数理・計算科学特論第二（2003年度後期）レポート問題

（講師：統計数理研究所 福水健次 fukumizu@ism.ac.jp）

Jan. 17, 2004.

以下の2問についてレポートを作成し提出せよ。**提出期限２月１３日（金）**

問題１　2次元ガウス混合分布モデル

$$\sum_{i=1}^{H} c_i \, \phi_2(x; \mu_i, V_i)$$

に対するEMアルゴリズムのプログラムを作成し、下の２つのデータセットに対してEMアルゴリズムの出力結果を求め、それらの対数尤度関数の値と得られたパラメータを記せ。

- レポートにはプログラムを添付すること。用いる言語は問わない。
- 各コンポーネントのガウス分布は、平均 μ および2次元の分散共分散行列 V をパラメータに持つとする。
- EMアルゴリズムの初期値の選び方をどのようにしたか説明せよ。もしランダム性を持つ方法を用いた場合は、それぞれのデータセットに対し5セットの異なる初期値を用いて、そのすべての結果を記すこと。
- 初期値の設定法、および分散共分散行列の特異化の排除について特に工夫した点があれば述べよ。

データセットA: http://www.ism.ac.jp/~fukumizu/class/report/data_a.txt
　コンポーネント数は３とせよ。

データセットB: http://www.ism.ac.jp/~fukumizu/class/report/data_b.txt
　コンポーネント数は４とせよ。

データフォーマットは、タブ区切りASCIIで、各行が1個の2次元データに対応する。

問題２　時系列データ $f(0), f(1), \ldots, f(N-1)$ があるとき、その離散フーリエ変換（Discrete Fourier Transform, DFT）は

$$C_f(k) = \sum_{m=0}^{N-1} f(m) e^{2\pi\sqrt{-1}\frac{mk}{N}} \qquad (0 \leq k \leq N-1)$$

により定義された。この演算を定義どおりに実行すると、$C_f(0), \ldots, C_f(N-1)$ を求めるのに N^2 の演算が必要になる。これを $N \log N$ のオーダーの演算量で実

行するアルゴリズムが高速フーリエ変換 (Fast Fourier Transtorm, FFT）である。（FFT については付録参照）

以下の各設問に答えることにより、$N = p^n$（p は素数）に対する FFT が Junction Tree アルゴリズムから導けることを示せ。

(i) FFT の計算式（(1)）を、ポテンシャル関数の積の周辺化とみなしたとき、それを分解しているグラフ（junciton tree のもとになるグラフ）はどのようなものか論ぜよ。また、$n = 3$ の場合にそのグラフの図を示せ。

(ii) 一般の n に対し、(i) のグラフは三角化可能であることを示せ。

(iii) 一般の n に対し、(i) のグラフの junction tree の図を書け。

(iv) (iii) の junction tree に対し junction tree アルゴリズムの伝搬則を適用し、それが FFT の計算と一致していることを確認せよ。

[ヒント： (i) 指数部は、計算に必要とされる項だけを考えよ。(iv) FFT ではすべてのマージナルを求める必要はない。また、ビットリバーサルなどは説明しなくてもよい。]

付録　FFT のアルゴリズム

FFT は一般の N に対して適用可能であるが、簡単のため、素数 p に対して $N = p^n$ なる場合について述べる。m および k （$0 \le m, k \le p^m - 1$）を p 進展開し、それぞれ

$$m = x_{n-1}p^{n-1} + \cdots + x_1 p + x_0, \qquad k = y_{n-1}p^{n-1} + \cdots + y_1 p + y_0$$

と書く。x_a, y_b （$0 \le a, b \le n-1$）は 0 から $p-1$ までの整数値を取る。この表現に対応して、DFT の定義式は

$$\tilde{C}_f(y_0, \ldots, y_{n-1}) = \sum_{x_0=0}^{p-1} \cdots \sum_{x_m=0}^{p-1} \tilde{f}(x_0, \ldots, x_{n-1}) \exp\left\{ 2\pi\sqrt{-1} \frac{\sum_{a=0}^{n-1} x_a p^a \sum_{b=0}^{n-1} y_b p^b}{p^n} \right\} \tag{1}$$

となる。ここで $\tilde{f}(x_0, \ldots, x_{n-1})$, $\tilde{C}_f(y_0, \ldots, y_{n-1})$ はそれぞれ対応する $C_f(k)$, $f(m)$ を意味する。

（1）式を計算する手順を $n = 3$ の場合に具体的に考えてみる。以下では 1 の p^r 乗根を A_{p^r} で表すことにする。$\tilde{C}_f$ は

$$\tilde{C}_f(y_0, y_1, y_2) = \sum_{x_0} \sum_{x_1} \sum_{x_2} \tilde{f}(x_0, x_1, x_2) \prod_{a=0}^{2} \prod_{b=0}^{2} e^{2\pi\sqrt{-1} x_a y_b p^{a+b-3}} \tag{2}$$

で与えられる。はじめに x_2 に関する和を考える。$a+b \ge 3$ のとき指数部が $2\pi\sqrt{-1}$ の整数倍になることに注意すると、$x_2 y_0$ だけを考慮すればよいことがわかる。こ

の和を $\tilde{f}_1(y_0, x_0, x_1)$ と書くと

$$\tilde{f}_1(y_0, x_0, x_1) = \sum_{x_2} \tilde{f}(x_0, x_1, x_2)(A_p)^{x_2 y_0}$$

である。これを（2）式に代入すると

$$\tilde{C}_f(y_0, y_1, y_2) = \sum_{x_0} \sum_{x_1} \tilde{f}_1(y_0, x_0, x_1) \prod_{a=0,1} \prod_{b=0}^{2} e^{2\pi\sqrt{-1} x_a y_b p^{a+b-3}}$$

が得られる。さらに x_1 に関する和を $\tilde{f}_2(y_0, y_1, x_0)$ と書くと、指数部では y_0, y_1 のみが和に関わることに注意して、

$$\tilde{f}_2(y_0, y_1, x_0) = \sum_{x_1} \tilde{f}_1(y_0, x_0, x_1)(A_{p^2})^{x_1(y_0 + p y_1)}$$

を得る。これをさらに（2）式に代入して

$$\begin{aligned} \tilde{C}_f(y_0, y_1, y_2) &= \sum_{x_0} \tilde{f}_2(y_0, y_1, x_0) \prod_{0 \le b \le 2} e^{2\pi\sqrt{-1} x_0 y_b p^{b-3}} \\ &= \sum_{x_0} \tilde{f}_2(y_0, y_1, x_0)(A_{p^3})^{x_0(y_0 + p y_1 + p^2 y_2)} \end{aligned}$$

が得られる。

$\tilde{f}_1, \tilde{f}_2, \tilde{C}_f$ を計算する各ステップでの演算量は $O(pN)$ であり、ステップの数は $n = \log N$ であることに注意すると、上のアルゴリズムは $O(pN \log_p N)$ の演算量であることがわかる。N が大きいとき、これは DFT の定義式どおりの演算量 $O(N^2)$ に比べてはるかに効率がよい。この計算手順は文献 [1]p.51 図 2・14 のように信号線図を使って表されることも多い。

参考文献

[1] 添田、中溝、大松「信号処理の基礎と応用」日新出版 (1986)