2I2-3in2

λ-スキャン法を用いたスパース基底選択とスペクトル分解への応用

Sparse Basis Selection By λ-scan Method
And Its Application To Spectral Deconvolution

本武 陽一 *1 Mototake Yohichi
五十嵐 康彦 *1 Igarashi Yasuhiko
竹中 光 *1 Takenaka Hikaru
永田 賢二 *2 Nagata Kenji
岡田 真人 *1 Okada Masato

*1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
*2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

Spectral data is useful data acquired by spectroscopic measurement. In particular, the number of peaks, the center position, width and intensity of each peak have important information. Therefore, spectral deconvolution that fitting spectral data as a sum of unimodal basis functions and estimating peak parameters is a useful method in analysis of spectral data. Nagata et al. Proposed a spectral deconvolution method using Bayesian estimation by the exchange Monte Carlo method and realized estimation of peak parameters and peak number. On the other hand, since the exchange Monte Carlo method requires a large calculation cost, it is difficult to apply it when the number of dimensions and the number of data become large. In order to overcome this problem, Igarashi et al. carried out a λ-optimization method using L1VM and one-standard error (1SE) rule, that method is much faster than Bayesian estimation method using exchange Monte Carlo. However, the 1SE rule used in this method is a heuristic method, and that theoretical basis is poor. That's why we tried to construct a spectrum deconvolution method, which called a "λ-scan method", which does not need to use 1SE rule, and we verified the λ-scan method has higher prediction performance than the *lambda*-optimization method.

1. 研究の背景と目的

スペクトルデータは物性物理学や生物学といった様々な分野において，物質の特性を反映する情報として取得される有用なデータである．スペクトルデータは，複雑な多峰性の構造を持っており，各ピークの中心位置や幅，強度から対象物質の性質に関する情報を取得することができる．そのため，スペクトルデータを単峰性の基底関数の和としてフィッティングし，ピークのパラメータを推定するスペクトル分解は，スペクトルデータの解析において有用な手法である．特に，基底関数の個数 K をいかに決めるかは重要な問題となる．

永田らは交換モンテカルロ法によるベイズ推定を用いたスペクトル分解法を提案し，ピークパラメータとピーク数の推定を実現した [Nagata 12]．一方で交換モンテカルロ法は大きな計算コストが必要となるため，次元数やデータ数が大きくなるような場合に適用することは難しい．

この問題に対して五十嵐らは，正則化項付きの回帰手法である L1VM を用いることによって，永田らと同程度の精度のスペクトル分解を高速に行えることを示した [Igarashi 16]．この高速化によって，高次元スペクトルデータのような，ベイズ推定法では計算量的に困難な系でのスペクトル分解が可能になると考えられる．しかしながらこの手法は，スペクトル分解に適した基底を選択するために，理論的根拠に乏しい one standard error ルール [Tibshirani 01]（以下，1SE ルール）と呼ばれるヒューリスティクス的手法を適用しなければならないという問題をもつ．

そこで本研究では，近似的に基底の組み合わせを全探索するλ-scan 法を提案し，1SE ルールのようなヒューリスティクスを用いずに，的確な基底選択とパラメータ推定が行えないかを検討する．

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

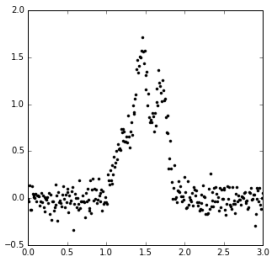

図 1: 分析対象とするスペクトルデータ

2. スペクトル分解

まずここで，今回対象とするスペクトル分解について記述する．今回対象とするスペクトルデータは，以下のように定式化されるガウス関数の和として生成されるものとする．

$$g(x) = \sum_{k=1}^{K} a_k exp(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2) \quad (1)$$

パラメータセットはピーク数 K を 3 とし，表 1 の通りとなる．具体的には図 1 のような分散がある程度違う，3 ピーク構造を持つスペクトルデータとなっている．

Param	値	Param	値	Param	値
a_1	0.587	a_2	1.522	a_3	1.183
μ_1	1.210	μ_2	1.455	μ_3	1.703
σ_1	0.1022	σ_2	0.0825	σ_3	0.0780

表 1: スペクトルデータのパラメータ

本研究では，このスペクトルデータをガウス基底関数を線形結合した線形モデル

$$G(x) = \sum_{j=1}^{N} w_j exp(-\frac{1}{2\sigma^2}(x-\mu_j)^2) \quad (2)$$

を用いて回帰し，ピーク数やピーク中心などのスペクトルのパラメータを推定することを考える．ここで，N はデータのサンプル数を表す．

3. 手法

3.1 λ-optimization 法 [Igarashi 16]

五十嵐らは，このスペクトル分解をガウス基底を用いた L1 正則化項付きの線形回帰問題（LASSO）として捉え，

$$E = \frac{1}{2N}\sum_{j=1}^{N}(g_j - G(x_j))^2 + \lambda\sum_{i=1}^{N}|w_i| \quad (3)$$

