State space model approach to analyse cDNA microarray time course data of yeast

山口　類 [1)]　　樋口　知之 [2) 3)]

1) 九州大学大学院数理学研究院

2) 情報・システム研究機構　統計数理研究所

3) 科学技術振興機構 CREST

当研究では，マイクロアレイにより得られた細胞周期に伴い変動するイースト菌 (*Saccharomyces cerevisiae*) の遺伝子発現量データに状態空間モデルを適用する．EM アルゴリズムによるパラメータ推定と，情報量規準による評価により，細胞周期を表す少数の潜在変数の次元を自動同定し，他の手法による既存研究の結果との比較も行う．

State space model approach to analyse cDNA microarray time course data of yeast

Rui Yamaguchi[1)]　　Tomoyuki Higuchi[2) 3)]

1) Faculty of Mathematics, Kyushu University

2) Research Organization of Information and Systems, The Institute of Statistical Mathematics

3) Japan Science and Technology Agency, CREST

In this study, we apply a state space model to time course gene expression data of yeast (*Saccharomyces cerevisiae*). We estimate parameters characterizing the system by using EM algorithm and determine the optimal dimension of hidden state variable by an information criterion. Then we compare the result with those in other papers using the same model and the same data but with different methods to estimate the parameters and to determine the internal dimension.

1　はじめに

遺伝子マイクロアレイ実験技術の開発により多数の遺伝子の発現状態を同時に計測することができるようになった．また，そのような測定をある期間に渡って複数回繰り返すことにより，時系列データ (time course data) も得られるようになってきている．それらの時系列データの顕著な特徴の一つは，通常，遺伝子数に比して時点数が非常に少ないことである．本研究の例では，遺伝子数約 800 に対し，時点数 19 である．そのような，高次元かつ短期間のデータをモデル化し有用な情報を引き出すことは，困難かつチャレンジングな問題である．その有力なモデルの一つとして状態空間モデル（e.g. [1, 2] ）が挙げられ，遺伝子データへの適用も行われている（e.g. [3, 4] ）．後でも述べるが，状態空間モデルでは，遺伝子発現データは，動的システム内部の，直接は観測されない状態変数から逐次生成されるものとして表現される．ここで必要となることは，データを得たもとで，時間変化する内部状態を推定すると共に，システム自体を規定するパラメータを推定し，また，内部状態の最適な次元数を同定することである．本研究では，イースト菌の細胞周期時系列データを例にとり，状態空間モデルをそのデータに適用

し，EM アルゴリズム [5] によるパラメータ推定と，情報量規準による評価により，細胞周期を表す少数の潜在変数の次元の自動同定と，潜在変数間の相互作用を表す行列の推定を行う．また他の手法による既存研究の結果 [6, 7] との比較も行う．

本稿の構成は，以下のようになっている．節 2 ではモデルの説明を行う．節 3 ではパラメータ推定の方法を述べ，節 4 で実データへモデルを適用した結果を示す．節 5 で考察を述べ，最後に節 6 でまとめを述べる．

2　モデル

本研究では，マイクロアレイによって観測された時刻 n，$(n = 1, \ldots, N)$ における，l 個の遺伝子発現時系列データ $y_n = [y_{n1}, y_{n2}, \ldots, y_{nl}]'$ が，直接観測されない k 次元状態変数 $x_n = [x_{n1}, x_{n2}, \ldots, x_{nk}]'$ の動的システムから生成されるというモデルを考える．それは，以下のような，状態空間モデルとして定式化される．

$$x_n = F x_{n-1} + v_n \,, \tag{1}$$

$$y_n = H x_n + w_n \,, \tag{2}$$

ここで，F は状態遷移行列 $(k \times k)$，H は観測行列 $(l \times k)$ である．また $v_n \sim N(0, Q)$ は，平均 0，分散共分散行列 $Q, (k \times k)$ をもつガウス分布に従う k 次元システムノイズであり，$w_n \sim N(0, R)$ は，平均 0，分散共分散行列 $R, (l \times l)$ をもつガウス分布に従う l 次元観測ノイズである．式 (1)，式 (2) は，それぞれシステムモデル，観測モデルと呼ばれる．状態ベクトルの初期分布 $x_0 \sim N(\mu_0, \Sigma_0)$ およびパラメータ F, H, R, Q が所与であれば，データを与えたもとで，カルマンフィルタおよび固定区間平滑化 (e.g., [8, 9]) と呼ばれるアルゴリズムにより，状態を推定することができる．しかし，当問題においては，状態変数 x_n と共に，それらのシステムを規定するパラメータおよび状態変数の次元 k が，未知である．

