

ベイズ的変数選択に基づく分光スペクトル分解

川島 貴大^{1,a)} 庄野 逸^{2,b)}

概要: 分光学において、測定したスペクトルデータからピーク数・ピーク形状・各ピークのパラメータを推定することは、試料の物性を知るために重要なタスクである。この分光スペクトルのピーク分離に関してレプリカ交換モンテカルロ法によるベイズ的スペクトル分解が提案されており、人工および実データに対して適切にピークを分離できることが示されている。しかしこの手法ではモデル選択の方法に基づいてピーク数を推定するため、複数のモデルを事前に用意する必要があり、さらに計算時間も大きくなる問題がある。そこで本研究では先行研究のモデルを拡張し、ベイズ的変数選択法の枠組みで階層モデリングを適用することで、人工データに対してより高速なピーク数とピーク形状およびパラメータの同時推定が可能であることを示した。さらに提案手法による実データへの応用としてコランダム・ラマンスペクトルの分析を行い、有効なピークを抽出できることを確認した。

キーワード: スペクトル分析, ベイズ的スペクトル分解, ベイズ的変数選択

Spectral Deconvolution Based on Bayesian Variable Selection

TAKAHIRO KAWASHIMA^{1,a)} HAYARU SHOUNO^{2,b)}

Abstract: In the field of spectroscopy, estimating the number of peaks and the shape and parameters of each peak from spectral data is a significant task. With respect to this task, Bayesian spectral deconvolution with replica exchange MCMC has been proposed and the effectiveness was shown. However, in Bayesian spectral deconvolution, we have to prepare multiple models and the computational time becomes large. Thus, in this study, we extended the previous model based on Bayesian variable selection, and we found that more efficient spectral deconvolution is possible for synthetic spectral data. In addition, we tried analyzing Raman spectrum of corundum and were able to extract valid peaks.

Keywords: Spectral Analysis, Bayesian Spectral Deconvolution, Bayesian Variable Selection

1. はじめに

分光分析では対象となる試料に電磁波をあて、吸収・反射した波の周波数特性などを測定し、その物性を調査する。測定されたデータから実際に知見を得るためには、スペクトルに現れるピークを分離してこれらのパラメータを適切に推定する必要がある。

スペクトルデータはしばしば RBF ネットワーク [7] の

形式、すなわち基底関数の線形和としてモデリングされる [3]。スペクトルに現れるピークの形状や数が既知なら各基底関数のパラメータを推定すればよいが、一般にはこれらも未知であり、モデルの推定が必要となることが多い。AIC[1] や WAIC[8] などのモデル選択手法や Lasso[9] のようなスパース推定手法を用いれば基底関数の選択が可能であるが、スペクトル分析において重要なのは、よりよくデータの生成モデルを近似する統計モデルを得ることであり、汎化能力の観点で議論されるこれらの手法の適用は妥当でないと考えられる。

これに対して Nagata らはベイズ的スペクトル分解による分光スペクトルの分析法を提案した [5], [10]。この手法

¹ 電気通信大学 情報理工学部
東京都調布市調布ヶ丘 1-5-1

² 電気通信大学大学院 情報理工学研究科
東京都調布市調布ヶ丘 1-5-1

a) kawashima@uec.ac.jp

b) shouno@uec.ac.jp

では RBF ネットワークの形式でスペクトルデータをモデリングし、ベイズ推論の処方箋に基づいて各基底のパラメータの事後分布を定式化する。このとき基底数の異なる複数のモデルを用意しておき、レプリカ交換モンテカルロ法により評価した周辺尤度に基づきモデル選択の観点でピーク数の推定を行う。しかしベイズ的スペクトル分解では事前に実験者が目星をつけて複数のモデルを準備する必要がある。さらにレプリカ交換モンテカルロ法は計算量が大きく、これを準備した各モデルに対して適用するために全体の計算時間も大きくなってしまいう問題がある。

そこで本研究ではベイズ的変数選択法 SSVS (Stochastic Search Variable Selection) [2] に基づくスペクトルデータのピーク数とパラメータの同時推定法を提案する。SSVS では変数選択を行いたいパラメータに対して spike and slab 事前分布を導入することで不要なパラメータを刈り込むために縮小推定が起こりづらく、また各基底の有効性を示す指示変数を導入するために結果の解釈が容易であるなどの特長がある。提案手法ではこの SSVS を各ピークに対する強度のパラメータに適用することでピーク数選択を試みる。