を最小化することを指標として回帰パラメータを推定した．ここで，$G(x)$，w_i は (2) 式によって与えられたものと同じであり，λ はスパース性をコントロールするハイパーパラメータである．ハイパーパラメータとしては，λ 以外に基底関数の幅を表す σ がある．このような手法を L1VM と呼ぶ．

さらに，そのハイパーパラメータは交差検証法によって決定された．交差検証法とは，オーバフィッティングしないパラメータを推定するために，データセットを何らかのルールで分割し，その一部を訓練データ，残る部分をテストデータとして用い，訓練結果の妥当性の検証を行う手法のことである．(後の説明のために，これで得られる交差検証誤差（CVE:Cross Validation Error）を CVE(λ, σ) と定義する．)

L1VM で用いられる lasso 法は，λ を CVE 最小化によって決定すると，真の基底数より多めの基底を選択する傾向を持っている [Tibshirani 01]．五十嵐らはこれに対応するため，交差検証で得られた最小の λ に対して，one standard error ルール（1SE ルール）と呼ばれるヒューリスティクス手法を適用している．1SE ルールとは，交差検証誤差の標準誤差を λ に加え，交差検証誤差のゆらぎの範囲内でよりスパースな基底を選択しようとする手法のことである．

全ての基底の分散を同じ値として固定しているため，回帰結果から直接ピーク数やピーク位置、ピーク分散などを推定することはできないが，これらの結果，ピーク数やピークの位置などのピーク構造を推定する上で有用な，選択された基底の中心位置というインディケータ情報が得られている（図 3）．

3.2 λ-scan 法

Lasso などの正則化法に基づくスパース推定法は，スパース化による基底選択と回帰を同時に行う枠組みと捉えることができる．一方 λ-scan 法は，この基底選択と回帰という機能を分離して段階的に行う枠組みに基づいており，さらにその基底選択を基底の組み合わせに関する近似的な全探索によって実現する手法である．基底数が多い場合，全探索を行えば組み合わせ爆発が生じる一方で，ほとんどの基底はピーク構造と関係しない（ピークから遠く離れたデータ点など）ため，実効的には全ての組み合わせを調べる必要はない．そこで λ-scan 法では，基底の全通りの組み合わせを，λ を走査しながら LASSO を実行し，その際に選ばれた基底セットを集めることで近似する．これにより，より適切な基底選択が実現されることが期待され，実際に λ-scan を線形回帰問題に適用した結果，より予測性能の高い回帰モデルが推定されることが報告されている [五十嵐 16]．

本研究ではこの λ-scan 法を，前節で説明したガウス基底による線形回帰問題に適用する．具体的には，正則化パラメータ λ や基底の分散 σ でグリッドサーチしながら，以下のような手順を繰り返し適用することで λ-scan 法が実現される．

step1. L1VM による基底選択
step2. 刈り込まれた基底空間での回帰
step3. CVE による (λ, σ) の評価

それぞれの step について，以下で詳細な説明を行う．

3.2.1 step1 L1VM による基底選択

各説明変数の回帰係数を 0 とすべきかを推定する問題は，L1 正則化による LASSO 回帰の推定結果を近似として用いることが可能であることが知られている [Tibshirani 96]．本研究ではこの知見を活用し，基底関数の空間での LASSO 回帰である L1VM を用いて基底選択を行う．具体的には，L1VM を用いて，与えられた λ と σ の元で，損失関数 (3) を最小化するような w_i を求め，その w_i のうち数値的に有限な値を持つものを，選択された基底とする．ちなみに本研究では L1VM の最適化アルゴリズムとして，各 w_i（w_0，w_1，w_2…）を順番に更新していく coordinate descent 法を用いた．

3.2.2 step2 刈り込まれた基底空間での回帰

step1 で選択された基底の元で，以下のように正則化項のない損失関数を最小化する単純な線形回帰を行なう．

$$E = \frac{1}{2N}\sum_{j=1}^{N}(g_j - G(x_j))^2 \quad (4)$$

この際 (λ, σ) の評価指標として CVE を算出する必要があるため，step3 で説明する方法で，データセット D を訓練データ D_{train} とテストデータ D_{test} に分け，訓練データで w_i を決定した後に，テストデータを用いて平均二乗誤差（MSE:Mean Squared Error）

$$MSE = \frac{1}{N_{test}}\sum_{j=1}^{N_{test}}(g_j - G(x_j))^2 \quad (5)$$

を求めた．

3.2.3 step3 CVE による λ の評価

CVE の算出方法について説明する．本研究では 10-fold 交差検証法を用いた．10-fold 交差検証法では，訓練データを 10 分割し，そのうちの 1 つをテストデータ、残る 9 個を訓練データに割り当てる．その上で，テストデータの割り当てを変えながら 10 回 MSE を計算し，その平均を交差検証誤差（CVE）とする．特に，選択された基底を表すインディケータベクトル C が与えられた元での CVE という意味を込めて，以降この CVE を $CVE(C(\lambda, \sigma))$ と呼称する．