そこで本研究では [10] にならい，上記パラメータを，EM アルゴリズムにより推定し，状態変数の次元 k を，BIC (Bayesian Information Criterion; [11]) にて同定する．次節において，パラメータの推定方法について，具体的に述べる．

3　パラメータ推定

パラメータの推定には EM アルゴリズムを用いる．式 (1)(2), のような状態空間モデルに対しては，[2, 12] によって，EM アルゴリズムが定式化されている．$Y_N = \{y_1, \ldots, y_N\}$，$X_N = \{x_1, \ldots, x_N\}$ とすると，完全データ $\{Y_N, X_N\}$ の尤度は，

$$P(Y_N, X_N; \theta) = P(x_0) \prod_{n=1}^{N} P(x_n | x_{n-1}) P(y_n | x_n) \,,$$

で与えられる．また $\theta = \{H, F, R, Q, \mu_0\}$ とし，Σ_0 は所与と仮定する [2]．ここで，

$$P(x_0) = \frac{\exp\left\{-\frac{1}{2}(x_0 - \mu_0)' \Sigma_0^{-1} (x_0 - \mu_0)\right\}}{(2\pi)^{k/2} |\Sigma_0|^{1/2}} \,,$$

$$P(x_n | x_{n-1}) = \frac{\exp\left\{-\frac{1}{2}(x_n - F x_{n-1})' Q^{-1} (x_n - F x_{n-1})\right\}}{(2\pi)^{k/2} |Q|^{1/2}} \,,$$

$$P(y_n | x_n) = \frac{\exp\left\{-\frac{1}{2}(y_n - H x_n)' R^{-1} (y_n - H x_n)\right\}}{(2\pi)^{l/2} |R|^{1/2}} \,,$$

である．故に，対数尤度は，

$$
\begin{aligned}
\log P(Y_N, X_N; \theta) \quad = \quad & -\frac{1}{2}\log|\Sigma_0| - \frac{1}{2}(x_0-\mu_0)'\,\Sigma_0^{-1}\,(x_0-\mu_0) \\
& -\frac{N}{2}\log|Q| - \frac{1}{2}\sum_{n=1}^{N}(x_n - Fx_{n-1})'\,Q^{-1}\,(x_n - Fx_{n-1}) \\
& -\frac{N}{2}\log|R| - \frac{1}{2}\sum_{n=1}^{N}(y_n - Hx_n)'\,R^{-1}\,(y_n - Hx_n) \\
& -\frac{k+N(k+l)}{2}\log 2\pi
\end{aligned}
$$

となる．EM アルゴリズムでは，E ステップおよび M ステップと呼ばれる操作を繰り返し

$$
q(\theta \mid \theta_0) = E\left[\log P(Y_N, X_N; \theta) \mid Y_N, \theta_0\right] \tag{3}
$$

を θ の関数として，収束するまで逐次最大化することにより，パラメータの最尤推定値 $\hat{\theta}$ を求める．ここで θ_0 は，直前の繰り返しでのパラメータベクトルである．次に，E ステップ，M ステップそれぞれの段階での操作を示す．

E ステップにおいては，対数尤度の条件付期待値を求める．i+1 回目の繰り返しにおけるそれは式 (3) より，次式で表される [2].

$$
\begin{aligned}
q(\theta \mid \theta_i) \quad = \quad & E\left[\log P(Y_N, X_N \mid \theta) \mid Y_N, \theta_i\right] \\
= \quad & -\frac{1}{2}\log|\Sigma_0| - \frac{1}{2}\mathrm{trace}\left\{\Sigma_0^{-1}\left(V_{0|N} + (x_{0|N}-\mu_0)(x_{0|N}-\mu_0)'\right)\right\} \\
& -\frac{N}{2}\log|Q| - \frac{1}{2}\mathrm{trace}\left\{Q^{-1}(C - BF' - FB' + FAF')\right\} \\
& -\frac{N}{2}\log|R| - \frac{1}{2}\mathrm{trace}\left\{R^{-1}\sum_{n=1}^{N}\left[(y_n - Hx_{n|N})(y_n - Hx_{n|N})' + HV_{n|N}H'\right]\right\} \\
& -\frac{k+N(k+l)}{2}\log 2\pi\,.
\end{aligned}
$$