実験として人工および実データに対して提案手法を適用し、真のピーク数より多いピーク数をもつ単一のモデルを事前に与えることで、ピーク数、ピーク形状、パラメータを同時に推定できることを示した。

2. 手法

本章では提案手法であるベイズ的変数選択に基づく分光スペクトル分解について紹介する。

2.1 モデル

まず N 個のスペクトルデータの組が得られたとする。すなわちデータ集合 \mathcal{D} は $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ と表される。ここで x_i は波数や化学シフト、束縛エネルギーなどの量であり、実際に使用した分光法によってその単位は異なる。 y_i は x_i に対応するスペクトル強度であり、この $\mathbf{x} = (x_i)_{i=1}^N$ から $\mathbf{y} = (y_i)_{i=1}^N$ への線形回帰基底モデルを考える。スペクトル強度 y_i をパラメータ集合 $\boldsymbol{\theta}$ とピーク数 K を用いて

$$y_i = f(x_i; \boldsymbol{\theta}) + \epsilon_i$$

$$= \sum_{k=1}^K a_k \phi_k(x_i; \mu_k, b_k) + \epsilon_i \quad (1)$$

$$\epsilon_i \sim \mathcal{N}(\epsilon_i; 0, \sigma^2) \quad (2)$$

とモデリングする。 a_k, μ_k, b_k はそれぞれ k 番目のピークの強度、位置、鋭さを示すパラメータである。すなわちパラメータ集合 $\boldsymbol{\theta}$ は $\boldsymbol{\theta} = \{a_k, \mu_k, b_k\}_{k=1}^K$ と表される。 ϕ_k は k 番目のピークの形状を示す基底関数である。ピーク形状は試料の物理的特性によって式 (3) のようなガウス型、式 (4) のようなローレンツ型、それらのたたみ込みであるフークト型などをとることが知られている。

$$\phi_k^G(x_i; \mu_k, b_k) = \exp\left(-\frac{b_k}{2}(x_i - \mu_k)^2\right) \quad (3)$$

$$\phi_k^L(x_i; \mu_k, b_k) = \frac{1}{1 + (b_k(x_i - \mu_k))^2} \quad (4)$$

さらに式 (2) のようにノイズ ϵ_i が平均が 0 で分散が既知の値 σ^2 の正規分布に従うとすると、ベイズの定理よりピーク数が K のときのパラメータ集合 $\boldsymbol{\theta}$ の事後分布は、 $i = 1, \dots, N$ に関してデータが独立であるという仮定のもとで

$$p(\boldsymbol{\theta} | \mathcal{D}, K) = \frac{p(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta}, K) \pi(\boldsymbol{\theta} | K)}{p(\mathbf{y} | \mathbf{x}, K)}$$

$$= \frac{\pi(\boldsymbol{\theta} | K)}{p(\mathbf{y} | \mathbf{x}, K)} \exp\left(-\frac{N}{\sigma^2} E(\boldsymbol{\theta})\right) \quad (5)$$

と表現できる。ここで $\pi(\boldsymbol{\theta} | K)$ はピーク数 K のもとでのパラメータ集合 $\boldsymbol{\theta}$ の事前分布であり、 $E(\boldsymbol{\theta})$ は以下のように二乗誤差によって定義されるエネルギー関数である。

$$E(\boldsymbol{\theta}) = \frac{1}{2N} \sum_{i=1}^N (y_i - f(x_i; \boldsymbol{\theta}, K))^2 \quad (6)$$

2.2 スペクトル分析における変数選択問題

スペクトル分析における変数選択問題では、ピーク数 K をある程度大きく見積もった上でスペクトル強度 y_i を式 (1) のように定式化し、各ピーク強度 a_k に関して変数選択を適用する。これによりピーク数を $0 \sim K$ の範囲で推定することができる。

