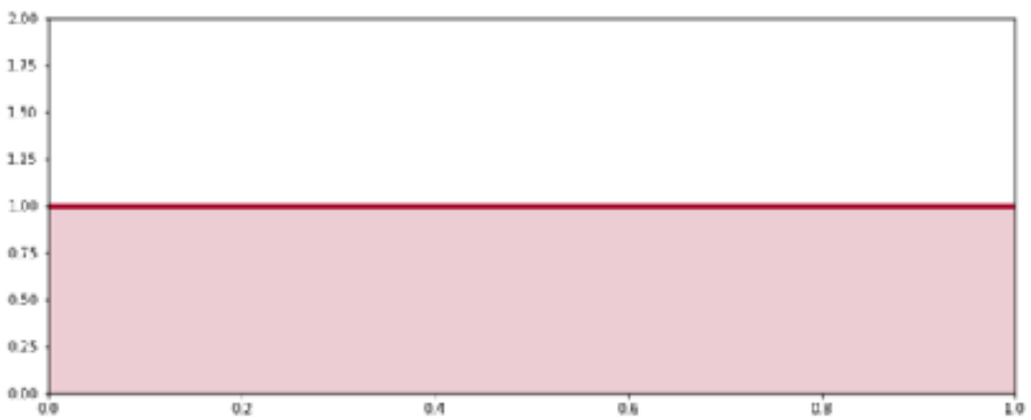


共役事前分布	母数が規定する確率分布	事後分布
ベータ分布	ベルヌーイ分布	ベータ分布
ベータ分布	二項分布	ベータ分布
正規分布	正規分布（σ^2既知）	正規分布
逆ガンマ分布	正規分布（σ^2未知）	逆ガンマ分布
ガンマ分布	ポアソン分布	ガンマ分布
ディリクレ分布	多項分布	ディリクレ分布

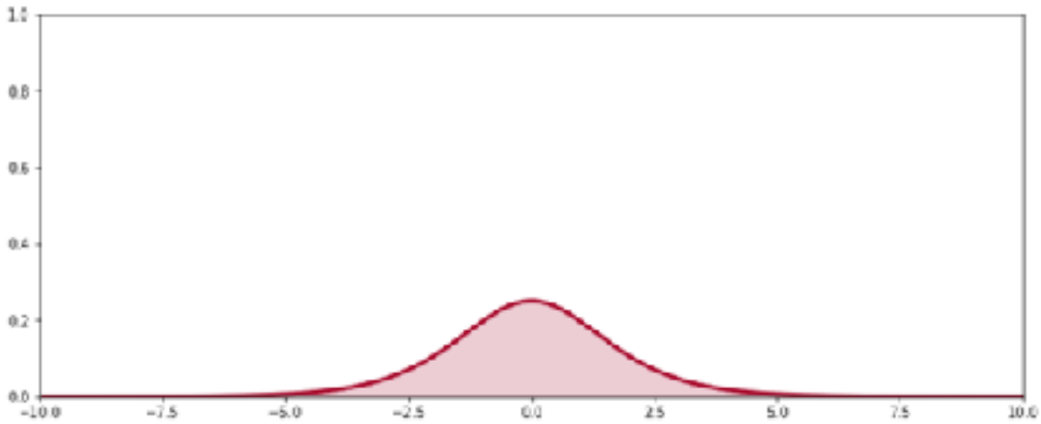
目次

Jeffreys 事前分布

- すべての値に等しく確率を割り当てる一様事前分布は、事後分布に偏りを生じさせない選択に思える
- しかし、一様事前分布は変換に対して不変ではない



θを変換



$$\varphi = log(\frac{\theta}{1-\theta})$$ (これは単にθを引き伸ばすだけの変換)

- 一様分布だったものが情報を含んでしまっている！
- こういった思いがけず情報のある事前分布ができてしまうのを防ぐ方法としてJeffreys事前分布がある

Jeffreys 事前分布の定義

（本書では全く解説がないため定義だけ載せときます）

Fisher情報行列 $I(w) = I_{ij}(w)$ を次で定義する。

$$I_{ij}(w) = \int \frac{\partial f(x,w)}{\partial w_i} \frac{\partial f(x,w)}{\partial w_j} p(x|w) dx$$

ただし $f(x,w)$ は対数密度比関数

$$f(x,w) = \log(q(x)/p(x|w))$$

である。

W 上の Jeffreys事前分布とは、次で定義される $\mathbb{R}^d$ 上の確率密度関数 $\varphi(w)$ のことをいう[1]。

$$\varphi(w) = \begin{cases} \frac{1}{Z}\sqrt{\det I(w)} & (w \in W) \\ 0 & \text{otherwise} \end{cases}$$