ここで，$\theta_i = \{H(i), F(i), R(i), Q(i), \mu_0(i)\}$ は，i 回目の繰り返しで得られた，パラメータベクトルであり．$A = \sum_{n=1}^{N}\left(V_{n-1|N} + x_{n-1|N}x'_{n-1|N}\right)$，$B = \sum_{n=1}^{N}\left(V_{n,n-1|N} + x_{n|N}x'_{n-1|N}\right)$，および $C = \sum_{n=1}^{N}\left(V_{n|N} + x_{n|N}x'_{n|N}\right)$ である．また，上式中，$x_{n|N} = E\{x_n|Y_N\}$，$V_{n|N} = E\{(x_n - x_{n|N})(x_n - x_{n|N})'|Y_N\}$，および $V_{n,n-1|N} = E\{(x_n - x_{n|N})(x_{n-1} - x_{n-1|N})'|Y_N\}$ は，カルマンフィルタ と固定区間平滑化を用いることによって得ることができる．更に，カルマンフィルタの副産物として周辺対数尤度関数 $\log L(Y_N|\theta_i)$ も得られる．

一方，M ステップにおいては，次式，$\partial_H q(\theta \mid \theta_i) = 0$，$\partial_F q(\theta \mid \theta_i) = 0$，$\partial_R q(\theta \mid \theta_i) = 0$，$\partial_Q q(\theta \mid \theta_i) = 0$，および $\partial_{\mu_0} q(\theta \mid \theta_i) = 0$ を満たすように，パラメータが更新される．結果，i+1 回目の繰り返しでは以下のようにパラメータが更新される．

$$
\begin{aligned}
H(i+1) &= \left(\sum_{n=1}^{N} E\{y_n x'_n \mid Y_N\}\right)C^{-1}\,, \\
F(i+1) &= BA^{-1}\,, \\
R(i+1) &= N^{-1}\sum_{n=1}^{N}\left[(y_n - Hx_{n|N})(y_n - Hx_{n|N})' + HV_{n|N}H'\right]\,, \\
Q(i+1) &= N^{-1}\left(C - BA^{-1}B'\right)\,, \\
\mu_0(i+1) &= x_{0|N}\,.
\end{aligned}
$$

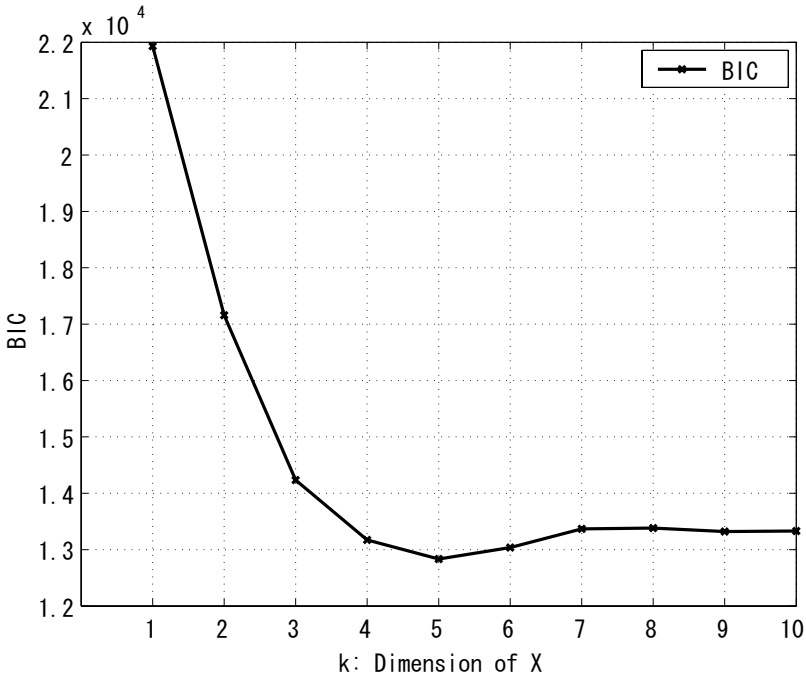

図 1: 状態変数 x_n の次元 k の変化に対する BIC の変化

また，ここでは詳述しないが，欠測値を含む場合も自然な方法で，状態推定，およびパラメータ推定を行うことが可能である [2, 12].

