

分光透過率推定のための EM アルゴリズムによる 不完全サンプルの画素値補正

竹谷 勝*,**・津村 徳道*・羽石 秀昭*・三宅 洋一*

*千葉大学自然科学研究科 〒263-8522 千葉市稻毛区弥生町 1-33

**農業生物資源研究所遺伝資源第一部 〒305-8602 つくば市観音台 2-1-2

Pixel Value Correction of Incomplete Sample by EM Algorithm for Estimation of Transmittance Spectra

Masaru TAKEYA*,**, Norimichi TSUMURA*, Hideaki HANEISHI* and Yoichi MIYAKE*

*Graduate School of Science and Technology, Chiba University, 1-33, Yayoi-cho, Inage-ku, Chiba 263-8522

**Department of Genetic Resources I, National Institute of Agrobiological Resources, 2-1-2, Kannondai, Tsukuba, Ibaraki 305-8602

In a previous research, Wiener estimation based method was applied to estimate transmittance spectra of fungi from fungal optical micrograph taken by multiband camera. Transparency samples used to obtain Wiener estimation matrix, however, sometimes become incomplete due to the overflow of limited dynamic range of CCD camera. This leads to the decreased accuracy in the estimation of transmittance spectra for the pixels with high pixel value of multiband image. In this paper we propose to increase the accuracy of estimation by correcting the pixel values of the incomplete samples using EM algorithm. The effectiveness of the proposed method is shown through computer simulation.

1. はじめに

大量の生物遺伝資源を保存・管理しているジーンバンク¹⁾では、遺伝資源情報を利用者に伝える効果的手法の開発が重要な作業のひとつとなる。今まででは遺伝資源情報をテキスト形式によりいかに効率的に表現するかが研究対象であったが、インターネットが普及した現在では、テキスト形式による表現が難しい形態に関する情報に対して、画像データを利用するシステムの開発が行われている。とくに菌類は、形態的特徴が多様であるため、顕微鏡画像をデータベース化しておき、利用者からの要求に対して画像情報を提供するのが有効である。以上の背景から筆者らは、菌類の顕微鏡画像データベースシステムの構築をすすめている。

光学顕微鏡を通じた菌類の測色値は、光源の特性に大きく依存するため、菌類の色を正確に伝えることはできない。そこで画像データには、光源による影響を受けない分

光透過率を用いることにした。菌類の分光透過率分布は、菌類のマルチバンド画像から、Wiener 法^{2,3)}により推定できる⁴⁾。マルチバンド画像は、光学顕微鏡（対物レンズ 40 倍）にモノクロ CCD カメラ（1536×1024 pixels, 8 bits/pixel）を取り付けた装置で、コンデンサーレンズと光源の間に広帯域フィルターを 1 枚ずつ挿入して撮影した。菌類のマルチバンド撮影のつど、同じ条件下で透過色票サンプルの撮影を行い、画像中心部の矩形領域（696×644 pixels, 全画素数の約 30%）の平均画素値（以下、単に「透過色票サンプルの画素値」と呼ぶ）と、事前に測定した分光透過率から Wiener 推定行列を計算する。

本システムでは、画像中に複数の色票を入れることは技術的に困難で、かつ測定誤差の原因となるため、空間的に一様な透過色票サンプルを多数枚用意し、これを差し替えながら撮影する方法をとっている。1 回の撮影で 1 つの透過色票サンプルしか画像化できないので、透過色票サンプルの数を m 、フィルター数を n とすると、対象とする菌類の画像を含めて、撮影する画像の数は $(m+1)n$ になる。

**E-mail: katu@gene.affrc.go.jp

菌類および透過色票サンプルの画素値は、CCD カメラのダイナミックレンジ（以下、D レンジと略す）の影響を受けるため、今まで撮影のつど、すべての画素値がレンジ内に収まるよう手動で露光量を調整していた。Fig. 1

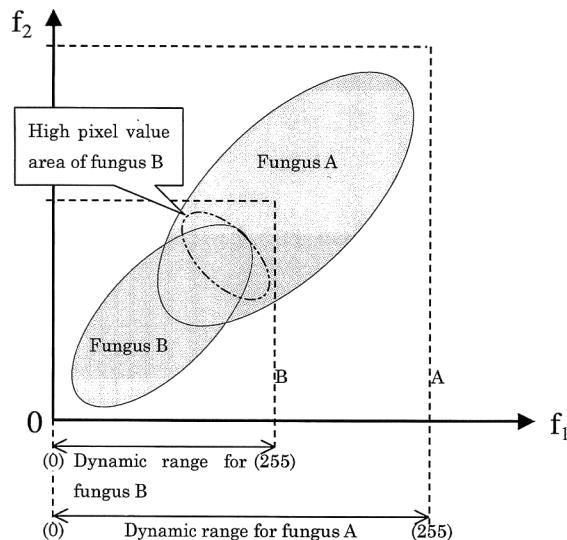


Fig. 1. Schematic diagram of fungal pixel value distribution and dynamic range.

