4F1-5in

スペクトルデータの潜在的ダイナミクス抽出

Extraction of latent dynamics from time-series spectral data

村田 伸 *1　永田 賢二 *1　岡田 真人 *1*2
Shin MURATA　Kenji NAGATA　Masato OKADA

*1東京大学大学院新領域創成科学研究科
Graduate School of Frontier Sciences, The University of Tokyo

*2独立行政法人理化学研究所脳科学総合研究センター
RIKEN Brain Science Institute

In a broad range of fields, spectral data is obtained from spectroscopy. Spectral data have complex structure. Spectral decomposition is a method to fit each peak of data to a unimodal basis function. Center, width and amplitude of each peak reflect the nature of the subject. In recent years, time-series spectral data is obtained. However, we usually analyze the data independently. In this research, we propose the method to analyze time-series spectral data by using Bayesian inference, and validate its efficacy by using synthetic data.

1. 序論

様々な科学分野で分光計測からスペクトルデータが得られている．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，強度に対象の性質が反映されている．そのため，スペクトルデータを単峰性の基底関数でフィッティングし，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，重要な手法である [Nagata 12]．

近年，スペクトルデータが時間的に計測された，時系列スペクトルデータが得られている．例えば，物性科学における時間分解 X 線光電子分光法や天文学におけるブラックホール観測が挙げられる．時間分解 X 線光電子分光法では，対象の物質で起きている化学反応を追跡することが可能である [Nugent-Glandorf 01]．また，ブラックホールに物質が吸い込まれる際に発する，短時間スケールで変化する光を観察することでブラックホールを観測することができる [Celotti 99]．このように，時系列スペクトルデータからその観測対象の背後にあるダイナミクスを抽出することは，広い分野にまたがる重要な課題である．

しかしながら，時系列構造を考慮したスペクトル分解手法は開発されておらず，各時刻で独立にスペクトル分解を行う方法が主流である．本研究では，時間構造を考慮したスペクトル分解手法を提案する．人工データを用いて，その性能を検証し，従来の各時刻独立な解析を行うより，性能が高いことを人工データを用いて示した．

本原稿は全 4 章で構成されている．2 章では提案する推定手法を説明する．3 章では人工データを用いて推定手法の有効性を検証する．4 章で得られた結果をまとめ，今後の展望を述べる．

2. 確率的定式化

2.1 時系列構造を考慮したスペクトル分解

本研究では，図 1(a) に示すような，時系列スペクトルデータを統合的に取り扱うベイズ推論の枠組みを提案する．まず，同時確率分布 $p(\boldsymbol{Y}, \Theta, \boldsymbol{W}, \boldsymbol{m})$ を考える．ここで，$\boldsymbol{W} = \{w_{k,\tau}\}$,

連絡先: 岡田真人 okada@k.u-tokyo.ac.jp

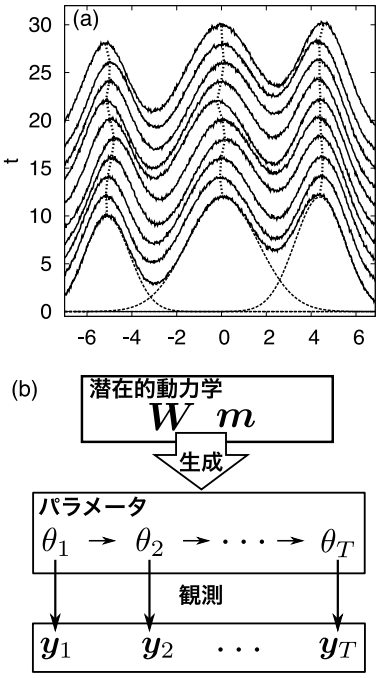

図 1: 本研究で考慮する階層構造．潜在的動力学からパラメータの時系列が生成され，それらのパラメータに従いスペクトルデータが観測される．

$\boldsymbol{m} = \{m_k\}$ は潜在的動力学を表すパラメータとする．また，各時刻でのスペクトルデータのピークを表すパラメータを $\theta = \{a_{k,t}, \mu_{k,t}, \sigma_{k,t}\}_{k=1}^{K}$ をとし，全時刻のパラメータセットを $\Theta = \{\theta_t\}_{t=1}^{T}$ とする．ここで，$a_{k,t}$ はピークの強度，$\mu_{k,t}$ はピーク中心，$\sigma_{k,t}$ はピークの幅を表す．各時刻で観測データ $\boldsymbol{y}_t = (y_{1t}, \cdots, y_{Nt})$ とし，時系列スペクトルデータを $\boldsymbol{Y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_T)$ とする．