4　解析

上記の手法を，[13] により公開されているイースト菌細胞周期遺伝子発現データ ($cdc15$, $N = 19$ 時点, 遺伝子数 $l = 763$ 個) に適用した [10]．元データには，800 個の遺伝子が含まれているが，19 時点のうち欠測値が，10 時点以下のものを採用した．ただし適用したモデルでは，Q，R に対角行列を仮定した，更に R の各対角成分は単一と仮定した．解析では，状態変数の次元 k を $\{1, ..., 10\}$ の範囲で変え，各 k において，EM アルゴリズムに与えるパラメータ $\theta^{(k)}$ の初期値を 5 組，乱数により用意し，最も大きな対数尤度を生じた場合のパラメータを，各次元の最尤パラメータ $\hat{\theta}^{(k)}, k = 1, \ldots, 10$ とした．

次に，最適な状態変数の次元を選ぶために，BIC を以下のように計算した．

$$BIC(k) = -2 \log L(Y_N|\hat{\theta}^{(k)}) + (\#\ of\ \ parameters) \cdot \log N\,, \tag{4}$$

ここで，N は時点数，$(\#of\,parameters) = k(k + l + 2) + 1$ は自由パラメータの数である．

結果，図 1 に示されているように，BIC が最小となる状態変数の次元は $\hat{k} = 5$ であった．

5　考察

関連する先行研究として，ほぼ同じモデル，データを用いて，内部次元の推定を行っている [6, 7] が挙げられる．これらの研究間には，欠測値の処理等にも細かな違いもあるが，より大きな違いは，パラ

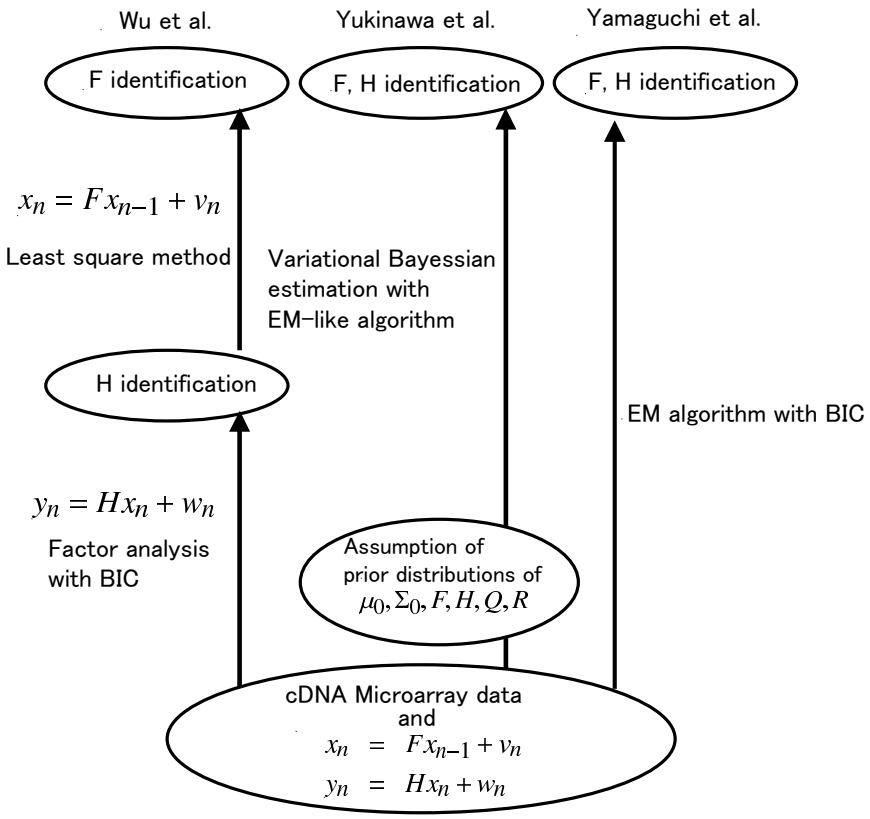

図 2: 関連研究の推定方法および次元選択法の概略

メータの推定方法と，モデル選択の基準である．前者は，まず H 行列を，あらかじめ因子分析モデルに EM アルゴリズムを用いて推定し，次元数の推定を BIC によって行っている．そして，その後 F 行列を，最小二乗法で求めている．また，後者は，パラメータに事前分布を仮定し，パラメータを変分ベイズ法（変分法的 EM）により推定している．次元数の推定は自由エネルギーの比較によって行っている．一方，本研究は，パラメータの事前分布を導入することなく，全てのパラメータの推定を上述の EM アルゴリズムで同時に行い，次元推定を BIC によって行っている．図 2 に，推定方法の概略を示す．