に、2種類の菌類 A と B の画素値分布と D レンジの関係を模式的に示す。破線 A は、菌類 A を撮影対象として露光量を調整した場合の D レンジを表している。 f_1 , f_2 は異なる 2 バンドの画素値である。この露光条件のまま対象を菌類 B とすると、全体的に低画素値しか得られない。そこで、D レンジを有効に活用するために露光量を増やすことになる。この場合の D レンジが破線 B である。しかし、このような調整により D レンジを超える透過色票サンプルが現れ、そのため Fig. 1 中の 2 点鎖線内の高画素値領域に対応する画素の分光透過率の推定精度が低下すると考えられる。

本論文では、D レンジを超えるサンプルの画素値を自動的に補正する方法を提案する。その結果、対象とする菌類だけを適正画素値領域に収まるように調節すれば、マルチバンド画像の撮影が可能となり、処理の効率化が図られる。画素値の補正には expectation-maximization (EM) アルゴリズムを用いる⁵⁻⁷⁾。これは、欠測値などを含む不完全データから、反復計算によりそのパラメータを最尤推定する手法である。Fig. 2 に提案法の概要を示す。

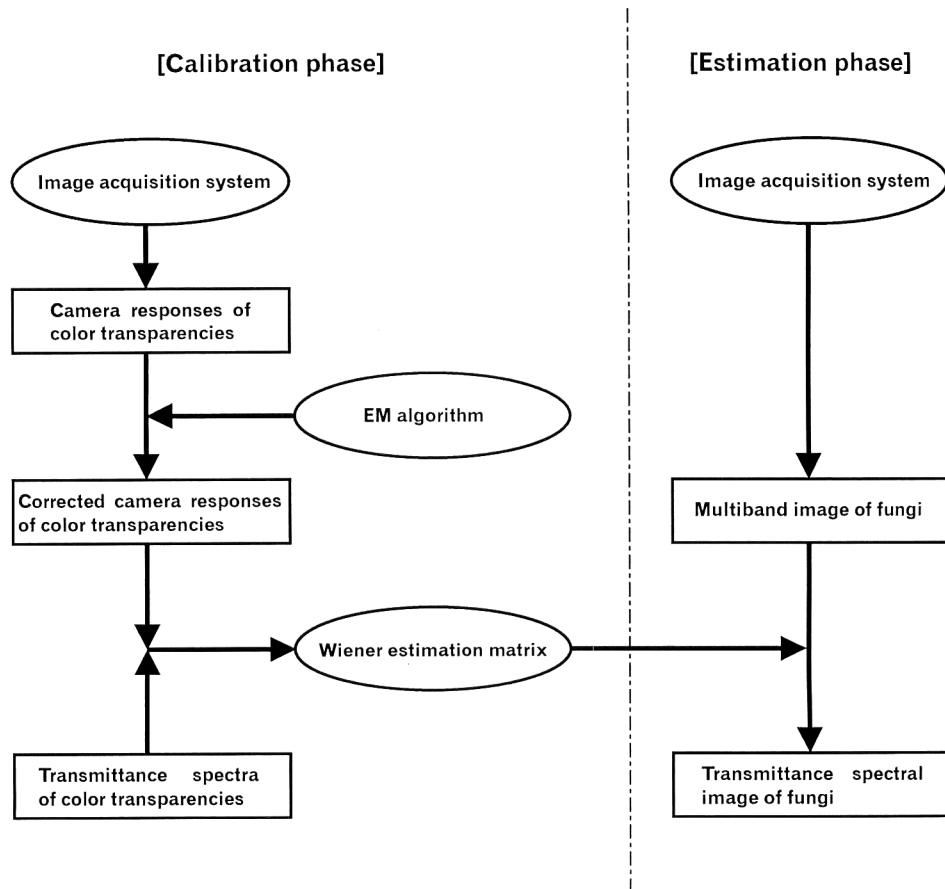


Fig. 2. Schematic diagram of proposed method to estimate the transmittance spectra of fungi.

2. 不完全サンプルの補正アルゴリズム

2.1 不完全サンプル

はじめに、任意の光量を線形に測定できる CCD カメラを仮定し、その出力値を「理想的な画素値」と呼ぶことにする。一方、現実の計測では D レンジにより測定限界が規定される。たとえば量子化レベル 8 bits の CCD カメラでは、画素値 255 を超える光量は D レンジにより 255 に打ち切られ出力される。このような「理想的な画素値」と異なる出力値を不良値と呼び、マルチバンド撮影の出力値が 1 つでも不良値となる透過色票サンプルを不良サンプルと呼ぶこととする。不完全サンプルとは、不良サンプルを含む透過色票サンプルの集合を意味する。Fig. 3 に不完全サンプルの模式図を示す。“×”はそれぞれ透過色票サンプルに対するある 2 バンドの画素値をプロットしたものである。右上の 2 つの “×” に対応する透過色票サンプルは、D レンジにより「理想的な画素値」 f_1 が 255 にシフトされ、不良サンプルとなることを表している。