回帰などの問題における変数選択手法としては Lasso が標準的に使用される [9]。しかし Lasso のような正則化手法により回帰パラメータを縮小推定すると、データへの過剰適合が抑制され汎化性能が向上することがある。一方スペクトル分析で興味の対象となるのは「新しいデータに対してどれだけ汎化性能を向上させられるか」ではなく「いかに真の生成分布をよく近似した統計モデルを得られるか」であり、これらの手法を用いることは適切でないと考えられる。そこで本研究では SSVS と呼ばれるベイズ的変数選択手法を導入する。

2.3 SSVS による分光スペクトル分解

SSVS は spike and slab 事前分布に基づくベイズ的変数選択手法である [2]。spike and slab 事前分布は 0 周辺で非常に鋭い領域 (spike) とそれ以外の緩やかな領域 (slab) からなる。SSVS では推定するパラメータの事前分布に補助的な指示変数 $\gamma_i \in \{0, 1\}$ を導入することで事前分布を spike または slab でスイッチングさせ、不要なパラメータを刈り込む。

提案手法ではこの SSVS を以下の式 (7) のようにピーク強度 a_k に対して適用することで有効な基底の選択を行う。

$$\pi(a_k|\gamma_k) = (1 - \gamma_k)\mathcal{N}^+(a_k; 0, \tau^2) + \gamma_k\mathcal{N}^+(a_k; 0, g\tau^2) \quad (7)$$

$$\pi(\gamma_k) = \text{Bernoulli}(p_k) \quad (8)$$

ここで τ^2 は小さい正の実数、 g は大きい正の実数で、どちらも定数である。 p_k は指示変数 γ_k が $\gamma_k = 1$ となる事前確率である。 $\mathcal{N}^+(0, s^2)$ は平均が 0 で分散が s^2 の正規分布から 0 以上の領域のみを取り出して正規化した半正規分布であり、ピーク強度 a_k の非負性を反映している。この事前分布の概形は図 1 のようになる。式 (7) において $\gamma_k = 0$ であるときは右辺第 1 項のみが有効となる。このとき a_k の事前分布は 0 周辺で非常に鋭いものであるため、 a_k の推定値は 0 にきわめて近くなり、対応する基底 ϕ_k は不要であるとみなされたことになる。一方で $\gamma_k = 1$ のときは緩やかな事前分布を表す右辺第 2 項のみが寄与するため、 a_k の推定値は値をもち、 ϕ_k は有効であるとみなせる。よってここではパラメータ集合 θ は a_k の事前分布を制御する階層パラメータ γ_k を加えて $\theta = \{\gamma_k, a_k, \mu_k, b_k\}_{k=1}^K$ となる。

このように提案手法ではピーク数が多く表現力の高い単一のモデルを用意し、そのモデルからピーク数と各ピークのパラメータを同時に推定する。これにより複数のモデルを用意する必要性を省くことができ、全体の計算時間は削減される。また基底関数 ϕ_k について各 k に対し異なるものを設定すれば、同様の枠組みでピーク形状を選択することも可能である。

2.4 事後分布の評価

提案手法では事後分布からのサンプリングを安定させるためにレプリカ交換モンテカルロ法 [4] を用いて事後分布 $p(\theta|D, K)$ の評価を行う。なお提案手法を用いて各基底 ϕ_k の有効性を判断する際には、指示変数 γ_k の事後確率 $p(\gamma_k = 1|D)$ が一定の閾値以上になるかを見ればよい。

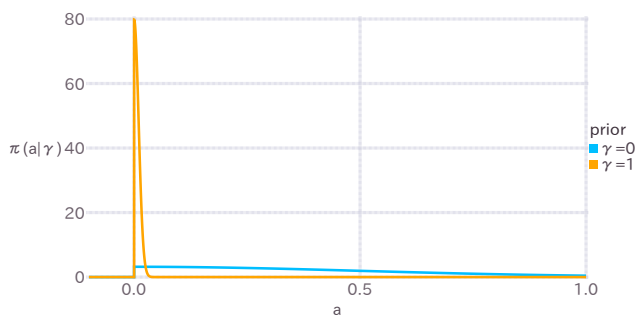


図 1 提案手法における spike and slab 事前分布

3. シミュレーションおよび考察

提案手法の有効性を評価するために人工データとコランダムのラマンスペクトルに対して実験を行った。本章ではこの実験の設定、結果および考察について述べる。

3.1 ガウス型人工データ

まず図 2 のような人工のスペクトルデータを用い、ピーク数とパラメータの推定を行った。入力は $\mathbf{x} = (0.00, 0.01, \dots, 3.00)^T$ 、データ数は $N = 301$ とした。このデータの真のピーク数は $K^* = 3$ 、基底関数はすべてガウス型である。すなわち真の生成モデルは以下の式 (9) のとおりである。

$$y_i = \sum_{k=1}^{K^*} a_k^* \phi_k^G(x_i; \mu_k^*, b_k^*) + \epsilon_i \quad (9)$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

