事後キュムラントを利用したブートストラップ法の近似

伊庭　幸人　　モデリング研究系　教授

ベイズ事後分布からの MCMCサンプルに豊富な頻度論的な情報が含まれていて，事後共分散や事後キュムラントの形で取り出せることが認識されてきた．この関係はデータの摂動に対する感受率の事後キュムラント表現に基づくもので，モデルが正しくなくても成立する．

　事後分布からのサンプルを用いて事後期待値の交差検証(CV)やブートストラップを行う方法としてはインポータンス・サンプリング法（IS）があるが，高次元の事後分布では重みの不均一によって計算が不安定になりがちで，それを回避できるのが事後キュムラントを用いる方法の利点である．

ベイズ感受率公式（local case sensitivity formula):　Millar and Stewart (2007), Perez et al (2006)

$$\frac{\partial}{\partial \epsilon_i}\mathbf{E}^{\epsilon}_{\mathrm{pos}}[A(\theta)]|_{\epsilon=0} = \mathbf{Cov}_{\mathrm{pos}}[A(\theta), \log p(y_i|\theta)]$$

$\mathbf{E}_{\mathrm{pos}}[\ ]$, $\mathbf{Var}_{\mathrm{pos}}[\ ]$, $\mathbf{Cov}_{\mathrm{pos}}[\ ]$: 事後期待値，事後分散 / 共分散

- 観測値を摂動したときの影響の表現として提案された
- 頻度論とベイズを結ぶ橋として機能（下の例）
- 統計物理でいう線形応答公式の一種（観測値の重み=外場）

$$\log p(y|\theta) = \sum_{i=1}^{N} \log p(y_i|\theta)$$

公式の証明は下の式を
積分記号のもとで微分するだけ

$$\mathbf{E}^{\epsilon}_{\mathrm{pos}}[A(\theta)] = \frac{\int A(\theta)\exp(\sum_m \epsilon_m \log p(y_m|\theta) + \log p(y|\theta) + \log p(\theta))\, d\theta}{\int \exp(\sum_m \epsilon_m \log p(y_m|\theta) + \log p(y|\theta) + \log p(\theta))\, d\theta}$$

2階微分と3次の事後キュムラントの関係に拡張できる

$$\frac{\partial^2}{\partial \epsilon_i \partial \epsilon_j}\mathbf{E}^{\epsilon}_{\mathrm{pos}}[A(\theta)]|_{\epsilon=0} = \boldsymbol{\kappa}_{\mathrm{pos}}\left[A(\theta), \log p(y_i|\theta), \log p(y_j|\theta)\right]$$

$$\boldsymbol{\kappa}_{\mathrm{pos}}[A, B, C] = \mathbf{E}_{\mathrm{pos}}\left[\left(A - \mathbf{E}_{\mathrm{pos}}[A]\right)\left(B - \mathbf{E}_{\mathrm{pos}}[B]\right)\left(C - \mathbf{E}_{\mathrm{pos}}[C]\right)\right]$$

: 3次の事後キュムラント

事後キュムラントを使った近似的ブートストラップ法：　提案手法

MCMCで計算した期待値をブートストラップしたい！→ 復元抽出ごとにMCMCを走らすのは大変

＊もとのデータのMCMCサンプルで以下を計算　1回だけ

$\mathbf{Cov}_{\mathrm{pos}}[A(\theta), \log p(y_i|\theta)]$

$\boldsymbol{\kappa}_{\mathrm{pos}}[A(\theta), \log p(y_i|\theta), \log p(y_j|\theta)]$

↓　ここだけ繰り返す

＊多項分布で観測の重みw_iを選ぶ
＊$\epsilon_i = w_i - 1$
＊ϵ_iについてのテーラー展開で復元抽出を近似

復元抽出の近似（2次の場合）

$w_i \sim multinomial\left(\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n}\right), \epsilon_i = w_i - 1, \quad i = 1,2,..,n$

$$\mathbf{E}_{\mathrm{pos}}[A(\theta)] + \mathcal{B}_1 + \mathcal{B}_2$$

$$\mathcal{B}_1 = \sum_i \epsilon_i\, \mathbf{Cov}_{\mathrm{pos}}[A(\theta), \log p(y_i|\theta)]$$

$$\mathcal{B}_2 = \frac{1}{2}\sum_{ij} \epsilon_i \epsilon_j \boldsymbol{\kappa}_{\mathrm{pos}}[A(\theta), \log p(y_i|\theta), \log p(y_j|\theta)]$$