実際の菌類のマルチバンド画像を多数取得して画素値の分布を調べたところ、任意の 2 バンド間でそれぞれ正規分布に従っているように推察された。このことから、母集団である菌類の「理想的な画素値」は多次元正規分布でモデル化できると考えられた。そのため菌類の分光特性を模擬する透過色票サンプルの「理想的な画素値」も多次元正規分布に従う必要がある。

今回の研究では、透過色票サンプルを以下のように準備した。まず OHP シートおよびリバーサルカラーフィルムを媒体とし、それぞれ銀塩写真方式カラープリンター、フィルムレコーダーを用いて、1 枚ずつシアン (C), マゼンタ (M), イエロー (Y) の各濃度をランダムに与えたサンプルを数十枚作成する。次にバンド間の画素値分布図の上で正規分布に近づくよう、分布から外れていると思われる透過色票サンプルを除いて、最終的なサンプルを選定した。その結果、OHP シート 25 枚、リバーサルカラーフィルム 11 枚の計 36 枚を今回の透過色票サンプルとした。

2.2 EM アルゴリズム

EM アルゴリズムは、E (expectation) ステップと M (maximization) ステップで構成される処理を反復することによって、尤度を最大にする解を得るアルゴリズムである。以下ではまず、EM アルゴリズムの一般的な手順を述べる。

いま、2 つの標本空間 X と Y を考え、X を理想的な観測値ベクトル \mathbf{x} (完全データ) の空間、Y を現実に観測されるベクトル \mathbf{y} (不完全データ) の空間とする。ここに X から Y への写像 $\mathbf{x} \rightarrow \mathbf{y}(\mathbf{x})$ が存在し、そのような \mathbf{x} の集

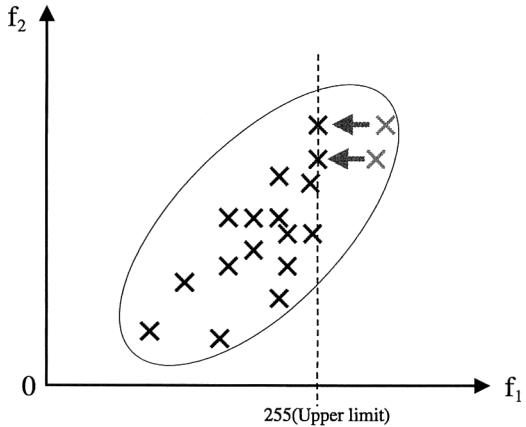


Fig. 3. Schematic diagram of incomplete sample.

合は $X(\mathbf{y}) = \{\mathbf{x} | \mathbf{y} = \mathbf{y}(\mathbf{x})\}$ として表される。 \mathbf{x} の確率密度関数を規定するパラメーターをベクトルで表し θ とする。 θ が与えられたときの \mathbf{x} の確率密度関数を $f(\mathbf{x}|\theta)$ とすれば、Y 上での \mathbf{y} の確率密度関数は

$$g(\mathbf{y}|\theta) = \int_{X(\mathbf{y})} f(\mathbf{x}|\theta) d\mathbf{x} \quad (1)$$

と表すことができる。ここで観測値ベクトル \mathbf{y} が与えられたときの θ の最尤推定を、EM アルゴリズムで行う。E ステップでは、観測値ベクトル \mathbf{y} と、現時点 (s 回目とする) での推定パラメーター $\theta^{(s)}$ を用いて、完全データの対数尤度関数 $\log f(\mathbf{x}|\theta)$ の期待値 $E(\log f(\mathbf{x}|\theta) | \mathbf{y}, \theta^{(s)})$ を計算し、M ステップでは、この期待値を最大にする θ を求め新たなパラメーター $\theta^{(s+1)}$ とする。この処理をパラメーターが収束するまで繰り返す。

本研究の解法では、正規分布が正則な指数分布族に属するため、簡略化された EM アルゴリズムの定式⁵⁻⁹⁾ (付録) を用いることができる。以下に解法を示す。

$\mathbf{Y} = (Y_1, \dots, Y_K)^T$ は、平均ベクトル $\mu = (\mu_1, \dots, \mu_K)^T$ 、共分散行列 Σ (i 行 j 列の要素を σ_{ij} と書く) の K 次元正規分布に従う確率変数ベクトルとする。 T は転置記号である。この問題の場合、パラメータベクトルは平均ベクトル μ および共分散行列 Σ の各要素からなるベクトルであり $\theta = (\mu_1, \dots, \mu_K, \sigma_{11}, \dots, \sigma_{KK})^T$ と書ける。透過色票サンプルの K 次元マルチバンド画像の画素値は、この K 次元正規分布に従う。すなわち、 i 番目の透過色票サンプルに対する画素値ベクトル $\mathbf{y}_i = (y_{i1}, \dots, y_{iK})^T$ は \mathbf{Y} から出た i 番目の観測値ベクトルである。ここで n 個の透過色票サンプルの K バンド撮影による nK 個の画素値をまとめて nK 次ベクトル $\mathbf{w} = (y_{11}, \dots, y_{1j}, \dots, y_{nK})^T$ とし、そのうち D レンジの影響を受けて欠測値とみなされる画素値だけを要素とするベクトルを \mathbf{w}_{mis} 、その他の画素値を要素とす