図 1(b) に示すような生成・観測プロセスを考慮する．すなわち，$\boldsymbol{W} = \{w_{k,\tau}\}$，$\boldsymbol{m} = \{m_k\}$ は独立に生成されると考え，$p(\boldsymbol{W}, \boldsymbol{m}) = p(\boldsymbol{W})p(\boldsymbol{m})$ である．スペクトルデータのパラメー

タセット Θ は，$\boldsymbol{W}$ と，$\boldsymbol{m}$ から生成され，$p(\Theta \mid \boldsymbol{W}, \boldsymbol{m})$ である．パラメータセット Θ が与えられたとき，全スペクトルデータ $\boldsymbol{Y}$ が観測される確率は，$p(\boldsymbol{Y} \mid \Theta)$ と表される．従って，同時確率分布は，

$$p(\boldsymbol{Y}, \Theta, \boldsymbol{W}, \boldsymbol{m}) = p(\boldsymbol{Y} \mid \Theta) p(\Theta \mid \boldsymbol{W}, \boldsymbol{m}) p(\boldsymbol{W}) p(\boldsymbol{m}) \quad (1)$$

となる．従来のスペクトル分解 [Nagata 12] を独立に T 回行うことは，同時確率分布で $p(\boldsymbol{Y}, \Theta) = \prod_t p(\boldsymbol{y}_t, \theta_t)$ と表され，$\boldsymbol{W}$ と $\boldsymbol{m}$ が存在せず，時間構造を考慮していないことが分かる．

本研究では，ピーク中心 $\{\mu_{k,t}\}$ が自己回帰モデル（AR モデル）

$$\mu_{k,t} = \sum_{\tau=1}^{d} w_{k,\tau} \mu_{k,t-\tau} + m_k + e_{k,t} \quad (2)$$

で生成されると考える．ここで，$w_{k,\tau}$ は，k 番目のピーク中心 $\mu_{k,t}$ が τ ステップ前のピーク中心 $\mu_{k,t-\tau}$ から受ける影響を表し，m_k は定数の入力，$e_{k,t}$ は $\mathcal{N}(0, \sigma_{AR}^2)$ の正規分布に従うノイズである [Akaike 69]．また，ピークの強度 $a_{k,t}$，幅 $\sigma_{k,t}$ は時間変化せず一定であるとする．このとき，$\boldsymbol{W}$，$\boldsymbol{m}$ が与えられたときの，パラメータセットの条件付き確率は，

$$p(\Theta \mid \boldsymbol{W}, \boldsymbol{m}) = p(a_k) p(\sigma_k) p(\{\mu_{k,t}\} \mid \boldsymbol{W}, \boldsymbol{m}) \quad (3)$$

となる．ピーク中心の時系列の条件付き確率 $p(\{\mu_{k,t}\} \mid \boldsymbol{W}, \boldsymbol{m})$ は，式 (2) のノイズ $e_{k,t}$ が正規分布に従うとき，二乗和誤差関数

$$E_{AR} = \frac{1}{2KT} \sum_{k=1}^{K} \sum_{t=1}^{T} \left| \mu_{k,t} - \left(\sum_{\tau=1}^{d} w_{k,\tau} \mu_{k,t-\tau} + m_k \right) \right|^2 \quad (4)$$

を考えると，次のボルツマン分布で表される．

$$p(\{\mu_{k,t}\} \mid \boldsymbol{W}, \boldsymbol{m}) \propto \exp\left[-\frac{KT}{\sigma_{AR}^2} E_{AR} \right] \quad (5)$$

図 1(b) にあるように，各時刻 $t = 1, \cdots, T$ では，パラメータセット $\theta_t = \{a_k, \mu_{k,t}, \sigma_k\}_{k=1}^{K}$ が与えられた下で，スペクトルデータ $\boldsymbol{y}_t = (y_{1t}, \cdots, y_{Nt})^{\mathrm{T}}$ は次のように観測される．

$$y_{it} = f(x_i; \theta_t) + e_{it}, \quad (6)$$

$$f(x_i; \theta_t) = \sum_{k=1}^{K} a_k \phi(x_i; \sigma_k, \mu_{k,t}), \quad (7)$$

$$\phi(x_i; \sigma_k, \mu_{k,t}) = \exp\left[-\frac{1}{2\sigma_k^2} (x_i - \mu_{k,t})^2 \right] \quad (8)$$

