データ駆動科学と近似アルゴリズム

東京大学　大学院新領域創成科学研究科
複雑理工学専攻
岡田真人

本スライドのまとめ

- データ駆動科学に用いるスパースモデリングやベイズ推論には、計算量の削減を目的とする近似アルゴリズムが多数提案されている。
- しかし、データ駆動科学では、これらの近似アルゴリズムを用いるのは適切ではない。
- 近似アルゴリズムを用いた際に起こる問題としては、複数ある近似アルゴリズムで解が異なる場合がある。そのような場合、どの解を信じるべきかの指針が存在しない。
- スパースモデリングに対しては、全状態探索を用いて、スパースモデリングの厳密解を求める。
- ベイズ推論おいては、モンテカルロ法を用いた数値的厳密解を求める。
- データ駆動科学では、近似アルゴリズムに関する知見は不必要であり、これがデータ駆動科学への参入を容易にしている。

スパースモデリングとは

スパースモデリング基本的な考え方は

(1)高次元データの説明変数が次元数よりも少ない(スパース(疎)である)と仮定し,

(2)説明変数の個数がなるべく小さくなることと,データへの適合とを同時に要請することにより,

(3)人手に頼らない自動的な説明変数の選択を可能にする枠組みである.

スパースモデリング(変数選択)に関する二つの戦略

原則：変数選択の問題の計算量は指数爆発する
(Cover and Van Campenhout, 1977)

＜変数選択に対する二つの戦略＞

1. 凸最適化や変分ベイズにもとづく緩和型アプローチ
   通常に用いられているスパース推定のアルゴリズム
   凸最適化：Lasso(1996, L1正則化)
   変分ベイズ：ARD(2008, 関連自動度決定)

2. 全状態探索またはモンテカルロ法により効率的にサンプリング.
   記述子の個数が20程度の場合，全記述子の組み合わせを評価

緩和型近似アルゴリズは沢山提案されていて、アルゴリズムごとに異なった解が得られる場合があるので、データ駆動科学では近似アルゴリズムを用いない。

スパースモデリングの具体例
線形回帰における定式化

目的変数 $y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_N x_N$ 説明変数
(機能)

$$= \boldsymbol{\beta}\mathbf{x}$$

サンプル数p, 説明変数の数 N

回帰係数 $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., 0, ..., 0, \beta_N)$

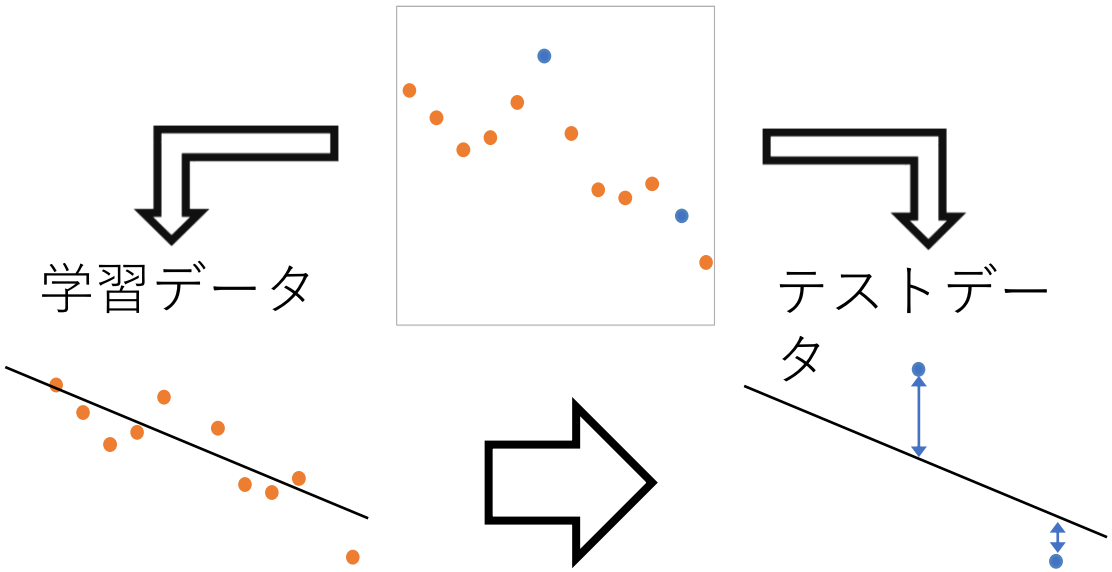
インディケータ $\mathbf{c} = (1, 1, ..., 0, ..., 0, 1)$

$\mathbf{c} = \{0,1\}^N$ $2^N - 1$ 通り

全状態探索(Exhaustive Search, ES) 法

Garside 1965, Ichikawa et al., 2014, Nagata et al., 2015, Igarashi et al., 2016, Igarashi et al., 2018