るベクトルを \mathbf{w}_{obs} と表すことにする。また、定式を解くうえで必要となる多次元正規分布の十分統計量を本手法では $t_j = \sum_{i=1}^n y_{ij}, j=1, \dots, K$ と $t_{jk} = \sum_{i=1}^n y_{ij}y_{ik}, j,k=1, \dots, K$ とし、これら 1 次と 2 次の統計量をまとめたものをベクトル $\mathbf{t} = (t_1, \dots, t_k, t_{11}, \dots, t_{KK})^T$ とする。

$s+1$ 回目の E ステップでは、 s 回目の反復処理によるパラメーターベクトル $\boldsymbol{\theta}^{(s)} = (\mu_1^{(s)}, \dots, \mu_K^{(s)}, \sigma_{11}^{(s)}, \dots, \sigma_{KK}^{(s)})^T$ を用いて、十分統計量を次式により推定する。

$$t_j^{(s)} = E(t_j | \mathbf{w}_{obs}, \boldsymbol{\theta}^{(s)}) = \sum_{i=1}^n y_{ij}^{(s)}, \quad j=1, \dots, K \quad (2)$$

$$t_{jk}^{(s)} = E(t_{jk} | \mathbf{w}_{obs}, \boldsymbol{\theta}^{(s)}) = \sum_{i=1}^n (y_{ij}^{(s)} y_{ik}^{(s)} + c_{jki}^{(s)}), \quad j,k=1, \dots, K \quad (3)$$

ここで、

$$y_{ij}^{(s)} = \begin{cases} y_{ij}, & \text{if } y_{ij} \text{ is observed;} \\ E(y_{ij} | \mathbf{w}_{obs,i}, \boldsymbol{\theta}^{(s)}), & \text{if } y_{ij} \text{ is missing.} \end{cases} \quad (4)$$

$c_{jki}^{(s)} = \begin{cases} 0 & \text{if } y_{ij} \text{ or } y_{ik} \text{ are observed;} \\ Cov(y_{ij}, y_{ik} | \mathbf{w}_{obs,i}, \boldsymbol{\theta}^{(s)}), & \text{if } y_{ij} \text{ and } y_{ik} \text{ are missing.} \end{cases}$ であり、 $Cov(y_{ij}, y_{ik} | \mathbf{w}_{obs,i}, \boldsymbol{\theta}^{(s)})$ は $\mathbf{w}_{obs,i}$ と $\boldsymbol{\theta}^{(s)}$ が与えられたときの y_{ij} と y_{ik} の条件付共分散を意味する。 $\mathbf{w}_{obs,i}$ は i 個目の観測値ベクトル $\mathbf{y}_i = (y_{i1}, \dots, y_{iK})^T$ の要素のうち、欠測値ではない画素値だけを要素としてもつベクトルを示す。

M ステップでは、パラメーターの更新が以下のようになされる。

$$\mu_j^{(s+1)} = t_j^{(s)} / n, \quad j=1, \dots, K \quad (5)$$

$$\sigma_{jk}^{(s+1)} = (t_{jk}^{(s)} - t_j^{(s)} t_k^{(s)} / n) / n, \quad j,k=1, \dots, K \quad (6)$$

以上のステップを、パラメーターベクトルのすべての要素 θ_i に関して、 $|\theta_i^{(s+1)} - \theta_i^{(s)}| / |\theta_i^{(s)}| < \delta$ 、ただし $\delta = 0.001$ が成立するまで繰り返し、収束後の $\boldsymbol{\theta}^{(s+1)}$ を用いて、式 (4) に従って最終的な補正值を出力する。

以上は本研究で行う画素値補正の基本手法であるが、実際の適用に際しては若干の改良を加えている。このことについて次章で述べる。

3. シミュレーション実験および補正法の改良

3.1 シミュレーションの方法

2.1 節で述べたように C, M, Y 濃度から作成した 36 枚の透過色票サンプルをもとにシミュレーション実験を行った。分光光度計により各透過色票サンプルの分光透過率を 410 nm から 700 nm まで 10 nm 間隔 (30 波長) で測定し、この分光透過率データから、理想的な CCD カメラを用いたマルチバンド撮影で撮られる画素値を計算し、それ

Table 1. An example of pixel value of incomplete sample at upper limit value 220.