ここで各パラメータの真値 $\{a_k^*, \mu_k^*, b_k^*\}_{k=1}^3$ を表 1 に示す。またノイズ分散は $\sigma^2 = 0.01$ で既知とした。

このデータに対して式 (1) のような RBF ネットワークの形式で表されるモデルを用意し、提案手法およびベイズ的スペクトル分解によりピーク分離を行った。このとき各基底関数 $\phi_k (k = 1, \dots, K)$ はすべて式 (3) で示されるガウス型とした。提案手法とベイズ的スペクトル分解でのレプリカ交換モンテカルロ法のレプリカ数 L はそれぞれ $L = 48$ 、 $L = 24$ とし、各レプリカの逆温度 $\beta_l (l = 1, \dots, L)$ は

$$\beta_l = \begin{cases} 0 & (l = 1) \\ 1.5^{l-L} & (\text{otherwise}) \end{cases} \quad (10)$$

と定めた。どちらの場合もモンテカルロステップは 10000 回、うち burn-in は 3000 回である。

提案手法では最大ピーク数は $K = 8$ とし、各パラメータの事前分布は以下のように設定した。

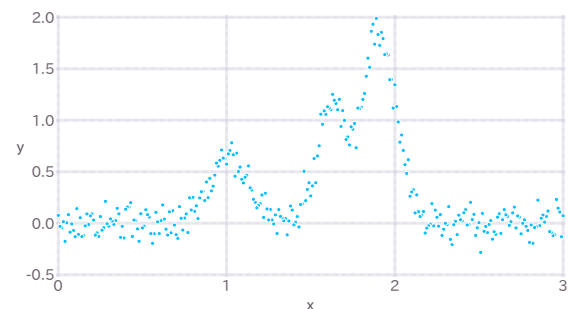


図 2 ガウス型の人工スペクトルデータ ($K^* = 3$)

表 1 ガウス型人工データの真のパラメータ

k	a_k^*	μ_k^*	b_k^*
1	0.62	1.02	69.44
2	1.21	1.62	152.02
3	1.87	1.91	100.00

$$\pi(\gamma_k) = \text{Bernoulli}(\gamma_k; \kappa_\gamma^{-\alpha_\gamma k}) \quad (11)$$

$$\pi(a_k | \gamma_k) = (1 - \gamma_k) \mathcal{N}^+(a_k; 0, \tau^2) + \gamma_k \mathcal{N}^+(a_k; 0, g\tau^2) \quad (12)$$

$$\pi(\mu_k) = \mathcal{N}(\mu_k; \nu_0, \xi_0^{-1}) \quad (13)$$

$$\pi(b_k) = \text{Gamma}(b_k; \eta_b, \lambda_b) = \frac{\lambda_b^{\eta_b}}{\Gamma(\eta_b)} b_k^{\eta_b-1} \exp(-\lambda_b b_k) \quad (14)$$

また各ハイパーパラメータの値は表 2 のとおりである。

ベイズ的スペクトル分解での実験では $K = 1$ から $K = 8$ の 8 つのモデルからピーク数選択を行った。ピーク強度 a_k の事前分布は

$$\pi(a_k) = \text{Gamma}(a_k; 5.0, 5.0) \quad (15)$$

とし、 $\pi(\mu_k)$ および $\pi(b_k)$ に関してはそれぞれ式 (13)、式 (14) と表 2 に示されているハイパーパラメータを使用した。

まず提案手法について $\gamma_k = 1$ となる事後確率 $p(\gamma_k | \mathcal{D}, K)$ を図 3 に示す。横軸は k 、縦軸は $p(\gamma_k = 1 | \mathcal{D}, K)$ であり、 $p(\gamma_k = 1 | \mathcal{D}, K) \geq 0.5$ であるピークを有効であると解釈すれば、真のピーク数 $K^* = 3$ が正しく推定されている。