1. $\mathbf{c}$ を決めると記述子の係数 βが決まる
2. その$\mathbf{c}$に対する回帰モデル(β)の交差検証誤差(CVE)を求める
3. 各記述子の組み合わせ($\mathbf{c}$)を評価



交差検証によって,
限られたデータから
汎化誤差を推定する

線形回帰における全状態探索(ES-LiR)法

インディケータ $\boldsymbol{c}$	$\boldsymbol{\beta}$ を用いて $\mathbf{y}$ の推定	CVE
$c=(1,0,0)$	$\hat{y}=b_0+b_1x_1$	$error=0.20$
$c=(0,1,0)$	$\hat{y}=b_0+b_2x_2$	$error=0.21$
$c=(0,0,1)$	$\hat{y}=b_0+b_3x_3$	$error=0.23$
$c=(1,0,1)$	$\hat{y}=b_0+b_1x_1+b_3x_3$	$error=0.16$
$c=(1,1,0)$	$\hat{y}=b_0+b_1x_1+b_2x_2$	$error=0.14$
$c=(0,1,1)$	$\hat{y}=b_0+b_2x_2+b_3x_3$	$error=0.18$
$c=(1,1,1)$	$\hat{y}=b_0+b_1x_1+b_2x_2+b_3x_3$	$error=0.20$

線形回帰　　CVEの計算

$2^3-1=8$ 通り，すべてのCVEをチェックする

記述子の個数Nが十分に大きい場合，計算量$O(2^N)$が膨大になる

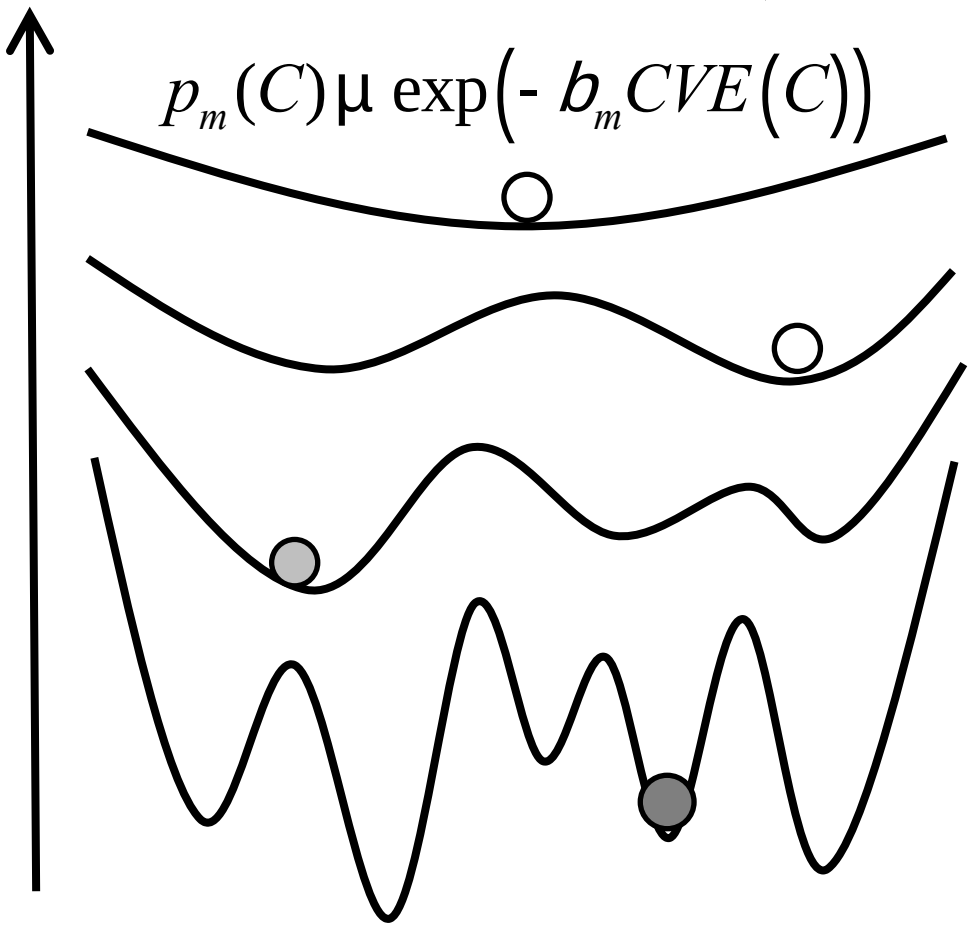
レプリカ交換法を用いた近似的全状態探索法

Similarity with random spin system (spin glass)