	Broadband filter					
	BP42	BP45	BP53	BP55	BP60	SC64
Ideal value	79	127	126	232	151	96
Recorded value	79	127	126	220	151	96
Estimated value	79	127	126	227.05	151	96

を「理想的な画素値」とした。次に不完全サンプルをシミュレートして、EM アルゴリズムによる推定法での画素値補正を行い、最後に、補正された画素値から Wiener 推定行列を算出して透過色票サンプルの分光透過率を推定した。マルチバンド撮影におけるバンド数は 6 と設定した。これは顕微分光計で測定した菌類 (*Trichoderma* 属^{10,11)}) の分光透過率を主成分分析したところ、第 6 主成分まで累積寄与率が 99.99% に達したことを根拠としている。

画素値 y_{ij} は次式により算出することができる。

$$y_{ij} = \mathbf{f}_j^T \mathbf{E} \mathbf{S} \mathbf{o}_i, \quad i=1, \dots, 36, \quad j=1, \dots, 6 \quad (7)$$

ここで、 $\mathbf{f}_j, j=1, \dots, 6$ は 6 枚の広帯域フィルター Fuji BP42, BP45, BP53, BP55, BP60, SC64 の各分光透過率を表す 30 次元ベクトル、 \mathbf{E} と \mathbf{S} は、それぞれ標準光源 D₆₅ の分光放射分布と典型的な CCD カメラの分光感度特性を対角成分にもつ 30 × 30 対角行列、 $\mathbf{o}_i, i=1, \dots, 36$ は透過色票サンプルの分光透過率を表す 30 次元ベクトルである。

実際の計測では、255 を超える画素値を測定することはできないので、シミュレーション実験では、便宜的に 255 未満の画素値を D レンジの上限値と仮定した。そのため上限値を超える画素値はすべてオーバーフローとみなされ上限値に置き換えられる。上限値を 220 とした場合の例を Table 1 に示す。例示したデータは、あるサンプルの 6 バンド画素値であり、BP55 フィルターでの画素値 (灰色部) が補正の対象となる。「理想的な画素値 (Ideal value)」は 232 であるが、上限値が 220 であるため不良サンプルとなり、220 (Recorded value) に置き換えられる。また、この値は以下で紹介する EM アルゴリズムを用いた推定法により 227.05 (Estimated value) と推定される。

3.2 画素値補正法の改良

3.2.1 Method 1

まず 2.2 節で紹介した画素値補正の基本手法にしたがって推定した結果について述べる。

Fig. 4 に透過色票サンプルの画素値のうち上位 3 位までの高い画素値 (255, 232, 206) を挙げて、上限値と推定画素値との対応を示した。各記号は推定画素値、破線は

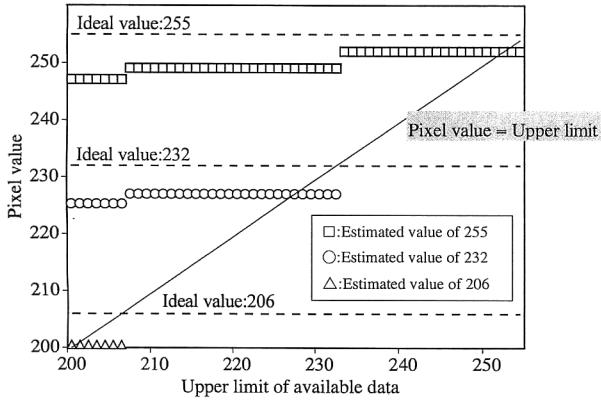


Fig. 4. Pixel value estimation using method 1.

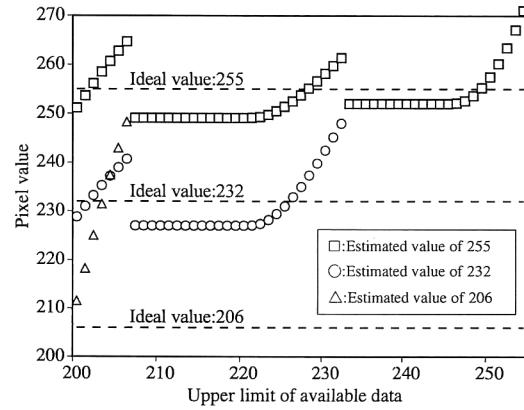


Fig. 5. Pixel value estimation using method 2.

「理想的な画素値」、横軸は上限値、縦軸は画素値を表す。推定は上限値以上の画素値を欠測値とみなして 2.2 節の EM アルゴリズムの処理手順により行った。上限値が 200 から 206 の区間では 3 つの画素値すべてが上限値以上のため推定対象となるが、207 から 232 の区間では 206 が上限値未満であるため推定対象とならない。上限値 233 以上では 255 だけが推定対象となる。またこれら区間内では推定のための条件が同一であるため推定結果は一定値となる。Fig. 4 から「理想的な画素値」をおおむね推定できたことがわかる。しかし“推定画素値=上限値”を表す実線の下側に現れる推定画素値は上限値未満を意味し、「理想的な画素値」は上限値以上であるという物理的知見に矛盾する。