各時刻での観測値と真の値の二乗和誤差

$$E_t(\theta_t) = \frac{1}{2N} \sum_{i=1}^{N} |y_{it} - f(x_i; \theta_t)|^2, (t = 1, \cdots, T) \quad (9)$$

ならびに，全時刻での二乗和誤差

$$E(\Theta) = \frac{1}{T} \sum_{t=1}^{T} E_t(\theta_t) \quad (10)$$

を考える．式 (6) のノイズ e_{it} が，$\mathcal{N}(0, \sigma_o^2)$ の正規分布に従うとき，各時刻での観測が独立であると仮定すると，パラメータセット Θ が与えられたときの全スペクトルデータ $\boldsymbol{Y}$ が観測される条件付き確率は，次のボルツマン分布に従う．

$$p(\boldsymbol{Y} \mid \Theta) = \prod_{t=1}^{T} p(\boldsymbol{y}_t \mid \theta_t) \propto \prod_{t=1}^{T} \exp\left[-\frac{N}{\sigma_o^2} E_t(\theta_t) \right] \quad (11)$$

$$\propto \exp\left[-\frac{NT}{\sigma_o^2} E(\Theta) \right] \quad (12)$$

以上の定式化から，スペクトルデータ $\boldsymbol{Y}$ が観測されたときの全時刻でのピークのパラメータ Θ，潜在的動力学構造を表す $\boldsymbol{W}$，$\boldsymbol{m}$，の事後確率は，

$$p(\Theta, \boldsymbol{W}, \boldsymbol{m} \mid \boldsymbol{Y}) \propto p(\boldsymbol{Y} \mid \Theta) p(\Theta \mid \boldsymbol{W}, \boldsymbol{m}) p(\boldsymbol{W}) p(\boldsymbol{m}) \quad (13)$$

$$\propto \exp\left[-\frac{NT}{\sigma_o^2} E(\Theta) \right] p(\Theta \mid \boldsymbol{W}, \boldsymbol{m}) p(\boldsymbol{W}) p(\boldsymbol{m}) \quad (14)$$

となる．この事後確率を計算することで，パラメータ Θ，ならびに潜在的時間構造を表す $\boldsymbol{W}$，$\boldsymbol{m}$ を推定する．

式 (14) の事後分布は一般に解析的に取り扱える形ではないため，レプリカ交換モンテカルロ法（REMC 法）を用いて，パラメータのサンプリングを行った．[Geyer 91, Hukushima 96].

2.2 ピーク数と AR モデルの次数のモデル選択

スペクトルデータをフィッティングするガウス関数の個数 K，および AR モデルの次数 d は，モデルの構造を決める重要なパラメータである．モデル (K, d) が変化すると，パラメータ $\{\theta_t\}$，$\boldsymbol{W}$，$\boldsymbol{m}$ も変化する．そのため，データから (K, d) を客観的に決定するモデル選択を行うことが必要である．

データ $\boldsymbol{Y}$ が与えられた元での，モデル (K, d) の周辺化事後確率は，

$$p(K, d \mid \boldsymbol{Y}) = \iiint p(\Theta, \boldsymbol{W}, \boldsymbol{m}, K, d \mid \boldsymbol{Y}) \mathrm{d}\Theta \mathrm{d}\boldsymbol{W} \mathrm{d}\boldsymbol{m} \quad (15)$$

$$\propto p(K, d) \iiint \mathrm{d}\Theta \mathrm{d}\boldsymbol{W} \mathrm{d}\boldsymbol{m} \exp\left[-\frac{NT}{\sigma_o^2} E(\Theta) \right] \times p(\Theta \mid \boldsymbol{W}, \boldsymbol{m}, K, d) p(\boldsymbol{W} \mid K, d) p(\boldsymbol{m} \mid K) \quad (16)$$

となる．式 (16) 中の積分の負の対数を取った自由エネルギー

$$F(K, d) = -\log \iiint \mathrm{d}\Theta \mathrm{d}\boldsymbol{W} \mathrm{d}\boldsymbol{m} \exp\left[-\frac{NT}{\sigma_o^2} E(\Theta) \right] \times p(\Theta \mid \boldsymbol{W}, \boldsymbol{m}, K, d) p(\boldsymbol{W} \mid K, d) p(\boldsymbol{m} \mid K) \quad (17)$$

を考える．モデルの事前分布 $p(K, d)$ が一様分布であるとき，周辺事後確率最大化は自由エネルギーの最小化と等価になる．本研究では自由エネルギー最小化で (K, d) のモデル選択を行う．

式 (17) の多重積分は一般に困難であるため，REMC 法を用いて，数値的に積分を行った [Nagata 12].