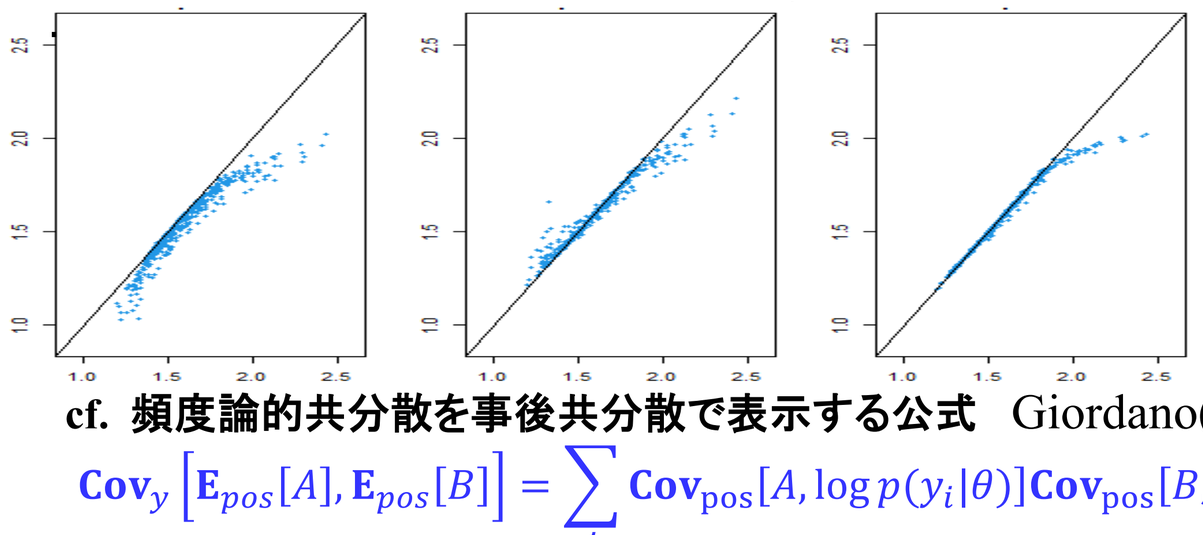
1次では$\mathcal{B}_1$のみを利用

Efron (2015)
十分統計量の分散が利用できる場合に頻度論的分散を事後共分散であらわす公式を提案．ABC近似の応用も提案

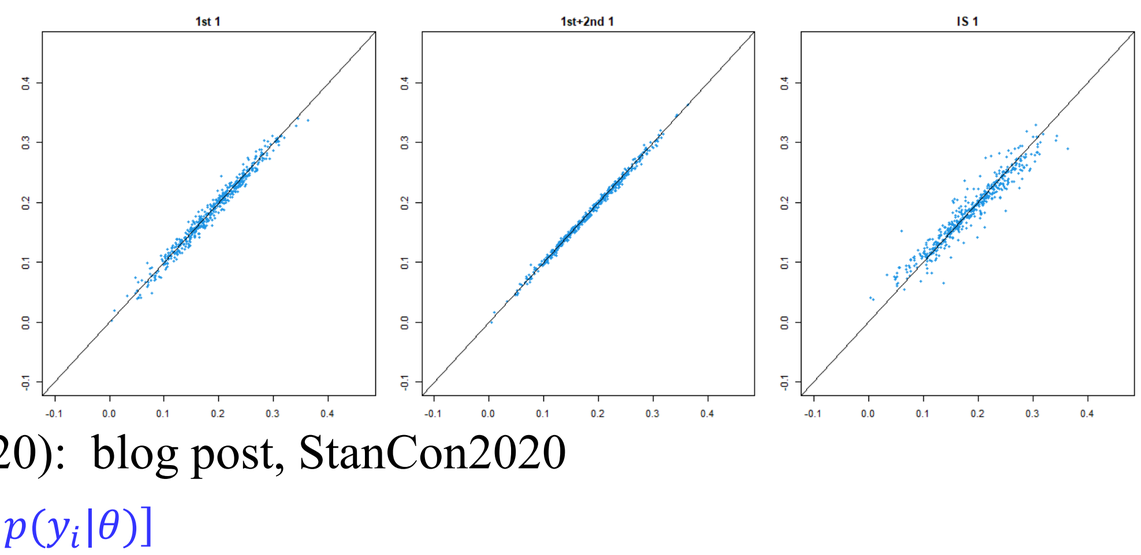
Lee et.al（2017)
インポータンス・サンプリングにもとづくブートストラップの近似を提案：**今回の比較対象（IS)**

Giordano (2020)
頻度論的分散を1次のlocal case sensitivityで表現する公式を提案(2020)　：**下記参照**

例1：古代エジプト人女性の寿命データへのワイブル分布のあてはめ
ワイブル分布の形状パラメータの事後期待値をブートストラップ
パラメータ数2，左から１次，２次の提案手法，最も右がIS
・各点が1組の復元抽出に対応
・横軸が各抽出ごとにMCMCを走らせた結果，縦軸が各近似手法の結果



例2：糖尿病データのリッジ回帰: Efron(2015)の例，ただしL1でなくL2
125人目の患者の予測値をブートストラップ
パラメータ数12，左から１次，２次の提案手法，最も右がIS
・各点が1組の復元抽出に対応．横軸・縦軸は例1と同様．



cf. 頻度論的共分散を事後共分散で表示する公式　Giordano(2020): blog post, StanCon2020

$$\mathbf{Cov}_y\left[\mathbf{E}_{pos}[A], \mathbf{E}_{pos}[B]\right] = \sum_i \mathbf{Cov}_{\mathrm{pos}}[A, \log p(y_i|\theta)]\mathbf{Cov}_{\mathrm{pos}}[B, \log p(y_i|\theta)]$$

競争的資金

- 科研費 基盤(C)　多様な予測に対応した情報量規準の開発：計算統計的アプローチ　　伊庭（代表）/矢野（分担）