3.2.2 Method 2

上述の物理的知見を考慮して、推定対象となる画素値を上限値で打ち切られた値とみなし、式(4)の条件付期待値の条件項に「 $y_{ij} \geq$ “上限値”」を加えて推定を行った。すなわち式(4)を

$$y_{ij}^{(s)} = \begin{cases} y_{ij}, & \text{if } y_{ij} \text{ is observed;} \\ E(y_{ij} | \mathbf{w}_{obs,i}, \boldsymbol{\theta}^{(s)}, y_{ij} \geq U), & \text{if } y_{ij} \text{ is missing.} \end{cases} \quad (8)$$

と修正する。ここで U は上限値である。この結果を Fig. 5 に示す。推定画素値は上限値を上げると徐々に「理想的な画素値」に近づくが、「理想的な画素値」を超えて増加し続けてしまう。これは、上限値 U が「理想的な画素値」に近い場合、正規分布の特性のため、すそ野領域 $y_{ij} \geq U$ で計算される期待値が「理想的な画素値」を超えるためである。

3.2.3 Method 3

以上の結果を勘案して、安定的に精度よく画素値を推定するには、推定対象を欠測値とみなして EM アルゴリズム

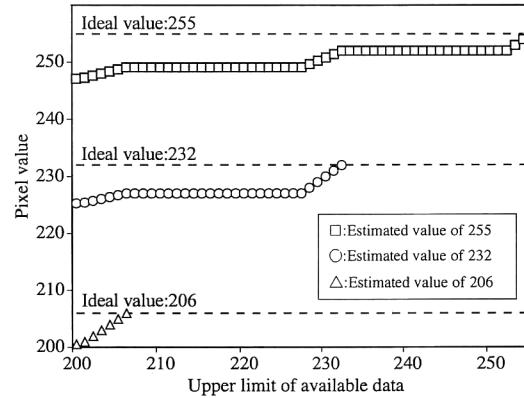


Fig. 6. Pixel value estimation using method 3.

を適用し、その計算過程で上限値未満の画素値が現れた場合には、上限値に置き換えるように矯正を施す方法が有効であると考える。すなわち式(4)を

$$y_{ij}^{(s)} = \begin{cases} y_{ij}, & \text{if } y_{ij} \text{ is observed;} \\ E(y_{ij} | \mathbf{w}_{obs,i}, \boldsymbol{\theta}^{(s)}), & \text{if } y_{ij} \text{ is missing and } E(y_{ij} | \mathbf{w}_{obs,i}, \boldsymbol{\theta}^{(s)}) \geq U; \\ U, & \text{if } y_{ij} \text{ is missing and } E(y_{ij} | \mathbf{w}_{obs,i}, \boldsymbol{\theta}^{(s)}) < U. \end{cases} \quad (9)$$

と修正する。Fig. 6 に推定結果を示す。物理的知見に反せず、推定画素値が「理想的な画素値」を超えて増加し続けることもなく、全上限値において良好であることがわかる。そこで、この推定法を本研究での最終的な提案法とする。

Fig. 7 に、上記 3 つの推定法による推定画素値と「理想的な画素値」との root mean square error (RMSE) を示す。また推定を行わなかったときとの比較のため、上限値と「理想的な画素値」との RMSE も併せて示す。RMSE は次式により計算した。

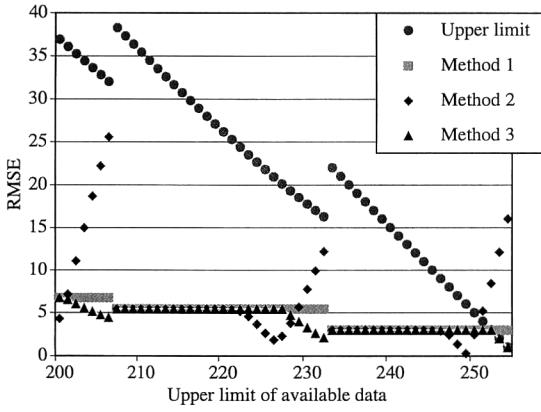


Fig. 7. RMSE between estimated pixel value and ideal value.

Table 2. RMSE between estimated transmittance spectra and measured at upper limit value 200. Case 1: Using color transparencies including upper limit, Case 2: Not using color transparencies including corrected pixel value, Case 3: Using color transparencies including corrected pixel value.

Case 1	Case 2	Case 3
0.147	0.089	0.073

$$RMSE = \sqrt{\frac{1}{N} \sum_{ij \in M} (y_{ij} - \hat{y}_{ij})^2} \quad (10)$$

ここで、 N は推定画素値の数、 M は推定対象となる画素値の添字 ij の集合、 y_{ij} は推定画素値（上限値と「理想的な画素値」との RMSE の場合は上限値）を表す。Fig. 7 から、推定を行うことによりほぼ全般に低い RMSE が得られ、推定の効果が表れていることがわかる。とくに提案法は、安定して低い RMSE となることから最も効果的な推定法であることが確認できる。