特徴量の次元が20個程度までであれば、厳密に全状態探索法を用いる

特徴量の次元がそれより多くなる場合は、レプリカ交換モンテカルロ法でサンプリングする

レプリカ交換モンテカルロ法
(Hukushima and Nemoto, 1996)



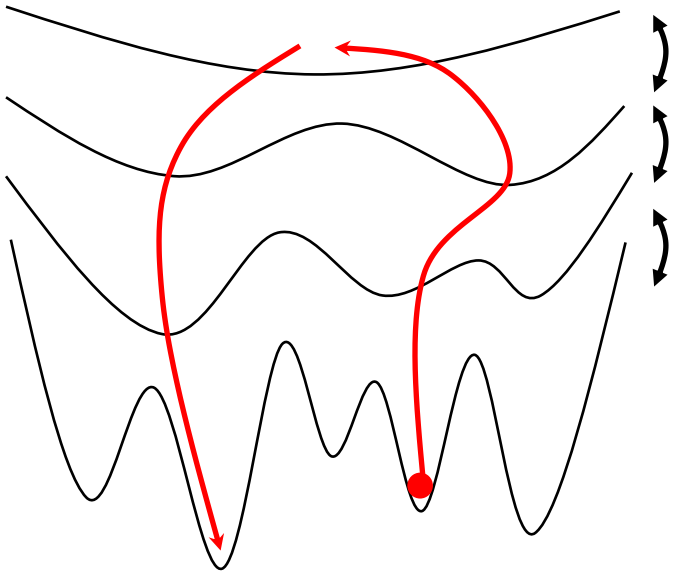
マルチヒストグラム法

Ferrenberg and Swendsen (1989), Hukushima (2002)

レプリカ交換モンテカルロ法によって、
全状態での状態密度を推定することも可能である.

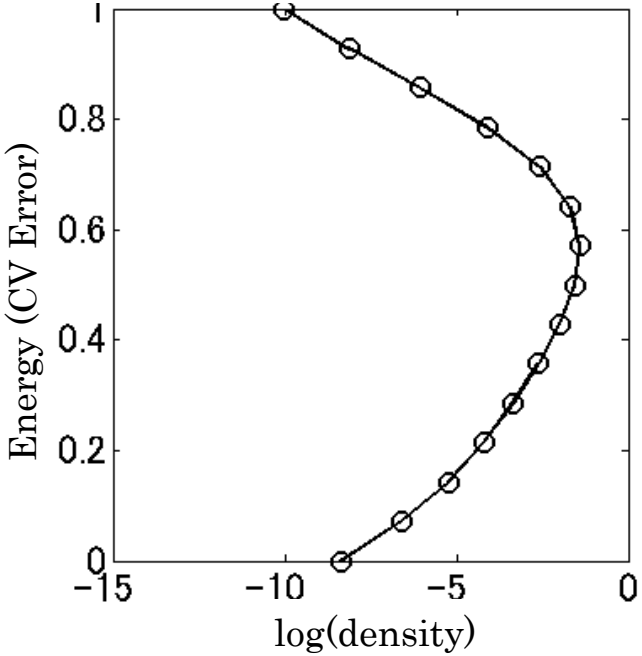
$$g(E) = \frac{\sum_{m=1}^{M} N_m(E)}{\sum_{m=1}^{M} n_m \mathrm{e}^{f_m - \beta_m E}}, \quad \Leftrightarrow \quad \mathrm{e}^{-f_m} = \sum_E g(E)\mathrm{e}^{-\beta_m E},$$

レプリカ交換モンテカルロ法



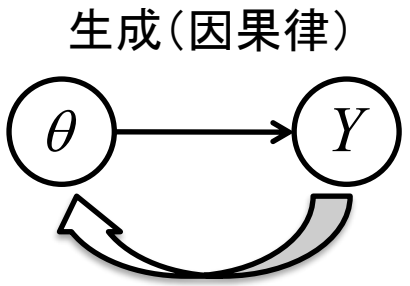
推定

CVEのヒストグラム



ベイズ推論：因果律を組み込んでデータ解析

$$p(Y,\theta) = p(Y|\theta)p(\theta) = p(\theta|Y)p(Y)$$



＜ベイズの定理＞

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)} \propto \exp\left(-nE(\theta)\right)p(\theta)$$

$p(\theta|Y)$ ：事後確率。データが与えられたもとでの，パラメータの確率.