3. 結果

本研究では提案手法の有用性を検証するため，人工データによる推定を行った．本章ではその結果について述べる．

3.1 数値実験条件

真のモデルとして $(K, d) = (2, 1)$ を考える．AR モデルのパラメータとして，$(w_{1,1}, w_{2,1}) = (-0.35, 0.35)$，$(m_1, m_2) = (1.0, -1.0)$ とし，式 (2) に従い，100 ステップの $\{\mu_{k,t}\}$ を生成した．このとき，$\sigma_{AR} = 1.0$ としている．

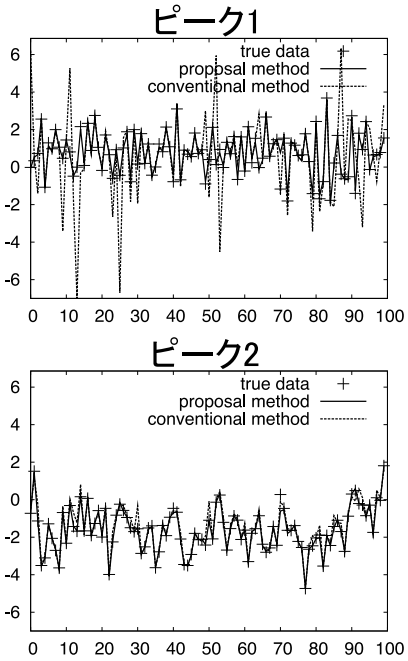

図 2: 各ピーク中心 $\{\mu_{k,t}\}$ の推定結果と，真の値の比較．マークが真の値を表し，実線が提案法による推定値を，点線が従来法による推定値を表す．

	提案法	従来法
k=1	0.0259	1.9032
k=2	0.0059	0.0463

表 1: 真のピーク中心の時系列 $\{\hat{\mu}_{k,t}\}$ と，推定されたピーク中心の時系列 $\{\mu_{k,t}\}$ の間の二乗和誤差．

各時刻 t において，スペクトルデータ $\boldsymbol{y}_t$ を生成する．ガウス関数 ϕ_k のパラメータは $(a_1, a_2) = (1.0, 2.0)$，$(\sigma_1, \sigma_2) = (0.816, 1.0)$ とした．ガウス関数に加算されるノイズの大きさは $\sigma_o = 0.22$ としている．また，ガウス関数の入力は $-7.0 \leq x_i \leq 6.86$ の範囲で等間隔に $N = 100$ 点用いた．

生成したデータを用いて，パラメータ推定ならびにモデル選択を行う．ここで，各パラメータについて，事前分布はそれぞれ $p(a_k) \in [0.00, 3.53]$，$p(\sigma_k^{-2}) \in [0.10, 100]$，$p(w_{k,\tau}) \in [-0.50, 0.50]$，$p(m_k) \in [-7.0, 6.86]$ の一様分布としている．σ_k に関しては，その二乗の逆数を推定するパラメータとする．また，モデル (K, d) は，ピーク数 $K = 1, 2, 3$，AR モデルの次数 $d = 0, 1, 2$ とし，9 通りのモデルを考え，それらのモデルは一様分布を事前分布として考える．パラメータをサンプリングするにあたり，最初の 10000 モンテカルロステップは burn-in にし，50000 モンテカルロステップでサンプリングをし，パラメータ推定・モデル選択を行った．