3.3 分光透過率推定精度の評価

提案法により推定された画素値を用いて Wiener 推定行列を計算し、その行列から Fig. 1 で示したような高画素値に対応する透過色票サンプルの分光透過率を推定した。先のシミュレーション実験で推定した 3 つの画素値はすべて BP55 フィルター由来である。BP55 フィルターを 4 番目のフィルターとすると、各透過色票サンプルの対応する画素値は $y_{i4}, i=1, \dots, 36$ で表すことができる。いま、 y_{i4} が i に対して単調減少するように番号付けされたとすると、推定画素値は $\hat{y}_{i4}, i=1, 2, 3$ と表せる。次に分光透過率の推定対象は、Fig. 1 に示すように上限値に近い画素値をもつサンプルとして、BP55 フィルター由來の画素値のうち上位 4 位から 6 位までの高画素値 $y_{i4}, i=4, 5, 6$ を要素にもつ 3 枚の透過色票サンプルとした。推定は以下の 3 通

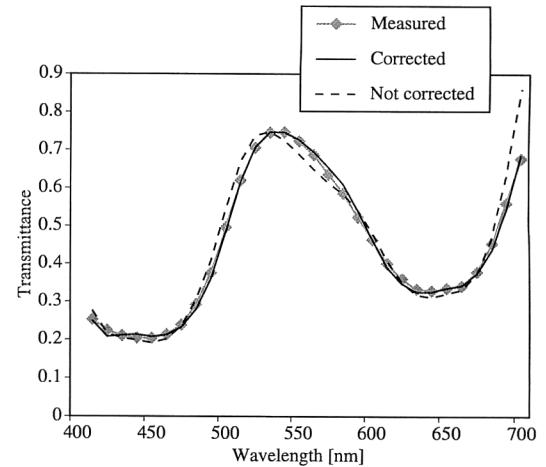


Fig. 8. Comparison between estimated and measured transmittance spectra. Example of upper limit value 200 was assumed.

りの方法で推定行列を作り行った。

[Case 1]

不完全サンプルを補正することなく推定を行うことを想定して、 $\hat{y}_{i4}, i=1, 2, 3$ を推定値ではなく上限値とし、推定対象を除いた 35 枚の透過色票サンプルから推定行列を作る。

[Case 2]

不完全サンプルの中で D レンジの影響を受けた不良サンプルを除いて推定を行うことを想定して、補正対象の 3 枚の透過色票サンプルも除く 32 枚で推定行列を作る。

[Case 3]

不完全サンプルの補正による推定を想定して、推定画素値、 $\hat{y}_{i4}, i=1, 2, 3$ を用いて 35 枚の透過色票サンプルから推定行列を作る。

Table 2 に上限値 200 での 3 通りの推定による、3 つの透過色票サンプルの分光透過率の実測値と推定値の RMSE を示す。Case 3 が最も低い RMSE であることから、高画素値を含む透過色票サンプルの分光透過率推定において、不完全サンプルの補正が有効であることが確認された。

例として、ある 1 つの透過色票サンプルの分光透過率の推定結果を Fig. 8 に示す。波長 410 nm から 700 nm の推定領域において、補正された透過色票サンプルを含めて推定した Case 3 による推定 (Corrected) のほうが、補正を施さない Case 1 による推定 (Not corrected) より実測分光透過率 (Measured) に類似していることがわかる。

4. おわりに

菌類顕微鏡画像の分光透過率推定において、マルチバンド画像に高画素値を含む画素の推定精度を向上させるため、EMアルゴリズムを用いた不完全サンプルの画素値補正法を提案した。シミュレーション実験により、提案法による画素値補正の効果を検証し、高画素値を含む透過色票サンプルの分光透過率の推定精度の向上を確認した。

今回のシミュレーション実験ではパラメーターが収束して推定値を得ることができたが、提案法はEMアルゴリズムの本来の仕様とは異なるので、パラメーターの収束性が保証されなくなることが懸念される。しかし、推定精度を上げるための現時点での最適法なので、優先的に適用し、パラメーターが収束しない場合に限り、他の推定法を用いることにより対応すればよいと考える。

今後は、推定精度をさらに向上させるための補正アルゴリズムの改良と、実際の顕微鏡画像に対する補正効果の確認実験を行う予定である。提案する画素値補正法は一般性をもっており、今後、画像上のノイズ補正など他のアプリケーションへの適用なども検討したい。

文 献

- 1) D. L. Plucknett, N. J. H. Smith, J. T. Williams and N. M. Anishetty: *Gene Banks and the World's Food* (Princeton University Press, Princeton, 1987).
- 2) H. Haneishi, T. Hasegawa, N. Tsumura and Y. Miyake: "Design of color filters for recording artworks," *IS&T's 50th Annual Conference* (Society for Imaging Science and Technology, Cambridge, 1997) pp. 369-372.
- 3) M. J. Vrhel and H. J. Trussell: "Filter considerations in color correction," *IEEE Trans. Image Process.*, **3** (1994) 147-161.
- 4) M. Takeya, N. Tsumura, H. Haneishi and Y. Miyake: "Estimation of transmittance spectra from multiband micrographs of fungi and its application to segmentation of conidia and hyphae," *Appl. Opt.*, **38** (1999) 3644-3650.
- 5) A. P. Dempster, N. M. Laird and D. B. Rubin: "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Stat. Soc. B*, **39** (1977) 1-38.
- 6) 宮川雅巳: "EMアルゴリズムとその周辺", *応用統計学*, **16** (1987) 1-19.
- 7) G. J. McLachlan and T. Krishnan: *The EM Algorithm and Extensions* (John Wiley, New York, 1997).