4. 結果

前節の λ-scan 法に基づき $CVE(c(\lambda, \sigma))$ を算出し，λ と σ の空間にヒートマップとして表現した（図 2）．図 2 中の各点はそれぞれの手法による推定パラメータの位置を表す．前節で説明したように，λ-scan 法の step2 で行う回帰には正則化項が含まれないため，過学習を抑える機能はない．したがって $CVE(c(\lambda, \sigma))$ は，真に必要な基底”のみ”が選択された場合に

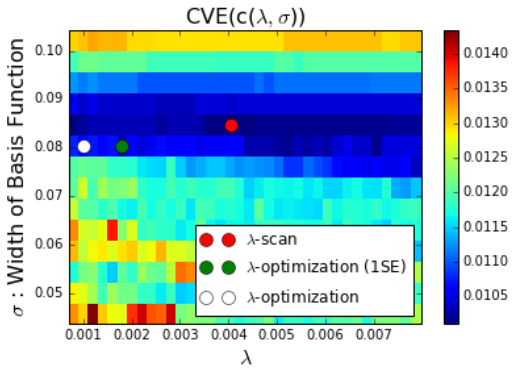

図 2: $CVE(c(\lambda,\sigma))$ による選択基底の比較.x 軸:λ、y 軸:基底関数の幅 σ、z 軸:$CVE(c(\lambda,\sigma))$

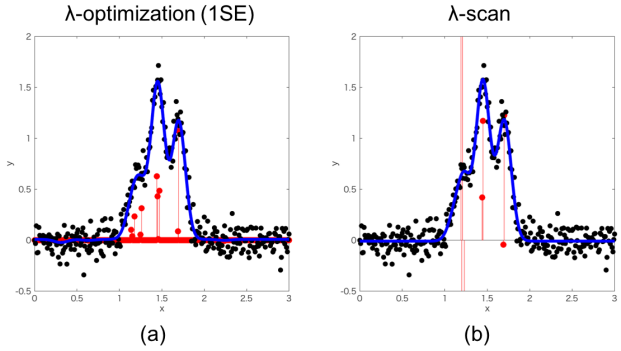

図 3: λ-optimization 法と λ-scan 法によるフィッティング結果の比較．(a).λ-optimization 法（1SE ルール適用後）による結果の再現図．(b).λ-scan 法によって選択された基底，λ，基底の幅 σ を用いて回帰を行なった結果．

最小化されるはずである．この視点から図 2 を見ると，λ-scan 法が，1SE ルール適用後の λ-optimization 法よりも，より良い基底選択を行なえているとわかる．実際にこの結果が得られるフィッティング結果を見ると（図 3），λ-optimization 法によるフィッティング結果（図 3(a)）と比較し，λ-scan 法による結果（図 3(b)）はより少ない基底数が選択されていることがわかる．さらに，一般的に成り立つ結果であるとは言い切れないが，λ-optimization 法では捉えきれていない最も幅の広いピーク（最左端）の中心を射抜いているようにも見える．

基底選択を含めた回帰モデル全体の予測誤差についても，λ-optimization 法によって最小化された予測誤差（1SE ルール適用前が最小）が $CVE(\lambda,\sigma) = 1.043 \times 10^{-2}$ となっている一方で，λ-scan 法によって最小化された予測誤差は $CVE(C(\lambda,\sigma)) = 1.011 \times 10^{-2}$ という値となっており，回帰モデル自体も λ-scan 法の方がより高い性能をもっていた．

このように，λ-scan 法を用いることで，1SE ルールのようなヒューリスティクスを用いることなく，それよりも良い基底選択と回帰モデルの構築が行えることがわかった．

5. まとめと展望

本研究によって，基底選択と回帰を分離し，基底選択を L1VM による近似的全探索によって行う λ-scan 法をスペクトル分解に適用することの有用性が確認された．特に，ヒューリスティクス的手法である 1SE ルールに従うことなく，それよりも良い予測性能を持つモデルを選択できたことは，非常に有用な知見であると考えられる．

参考文献

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

[Igarashi 16] Igarashi, Yasuhiko, et al. ”Three levels of data-driven science.” Journal of Physics: Conference Series. Vol. 699. No. 1. IOP Publishing, 2016.

[Tibshirani 01] Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning. Vol. 1. Springer, Berlin: Springer series in statistics, 2001.

[五十嵐 16] 五十嵐康彦, et al. ”全状態探索による線形回帰のスパース変数選択 (情報論的学習理論と機械学習)–(情報論的学習理論ワークショップ (IBIS2016)).” 電子情報通信学会技術研究報告= IEICE technical report: 信学技報 116.300 (2016): 313-320.

[Tibshirani 96] Tibshirani, R. (1996), Journal of the Royal Statistical Soci- ety. Series B (Methodological), 267-288.