3.2 数値実験結果

まず，真のモデルと同じ $(K, d) = (2, 1)$ の条件下でパラメータの推定を行う．

図 2 は，ピーク中心 $\{\mu_{k,t}\}$ に関する推定結果である．2 つのピークについてそれぞれ真の時系列データと従来手法，提案手法を比較している．マークが真の値である．実線が提案手法による推定値であり，点線が従来手法による推定値を表している．提案手法により，真の時系列が推定できていることが分かる．真のピーク中心の時系列 $\hat{\mu}_{k,t}$ と，推定されたピーク中心の時系列の間の二乗和誤差 $E_k = (2T)^{-1} \sum_t |\hat{\mu}_{k,t} - \mu_{k,t}|^2$ を表 1 に示している．提案手法の時間構造を考慮してピーク中心を推定する方が，各時刻で独立にピーク中心を推定するより良い性能であることが分かる．

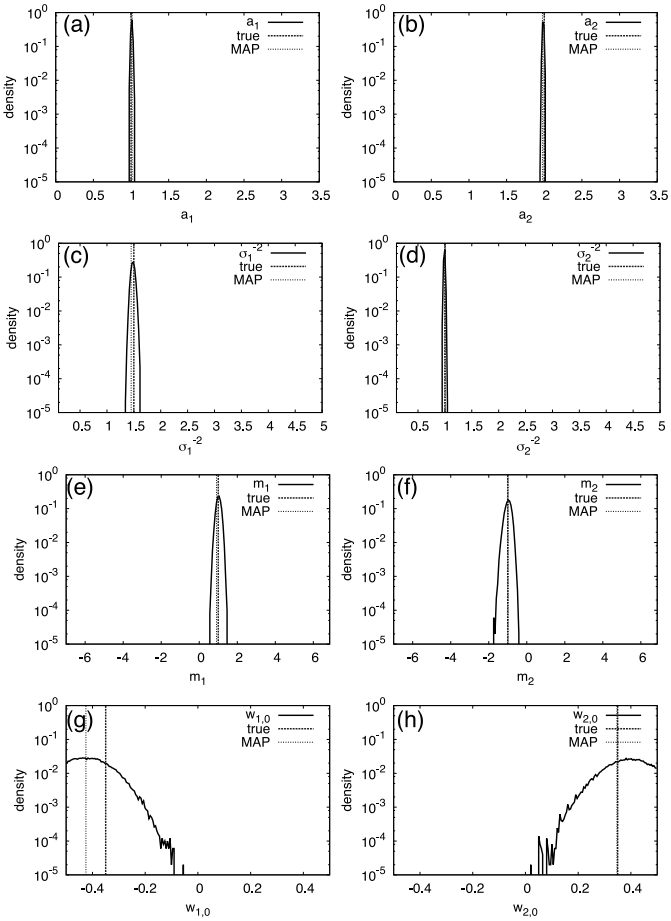

図 3: REMC 法によりサンプリングされた各パラメータのヒストグラム．横軸が各パラメータ，縦軸が度数分布の対数プロットである．実線がサンプリングされた分布，太点線が真の値，点線が事後確率最大となるような推定値である (a)(b) がそれぞれのピークのガウス関数の強度，(c)(d) がそれぞれのピークのガウス関数の分散の逆数である．(e)(f) がそれぞれ AR モデルの定数項に対応する．(g)(h) がそれぞれ AR モデルの係数に対応する．

図 3(a)–(d) は，ガウス関数の強度 $\{a_k\}$ と，精度 $\{\sigma_k^{-2}\}$ の周辺事後分布を度数分布で表している．横軸が各パラメータの値，縦軸が度数分布の対数プロットである．各図において，実線がサンプリングされた分布，太点線が真の値，点線が事後確率最大となるような推定値である．いずれのパラメータも，一様な事前分布と比較して，真の値周辺で急峻にピークを持ち，さらに，事後分布を最大にする MAP 解の値も真の値と一致していることが分かる．提案手法を用いて，パラメータ推定を精度よく推定できることが分かる．

図 3(e)–(h) は，潜在的動力学を表すパラメータ $\{w_{k,\tau}\}$ ならびに $\{m_k\}$ の周辺事後分布を度数分布で表している．各図において，実線がサンプリングされた分布，太点線が真の値，点線が事後確率最大となるような推定値である．横軸が各パラメータの値，縦軸が度数分布の対数プロットである．図 3(e)(f)