- 8) R. Sundberg: "Maximum likelihood theory for incomplete data from an exponential family," *Scand. J. Stat.*, **1** (1974) 49-58.
- 9) R. J. A. Little and D. B. Rubin: *Statistical Analysis with Missing Data* (John Wiley, New York, 1987) pp. 142-170.
- 10) M. A. Rifai: "A revision of the genus *Trichoderma*," *Mycological Papers*, **116** (1969) 1-56.
- 11) J. Bissett, "A revision of the genus *Trichoderma*. I. Section *Longibrachiatum* sect. nov.," *Can. J. Bot.*, **62** (1984) 924-931.

付録 正則な指数分布族のEMアルゴリズム

$f(\mathbf{x}|\boldsymbol{\theta})$ が以下のような形で表せる場合、 $f(\mathbf{x}|\boldsymbol{\theta})$ を正則な指数分布族であるという。

$$f(\mathbf{x}|\boldsymbol{\theta}) = b(\mathbf{x}) \exp(\boldsymbol{\theta}^T \mathbf{t}(\mathbf{x})) / a(\boldsymbol{\theta}) \quad (11)$$

ここで $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^T$ は自然パラメーター、 $\mathbf{t}(\mathbf{x}) = (t_1(\mathbf{x}), \dots, t_p(\mathbf{x}))^T$ は理想的な観測値ベクトル \mathbf{x} の空間 X での十分統計量である。十分統計量とは、標本全体の代わりに用いても、推測には十分である統計量をいい、 $\mathbf{t}(\mathbf{x})$ が十分統計量であるための必要十分条件は密度関数が、

$$f(\mathbf{x}|\boldsymbol{\theta}) = f_1(\mathbf{x}|\mathbf{t}(\mathbf{x})) f_2(\mathbf{t}(\mathbf{x})|\boldsymbol{\theta})$$

と分解されることである。

$f(\mathbf{x}|\boldsymbol{\theta})$ が正則な指数分布族である場合、EMアルゴリズムは以下のように簡略化できる。

正則な指数分布族の特性から、十分統計量 $\mathbf{t}(\mathbf{x})$ の期待値は

$$E(\mathbf{t}(\mathbf{x})|\boldsymbol{\theta}) = \partial \log a(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \quad (12)$$

となる。一方、 \mathbf{y} が与えられたときの完全データの対数尤度関数 $l_c(\boldsymbol{\theta}) = \log f(\mathbf{x}|\boldsymbol{\theta})$ の期待値は、式(11)より、

$$E(l_c(\boldsymbol{\theta})|\mathbf{y}, \boldsymbol{\theta}^{(s)}) = \log b(\mathbf{x}) + \boldsymbol{\theta}^T \mathbf{t}^{(s)} - \log a(\boldsymbol{\theta}) \quad (13)$$

となる。ただし

$$\mathbf{t}^{(s)} = E(\mathbf{t}(\mathbf{x})|\mathbf{y}, \boldsymbol{\theta}^{(s)})$$

とおいた。ここで $\boldsymbol{\theta}^{(s)}$ は s 回目の反復処理後の $\boldsymbol{\theta}$ の推定値である。Mステップでは式(13)を最大化する $\boldsymbol{\theta}$ を求めるために $\boldsymbol{\theta}$ で微分して 0 とおいて、 $\boldsymbol{\theta}$ について解くことになる。すなわち

$$\partial E(l_c(\boldsymbol{\theta})|\mathbf{y}, \boldsymbol{\theta}^{(s)}) / \partial \boldsymbol{\theta} = \mathbf{t}^{(s)} - \partial \log a(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} = 0 \quad (14)$$

ここで式(12)を代入すると、式(14)は

$$E(\mathbf{t}(\mathbf{x})|\boldsymbol{\theta}) = \mathbf{t}^{(s)} \quad (15)$$

と書き直せる。すなわち Mステップは式(15)を $\boldsymbol{\theta}$ について解くことと同意である。

以上をまとめると、正則な指数分布族に対して Eステップと Mステップは次のようになる。

Eステップ: X での十分統計量 $\mathbf{t}(\mathbf{x})$ を

$$\mathbf{t}^{(s)} = E(\mathbf{t}(\mathbf{x})|\mathbf{y}, \boldsymbol{\theta}^{(s)})$$

により推定する。

Mステップ:

$$E(\mathbf{t}(\mathbf{x})|\boldsymbol{\theta}) = \mathbf{t}^{(s)}$$

の解となる $\boldsymbol{\theta}$ を求め $\boldsymbol{\theta}^{(s+1)}$ とする。