次に提案手法およびベイズ的スペクトル分解での a_k 、 μ_k 、 b_k の事後分布を図 4 に示す。図 4(a) は提案手法、図 4(b) はベイズ的スペクトル分解において $K = 3$ であるときのものである。青線、黄線、赤線はそれぞれ $k = 1, 2, 3$ に対応し、点線は真の値である。提案手法は $K = 8$ の表現力の高いモデルであるためにベイズ的スペクトル分解よりも

表 2 ガウス型人工データに対するハイパーパラメータ

パラメータ	値	パラメータ	値
κ_γ	e	η_b	5.0
α_γ	5.0	λ_b	0.04
ν_0	1.5	τ^2	0.1
ξ_0	5.0	g	$\sqrt{2000.0}$

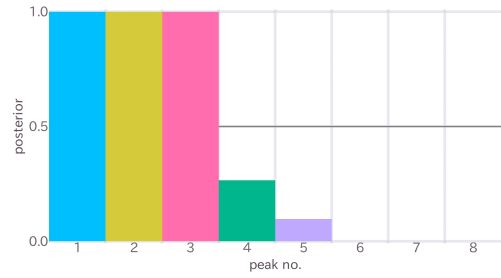


図 3 ガウス型人工データに対する提案手法でのピーク数選択結果

事後分布の分散は大きくなっている。表 3 は提案手法およびベイズ的スペクトル分解の MCMC サンプルからカーネル密度推定を用いて算出した MAP 推定結果である。また図 5 は表 3 による人工スペクトルデータへのフィッティング結果であり、図 5(a) が提案手法、図 5(b) がベイズ的スペクトル分解によるものである。提案手法では MAP 推定値としてベイズ的スペクトル分解にきわめて近い値が得られており、各パラメータに点推定を行うことを考えると、ピーク数 K が正しく推定できた場合にはベイズ的スペクトル分解と同等のピークパラメータに対する推定性能が得られることがわかる。

図 6 は計算時間の比較である。提案手法では単一のモデルから基底を選択するため計算時間が削減されている。より大きなモデルを用意する必要性のあるデータに対しては、計算時間に関する優位性はさらに増加することが見込める。

3.2 コランダムのラマンスペクトルの分析

提案手法の実データへの応用として、コランダムのラマンスペクトルに対して実験を行った。データは無機物のラマンスペクトルのデータベースである RRUFF Project 上で公開されているものであり、波長 532nm の励起光を用いて