は，それぞれ定数項 m_1，m_2 の結果を示している．事前分布と比較して真の値周辺でピークを持ち，さらに，MAP 解も真の値と良く一致していることが分かる．しかしながら，スペクトル分解のパラメータと $\{a_k\}$ や，$\{\sigma_k\}$ と比較して，事後分布が広がっており，推定精度にばらつきがあることが分かる．図 3(g)(h) は，それぞれ係数 $w_{1,1}$，$w_{2,1}$ の結果を示している．事前分布 $w_{k,\tau} \in [-0.5, 0.5]$ の一様分布と比較すると，真の値周辺でサンプリングされているが，他のパラメータと比較すると．$w_{1,1}$，$w_{2,1}$ の事後分布は推定精度にばらつきがあることが分かる．これは，$w_{k,\tau}$ がより深い構造のパラメータであるためと考えられる．

これまでのパラメータ推定は，真のモデルである $(K, d) = (2, 1)$ を既知とした上で行ってきた．そこで，(K, d) をデータから客観的に決定することを考える．

表 2 に，REMC 法を元に自由エネルギー $F(K, d)$ の式 (17) を数値的に計算した結果を示す．このとき，候補となるモデルは $K = 1, 2, 3$，$d = 0, 1, 2$ の組み合わせで，9 通りのモデルを考えた．モデル $(K, d) = (2, 1)$ で自由エネルギー最小となり，真のモデルを正しく選択できたことが分かる．また，表 3 に自由エネルギーを元に計算した事後確率 $p(K, d \mid \boldsymbol{Y})$ の値を示している．真のモデルの事後確率が 59.6% であり，他のモデルと比較して高い確率であることが分かる．

以上の結果から，提案手法は時系列スペクトルデータのスペクトル分解，潜在的動力学の推定，さらにモデル選択を正しく行えることが分かった．

4. 考察・結論

本研究では，時系列スペクトルデータを，時間構造を考慮して解析するためのベイズ推論の枠組みを構築した．

従来，時系列スペクトル分解の解析は各時刻で独立にスペクトル分解を行い，時間構造は考慮されていなかった．本研究では，特にピーク中心が時間的に変動する場合を考え，パラメータに AR モデルから考えられる事前分布を導入し，時系列スペクトルデータから，スペクトル分解と時系列構造抽出を同時に行う手法を提案した．さらに，提案手法の有用性を人工データを用いて検証し，各時刻で独立にスペクトル分解を行う場合より，高い精度でピーク中心の時系列を推定できることを示した．

さらに，フィッティングするピークの個数ならびに AR モデルの次数という，推定するパラメータの数を規定するモデルをデータだけから客観的に決定する枠組みを，提案手法に関して開発し，実際に人工データで推定し有効性を検証した．

実計測データへの適用を目指し，時系列構造の導入の仕方を発展させることが今後の課題である．

謝辞

本研究の一部は文部科学省　科学研究費補助金新学術領域研究 [課題番号　25120009(岡田)]，基盤研究（C）[課題番号　25330283(永田)] の下で行われた．

参考文献

[Akaike 69] Akaike, H.: Fitting autoregressive models for prediction, *Annals of the institute of Statistical Mathematics*, pp. 243–247 (1969)

[Celotti 99] Celotti, A., Miller, J. C., and Sciama, D. W.: Astrophysical evidence for the existence of black holes, *Classical and Quantum Gravity*, Vol. 16, No. 12A, pp. A3–A21 (1999)

[Geyer 91] Geyer, C. J.: Markov chain Monte Carlo maximum likelihood, in *Proceedings of the 23rd Symposium on the Interface*, p. 156 (1991)

[Hukushima 96] 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)

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

[Nugent-Glandorf 01] Nugent-Glandorf, L., Scheer, M., Samuels, D. a., Mulhisen, a. M., Grant, E. R., Yang, X., Bierbaum, V. M., and Leone, S. R.: Ultrafast time-resolved soft x-ray photoelectron spectroscopy of dissociating Br2., *Physical review letters*, Vol. 87, No. 19, p. 193002 (2001)

	$d=0$	$d=1$	$d=2$
K=1	11718.591465	11711.832820	11711.363701
K=2	5555.847529	5540.367995	5541.263794
K=3	5558.059545	5541.980823	5543.034303

表 2: ピーク数 K，AR 次数 d と自由エネルギー F(K,d) の関係

	$d=0$	$d=1$	$d=2$
K=1	0%	0%	0%
K=2	0%	59.6%	24.3%
K=3	0%	11.9%	4.1%

表 3: ピーク数 K，AR 次数 d と事後確率 $p(K, d \mid \boldsymbol{Y})$ の関係