$p(\theta)$ ：事前確率。あらかじめ設定しておく必要がある。
これまで蓄積されてきた科学的知見

ベイズ推論の近似アルゴリズム

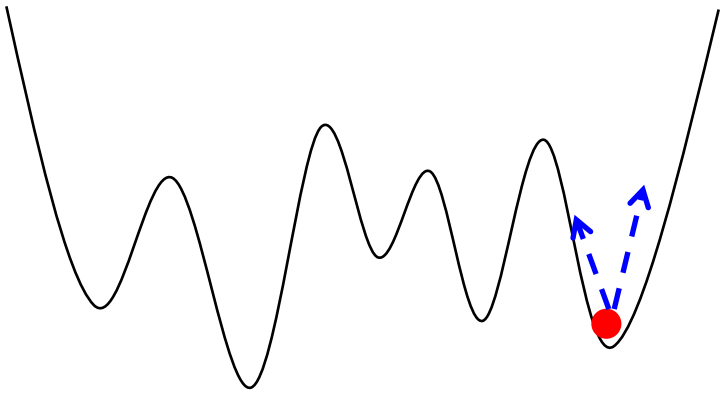
- ラプラス近似
  - 事後分布をガウス分布で近似する
- 変分ベイズ法
  - 事後分布の形を制限して、その範囲内で解を求める。

- 近似のために、間違った解が得られる。
- データ駆動科学には不適切

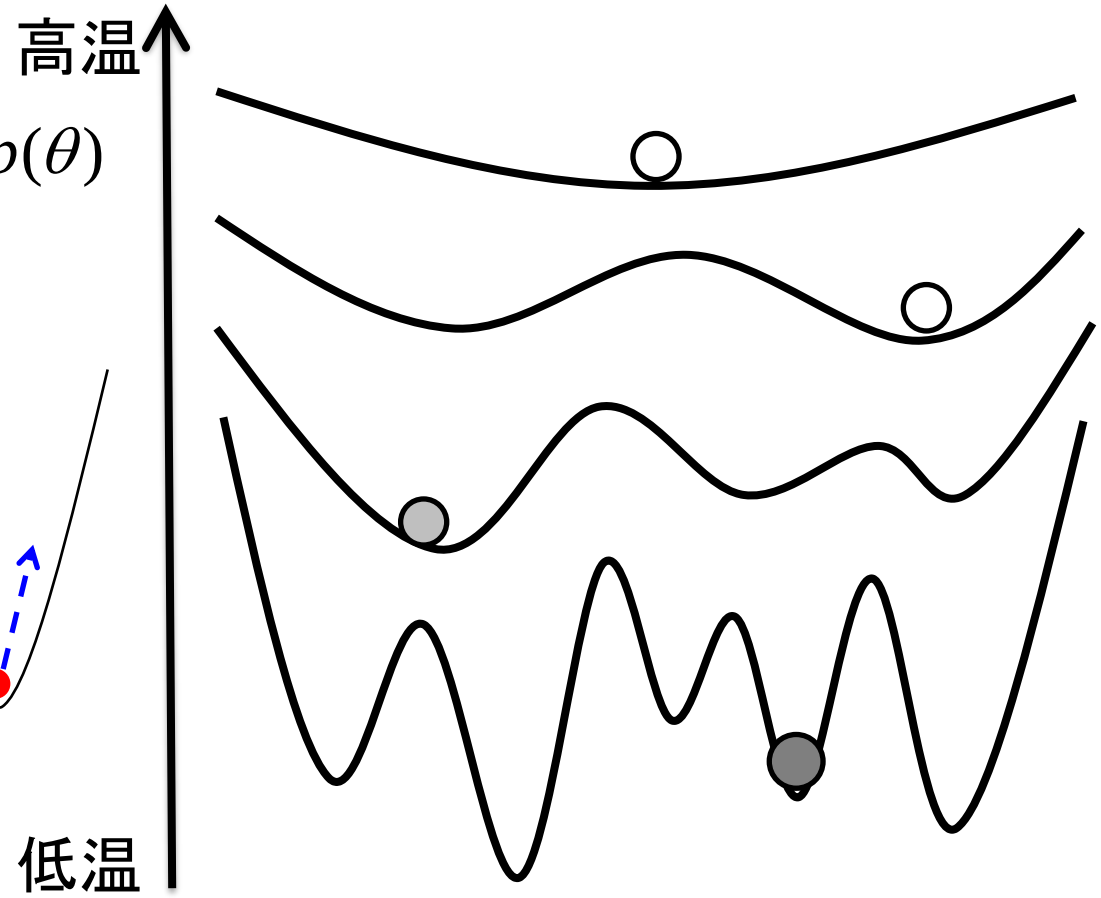
レプリカ交換モンテカルロ法
確率分布の数値的厳密解法

メトロポリス法

$$p_\beta(\theta) \propto \exp\left(-\frac{n}{\sigma^2}\beta E(\theta)\right) p(\theta)$$



レプリカ交換モンテカルロ法



K. Hukushima, K. Nemoto, *J. Phys. Soc. Jpn.* **65** (1996).