2I4-OS-10b-2

マルチピークスペクトル分解の高速化

Accelerating Multi-peak Spectral Deconvolution

永田賢二 *1　本武陽一 *2　田中顕至 *2　岡田 真人 *2
Nagata Kenji　Mototake Yohichi　Tanaka Kenji　Okada Masato

*1国立研究開発法人産業技術総合研究所人工知能研究センター
Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology
*2東京大学大学院新領域創成科学研究科複雑理工学専攻
Graduate School of Frontier Science, The University of Tokyo

This study focuses on Bayesian spectral deconvolution with the Exchange Monte Carlo method as an example of data-driven approach. We propose a parallelization algorithm of Bayesian spectral deconvolution in order to reduce its computational cost, and show its effectivity by simulation of synthetic data.

1. 背景と目的

ナノセンサ，量子計測など，微細・微量な対象を計測する技術が次々と開発されつつある先端デバイス分野において，高度な計測により得られる大量のデータから背後に潜む物理特性や構造を抽出するためのデータ駆動的アプローチは計測技術と高度情報処理の融合のキーテクノロジーである．視覚脳科学者である David Marr は，複雑な情報処理装置を考えるには，「計算理論」「表現・アルゴリズム」「ハードウェア」の三つのレベルに分けて議論すべきと説いている [Marr 82]．本研究では，David Marr の三つのレベルがデータ駆動科学における指針であると考え，この指針に基づき，マルチピークスペクトル分解のベイズ推定を議論する．

Nagata らは，図 1 に示すような多峰のスペクトルデータをガウス関数などの基底関数の和に分解するスペクトル分解におけるベイズ推定を表現のレベルとして用いることを提案した [Nagata 12]．これにより，従来恣意的に決定されていたピークの個数もモデル選択の枠組みでデータから客観的に決定することができることを示した．また，用いるアルゴリズムとして，交換モンテカルロ法を利用することで，ピークパラメータ最適化における局所最適化の問題を解決するとともに，モデル選択に必要となる自由エネルギーも効率的に計算できることを示した．一方で，交換モンテカルロ法は計算量が膨大になる問題があり，計算量削減は重要な課題である．

そこで本研究では，交換モンテカルロ法における並列プログラミング法を提案する．特に，ハードウェア拘束として，スパコンを代表とするメモリ分散型並列計算機ではなく，マルチコア CPU や GPGPU などのようなメモリ共有型の並列計算機に特化した手法を提案し，その有効性を計算機シミュレーションにより明らかにする．

2. スペクトル分解のベイズ推定

スペクトル分解とは，図 1 のような多峰性スペクトルを，ガウス関数に代表される単峰性の基底関数の線形和で表現する問題である [Nagata 12]．

$$G(x;\theta,K) = \sum_{k=1}^{K} a_k \exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right) \quad (1)$$

連絡先: 岡田真人，東京大学大学院新領域創成科学研究科 〒277-8561 千葉県柏市柏の葉 5-1-5, okada@k.u-tokyo.ac.jp

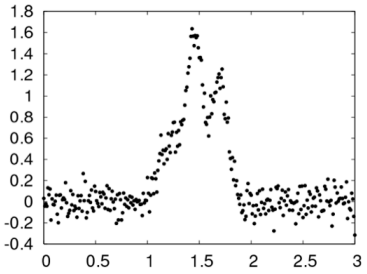

図 1: スペクトルデータの例

ここで K はガウス関数の個数であり，パラメータセット θ は $\theta = \{a_k, \mu_k, \sigma_k\}_{k=1}^{K}$ である．スペクトル分解の計算理論の目標は，図 1 のような多峰性スペクトルから，ピーク数 K と，それにともなうパラメータセット θ を適切に決めることである．本研究では，スペクトル分解の表現として，ベイズ推定を取り扱う．ベイズ推定では，観測データ y が生成された因果律を確率的に定式化する．まず確率 $p(K)$ に従って，ピーク数 K が生成され，次に，その K に従い，パラメータセット θ が条件付き確率 $p(\theta|K)$ で決定されると考える．ここでは簡単のために $p(K)$ と $p(\theta|K)$ は一様であるとする．観測データ y は，式 (1) に従う真の値に平均 0 分散 1 のガウスノイズが重畳されて観測されるとする．これらの仮定より，観測データ y は，ピーク数 K とパラメータセット θ が与えられた元で，入力 x の条件付き確率で生成される．

$$p(y|x,\theta,K) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-G(x;\theta,K))^2}{2}\right) \quad (2)$$

それぞれのデータ (x_i, y_i) が独立に得られたとすると，パラメータ θ が与えられた元での，N 個のデータ D の条件つき確率は，

$$p(D|\theta,K) = \prod_{i=1}^{n} p(y_i|x_i,\theta,K) \propto \exp\left(-NE(\theta,K)\right) \quad (3)$$

$$E(\theta,K) = \frac{1}{2N}\sum_{i=1}^{N}(y_i - G(x_i;\theta,K))^2 \quad (4)$$

となる．ベイズ推定では，ベイズの定理と周辺化の手続きを用いて，データ D と基底関数の個数 K が与えられた場合の，

パラメータセットの事後確率 $p(\theta|D,K)$ を導出し，推定に用いる．

$$p(\theta|D,K) \quad = \quad \frac{p(D,\theta,K)}{p(D,K)} \propto \exp\left(-NE(\theta,K)\right) \quad (5)$$

また，データ D が与えられた場合の，基底関数の個数 K の事後確率 $p(K|D)$ を用いて，ピーク数 K を推定する．

$$p(K|D) \quad = \quad \frac{p(D,K)}{p(D)} \propto \int d\theta \exp\left(-NE(\theta,K)\right) \quad (6)$$