目次

Nが大きくなった時の事前分布の影響

- 観測データが増えるにつれて、事前分布は重要でなくなっていく

  - 事前分布はこれまでの情報に左右されるため、新しいデータが十分にあると、その価値は低くなるのが直感的に理解できるはず

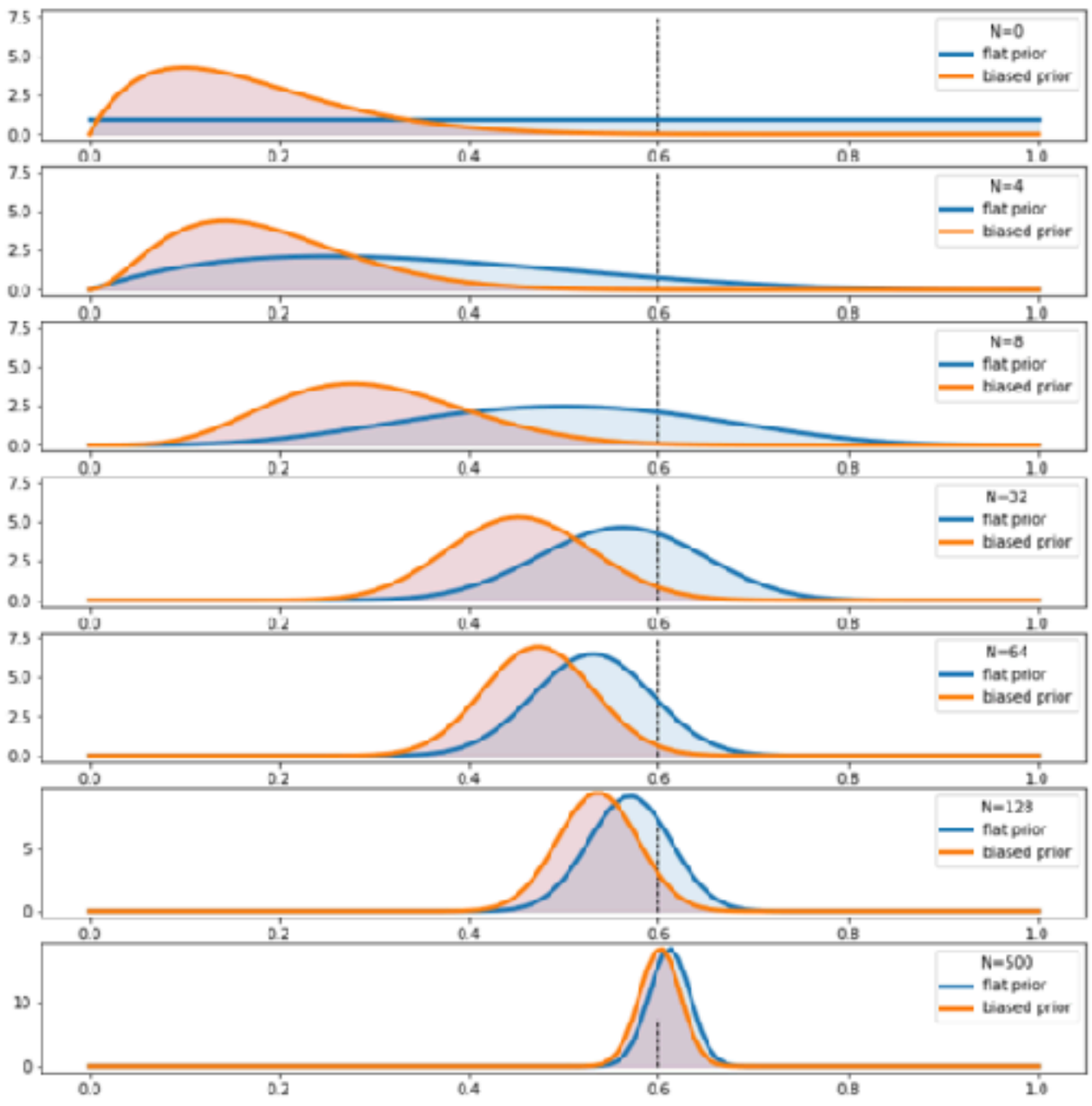
- 事前分布の影響を消してしまうほどのデータ量があるのが望ましい

  - 事前分布が致命的に間違っていても、データがそれを修正してより間違いの少ない事後分布が得られるため

事前分布が忘れられる例

サンプルサイズが増えることによって事前分布の影響が小さくなる

つまり、選んだ事前分布によらず、推論結果は同じものになる



全ての事後分布がこれほど早く事前分布を忘れるわけではない

最終的に事前分布が忘れられることを例として示している

おわりに

- 本章では事前分布の使い方を再確認した
- 事前分布はモデルに組み込まれる要素であり、慎重に選ばなければならない
- 事前分布にはメリット、デメリットがある
    - 事前分布のメリット：どんなデータに対しても柔軟にモデルを設計できる点
    - 事前分布のデメリット：主観と意見に左右されてしまう点

- 主観的な事前分布に関する論文は数百本ある

本書が適切な事前分布を選ぶ手助けとなることを願っている