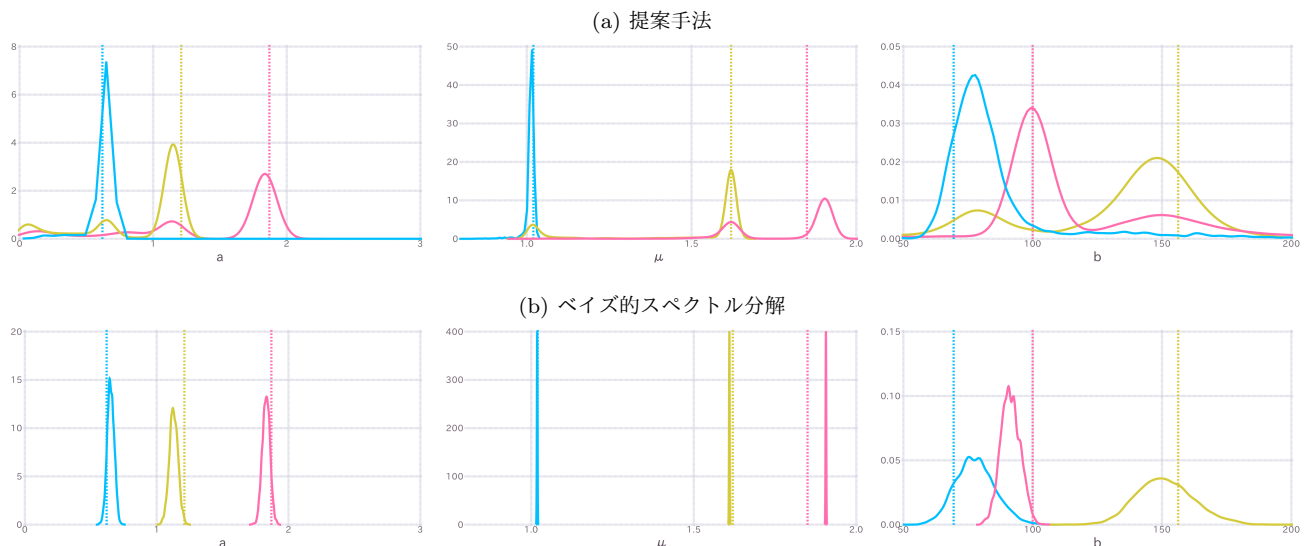


図 4 推定された a_k 、 μ_k 、 b_k の事後分布

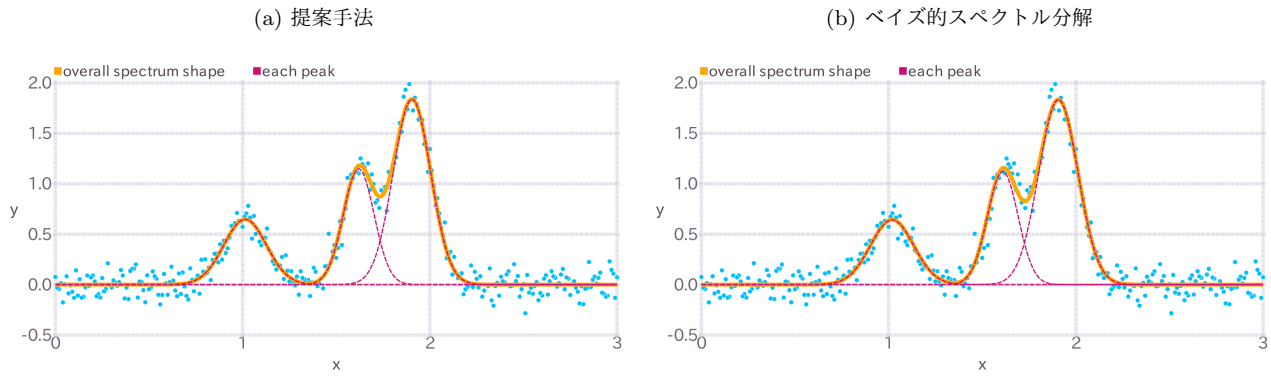


図 5 MAP 推定値によるフィッティング結果

表 3 ガウス型人工データのパラメータ推定結果

パラメータ	真の値	提案手法	ベイズ的 スペクトル分解
a_1	0.62	0.65	0.64
a_2	1.21	1.14	1.12
a_3	1.87	1.83	1.83
μ_1	1.02	1.01	1.02
μ_2	1.62	1.62	1.61
μ_3	1.91	1.90	1.90
b_1	69.44	77.42	75.39
b_2	152.02	148.12	149.58
b_3	100.00	99.64	90.67

表 4 コランダムのランンスペクトルに対するハイパーパラメータ

パラメータ	値	パラメータ	値
κ_γ	2.0	η_b	1.0
α_γ	1.0	λ_b	0.1
ν_0	600	τ^2	1.0×10^{-4}
ξ_0	5.0×10^{-5}	g	2.0×10^4