3. 交換モンテカルロ法

ベイズ推定を実行するアルゴリズムとして交換モンテカルロ (Exchange Monte Carlo, EMC) 法を考える [Geyer 91, Hukushima 96]．交換モンテカルロ法では，式 (5) の事後確率 $p(\theta|D,K)$ に擬似的な逆温度 $\beta = 1/T$ を導入し，以下の確率分布を考える．

$$p_\beta(\theta|D,K) \quad \propto \quad \exp\left(-N\beta E(\theta,K)\right) \quad (7)$$

温度 T が違う系を複数用意し，以下の 2 種類の手続きを交互に実行することで，同時並列にサンプリングを行う．

1. **それぞれの分布におけるサンプリング**
   メトロポリス法により，それぞれの確率分布 $p_{\beta_l}(\theta_l|D,K)$ からのサンプリングを並列に実行する．
2. **２つの分布間における確率的な交換**
   適当なステップ毎にサンプル θ_l と θ_{l+1} を確率 $u = \min(1,v)$ で交換する．

$$v \quad = \quad \exp\left((\beta_{l+1}-\beta_l)\left(E(\theta_{l+1},K)-E(\theta_l,K)\right)\right)$$

これら複数の系をレプリカと呼ぶ．交換モンテカルロ法を利用すると，隣り合った温度間で確率的にレプリカを交換することで，局所解から脱却し，効率的に最小解を探索することができる．また，複数温度でモンテカルロ法を行うため，ピーク数の決定に必要な自由エネルギー $F(K)$ も同時に計算できる利点もある．

4. 提案手法および検証

本研究では，EMC 法の計算量削減を目的とした並列プログラミング手法を提案する．特に，マルチコア CPU や GPGPU を想定したメモリ共有型並列計算機に特化した手法を提案する．

従来の並列プログラミングでは，EMC 法の 2 つの手続きのうち，1 番目のサンプリングについては，レプリカごとに完全に独立に実行することができることに着目し，図 2(a) のよう並列化される．手続き 1 の各レプリカでの状態更新を並列実行し，交換のプロセスの際に同期を取るというのを繰り返す方法である．このアルゴリズムでは，メモリ分散にて実行できるため，クラスタマシンなどで実装される．一方で，毎ステップで並列ジョブの投入の必要があり，この部分が計算の律速になるという問題がある．

そこで本研究では，メモリ共有型の並列化として，図 2(b) のアルゴリズムを提案する．はじめに各プロセスに担当させるレプリカを割り振り，その後，交換も含めて独立に実行するようなアルゴリズムである．この際，他のプロセスが担当するレプリカとの間の交換が必要なため，対象となるレプリカの状態についての同期更新を避けるための適切な排他制御を組み込む必要がある．この手法では，レプリカ全体ではなく，交換に関係する 2 つのレプリカ間だけで同期をとるだけで実行することが可能となり，計算資源の効率的利用につながる．

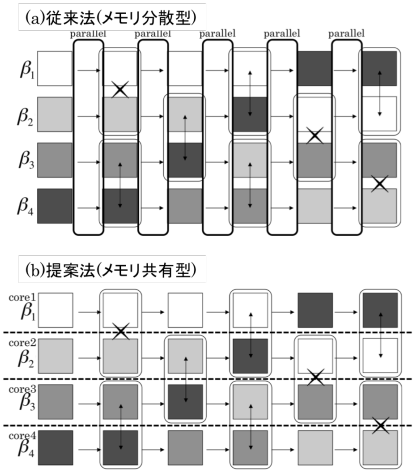

図 2: 交換モンテカルロ法の並列プログラミングの概要．(a) は従来法 (メモリ分散型) で (b) は提案法 (メモリ共有型) を表す．

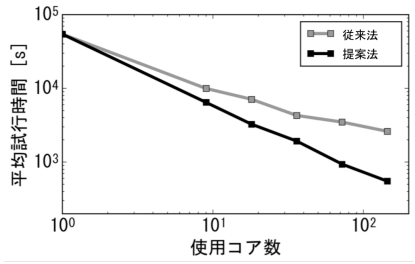

図 3: 二つの手法について計算時間の比較．

本提案手法の有効性を示すために，図 1 のデータに対するスペクトル分解を適用した場合で検証した．図 1 は人工データであり，$K = 3$ として生成された．図 3 は，メモリ分散型のアルゴリズムとメモリ共有型のアルゴリズムを実行した際の計算時間を比較したものである．用いた計算機は，ProLiant DL580 Gen9 であり，144 コアの並列化が可能である．EMC 法の設定としてレプリカの個数を 288 個とした．計算時間の計測には，乱数シードを変化させて 10 試行独立したシミュレーションを行いその平均時間をプロットしている．図 3 より，二つのアルゴリズムの計算時間の差はコア数が増えるにつれて大きくなり，並列コア 144 の時点では 4.74 倍の高速化が実現されたことがわかる．これは提案手法の有効性を示している．

参考文献

[Marr 82] Marr, D., “Vision.” MIT press (1982).

[Nagata 12] Nagata, K., Sugita, S., and Okada, M., “Bayesian spectral deconvolution with the exchange Monte Carlo method.” Neural Networks 28 (2012): 82-89.

[Geyer 91] Geyer, C. J., “Markov Chain Monte Carlo Maximum Likelihood.” in Computing Science and Statistics: Proc. of the 23rd Symposium on the Interface (1991): 156-163.

[Hukushima 96] Hukushima, K. and Nemoto, K. “Exchange Monte Carlo Method and Application to Spin Glass Simulations.” Journal of the Physical Society of Japan, 65 (1996): 980-988.