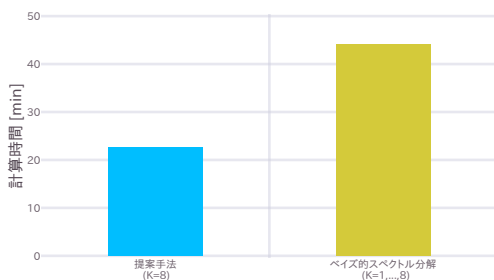


図 6 ガウス型人工データに対する計算時間

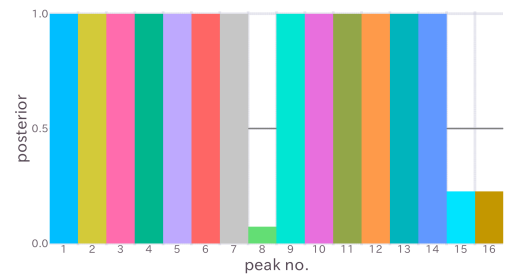


図 7 コランダムのランンスペクトルに対するピーク選択結果

測定されている。実際のデータではラマンシフト x として $176.3\text{cm}^{-1} \sim 1303.0\text{cm}^{-1}$ の計 2338 点においてスペクトル強度が測定されているが、ここでは $250\text{cm}^{-1} \sim 850\text{cm}^{-1}$ のラマンシフトに対応するデータのみを用い、データ数は $N = 1245$ となっている。またスペクトル強度は $0 \sim 2$ に正規化し、ノイズの分散は $\sigma^2 = 2.5 \times 10^{-5}$ で定数とした。モデルとしてガウス型およびローレンツ型の基底関数をそれぞれ 8 個ずつ用意し、ハイパーパラメータの値は表 4 に設定した。ガウス型およびローレンツ型の基底に対するパラメータ a_k, μ_k, b_k の事前分布はおのおの共通である。レプリカ交換モンテカルロ法のレプリカ数は $L = 128$ とし、各逆温度は

$$\beta_l = \begin{cases} 0 & (l = 1) \\ 1.225^{l-L} & (\text{otherwise}) \end{cases} \quad (16)$$

とした。

図 7 は各ピークの有効性を示す指示変数 γ_k の事後確率 $p(\gamma_k = 1|\mathcal{D})$ であり、 $k = 1, \dots, 8$ はガウス型基底、 $k = 9, \dots, 16$ はローレンツ型基底に対応する。また Porto らによるコランダムのラマンモード解析の結果 [6] と本実験によるピーク分離の結果を表 5 に示す。表中の (G) および (L) はそれぞれガウス型基底とローレンツ型基底を表す。ここで Porto らによる値は本実験で用いたデータとは異なる個体の試料に対して解析を行った結果であり、真の値ではないことに注意されたい。図 8 は各パラメータの MAP 推定値によるフィッティングの結果である。

表 5 に示されているように、提案手法ではコランダムのラマンモードとして知られているピークをすべて抽出することができている。提案手法では Porto らによる報告よりもピーク数が大きく推定されているが、図 8 から、これはデータに現れるピークの非対称性に起因していると考えられる。一方表 5 と図 8 より、 500cm^{-1} 以上のラマンシフトに現れる非ガウス型のピークに対し、適切にローレンツ型の基底を選択できていることが読みとれる。

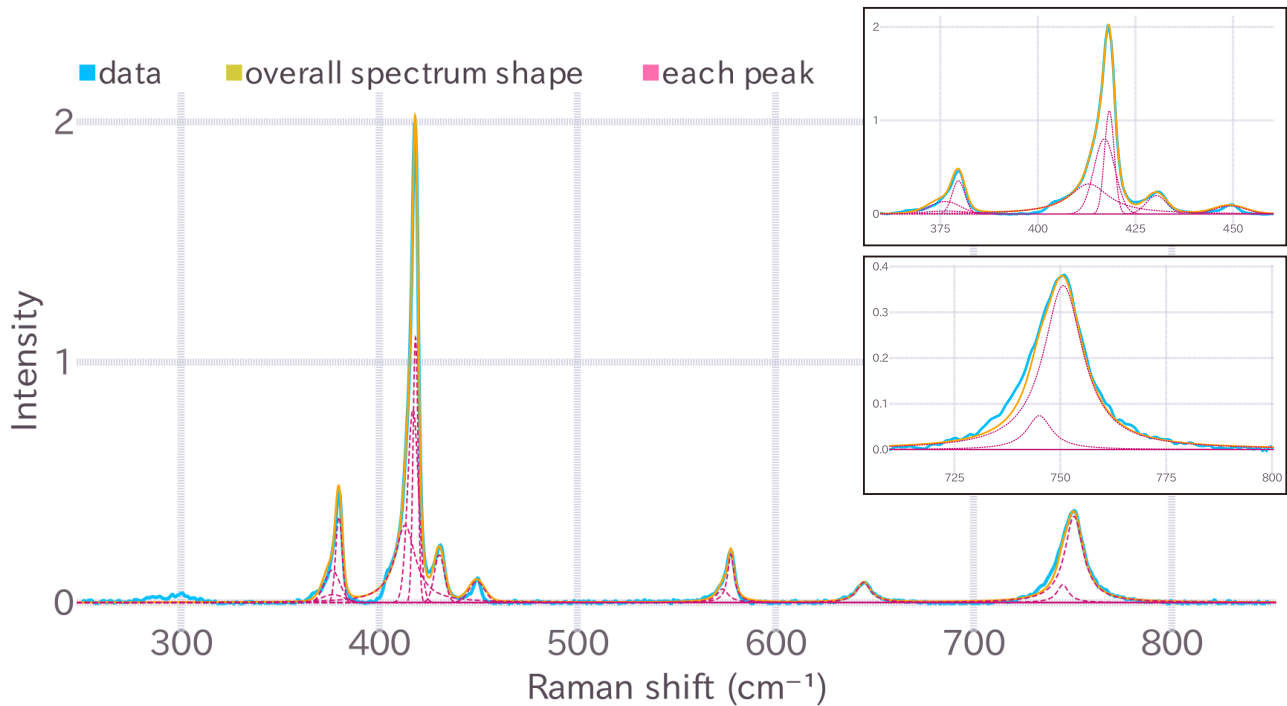


図 8 コランダム・のラマンスペクトルに対するフィッティング結果

表 5 コランダム・のピーク分離結果

Porto ら (1967)		Porto ら (1967)	
ラマンシフト (cm ⁻¹)	提案手法	ラマンシフト (cm ⁻¹)	提案手法
378	376(G)	451	449(G)
—	377(G)	578	573(L)
—	380(G)	—	577(L)
418	413(L)	645	645(L)
—	417(G)	751	745(L)
—	418(G)	—	750(L)
432	430(G)		

4. 結論

本研究では Nagata らによるベイズ的スペクトル分解 [5] のモデルを拡張し、ベイズ的変数選択法 SSVS[2] に基づいた分光スペクトル分解手法を提案した。提案手法では各ピークの強度パラメータに対して spike and slab 事前分布を設定することで、単一のモデルからピーク数、ピーク形状、およびそれらのパラメータの同時推定が可能である。人工データに対する実験により、提案手法では先行研究である Nagata らによるベイズ的スペクトル分解より計算時間が削減され、かつパラメータの MAP 推定値は同等なものが得られることを確認した。さらにコランダム・のラマンスペクトルに対して提案手法を適用し、実データに対してもその有効性を示すことができた。

参考文献

- [1] Akaike, H.: A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, Vol. 19, No. 6, pp. 716–723 (1974).
- [2] George, E. I. and McCulloch, R. E.: Variable Selection via Gibbs Sampling, *Journal of the American Statistical Association*, Vol. 88, No. 423, pp. 881–889 (1993).
- [3] Griffiths, P. R. and Pariente, G. L.: Introduction to spectral deconvolution, *TrAC Trends in Analytical Chemistry*, Vol. 5, No. 8, pp. 209–215 (1986).
- [4] Hukushima, K. and Nemoto, K.: Exchange Monte Carlo Method and Application to Spin Glass Simulations, *Journal of the Physical Society of Japan*, Vol. 65, No. 6, pp. 1604–1608 (1996).
- [5] Nagata, K., Sugita, S. and Okada, M.: Bayesian spectral deconvolution with the exchange Monte Carlo method, *Neural Networks*, Vol. 28, pp. 82–89 (2012).
- [6] Porto, S. P. S. and Krishnan, R. S.: Raman Effect of Corundum, *The Journal of Chemical Physics*, Vol. 47, No. 3, pp. 1009–1012 (1967).
- [7] Powell, M. J. D.: Radial Basis Functions for Multivariable Interpolation: A Review, *Algorithms for Approximation*, pp. 143–167 (1987).
- [8] Sumio, W.: Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory, *Journal of Machine Learning Research*, Vol. 11, pp. 3571–3594 (2010).
- [9] Tibshirani, R.: Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 58, No. 1, pp. 267–288 (1996).
- [10] Tokuda, S., Nagata, K. and Okada, M.: Simultaneous estimation of noise variance and number of peaks in Bayesian spectral deconvolution, *Journal of the Physical Society of Japan*, Vol. 86, No. 2 (2017).