今回，得られた結果は，[6] の結果と次元数では一致しており，[7] の結果 $k = 2$ とは異なっている．しかしながら [6] に記載されている BIC の計算では，BIC の罰則項中の，対数変換を受ける値，すなわち “sample size” として，遺伝子数が使われている．時系列データの場合，“sample size” としては，通常，式 (4) のように時点数を採用するので，それに比べて過剰に罰則がかかっている可能性がある．これにより，結果が変わることもありえるので，同様の方法を用いて確認する必要がある．また，このように，推定方法により，最適な内部次元の値が異なる結果が得られたが，どの方法が良いかは，次元選択に使用する基準が異なることもあり一概に比べることはできない．今後，遺伝子数や時点数の数の異なる，いろいろなタイプの実データ，人工データとの比較により，どの場合にどの方法が良い結果を出す傾向があるか調べる必要がある．しかしながら，同様のモデルを用いながら，一旦，静的モデルとみなし，かつ段階的にパラメータを推定する [6] の方法よりも，一貫して，動的モデルの枠組みで解析を行っている [7] や，本稿で用いた手法の方が，よりデータの構造に対して素直に従っているように思われる．

6 まとめ

本研究では，遺伝子マイクロアレイにより取得された，細胞周期中のイースト菌遺伝子発現の時系列データ（高次元，短期間）をモデル化するために，状態空間モデルを適用した．モデルのパラメータ推定には，EM アルゴリズムを用いた．また，最適な状態変数の次元 $\hat{k}$ を BIC を用いて選択した．その結果，$\hat{k} = 5$ という結果が得られた．次に，ほぼ同様のモデル，データに対して，異なる推定方法および異なるモデル選択基準を用いた，二つの研究 [6], [7] の結果と比較した．本研究の結果は [6] と同じ次元を選択し，[7] とは，異なる，次元を選択した．しかしながら，BIC の計算仕方などに異論があり，今後，確認のための解析を進める必要がある．

また，更に重要なのは，推定されたパラメータ $\hat{\theta}$ や状態変数の，生物学的知見との比較である．特に，内部変数の個々の要素が，生物システム内の要素へ対応付けられれば有用であろう．また，パラメータ F や H は，遺伝子間の繋がりや，内部要素間のつながりに関する，重要な情報を持っていることがモデルの構造から期待される．今後，推定されたパラメータを用いて，ネットワーク構造の推定および，他の有用な情報の抽出を目指して解析を進める予定である．

謝辞

本研究の一部は，文部科学省 21 世紀 COE プログラム「機能数理学の構築と展開」（九州大学大学院数理学府）からの援助による．

参考文献

[1] A. C. Harvey. *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press, New York, 1989.

[2] R. H. Shumway and D. S. Stoffer. An approach to time series smoothing and forecasting using the EM algorithm. *J. Time Series Analysis*, (3):253–264, 1982.

[3] K. Murphy and S. Mian. Modelling gene expression data using Dynamic Bayesian Networks. Technical report, University of Calfornia, Berkeley, CA, 1999.

[4] C. Rangel, J. Angus, Z. Ghahramani, M. Lioumi, E. Sotheran, A. Gaiba, D. L. Wild, and F. Falciani. Modelling T-cell activation using gene expression profiling and state space models. *Bioinformatics*, 20:1361–1372, 2004.

[5] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B,*, 39:1–38, 1977.

[6] F. X. Wu, W. J. Zhang, and A. J. Kusalic. Modeling gene expression from microarray expression data with state-space equations. *Pacific Symposium on Biocomputing*, 9:581–592, 2004.

[7] 行縄 直人, 吉本 潤一郎, 大羽 成征, and 石井 信. 線形ダイナミカルシステムモデルの変分ベイズ推定による遺伝子発現時系列のシステム同定. In **情報処理学会研究報告、数理モデル化と問題解決研究会**, number 51, pages 13–16, 2004.

[8] R. E. Kalman. A new approach to linear filtering and prediction problems. *Trans. Amer. Soc. Mech. Eng., J. Basic Engineering*, 82:35–45, 1960.

[9] 北川 源四郎. **時系列解析入門**. 岩波書店, 2005.

[10] R. Yamaguchi, S. Yamashita, and T. Higuchi. Estimating gene networks with cDNA microarray data using state-space models. In *Computational Science and Its Applications – ICCSA 2005*, volume 3482 of *Lecture Notes in Computer Science*, pages 381–388. Springer-Verlag GmbH, 2005.

[11] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6:461–464, 1978.

[12] R. H. Shumway. Dynamic mixed models for irregularly observed time series. *Resenhas- Reviews of the Institute of Mathematics and Statistics*, 4(4):433–456, 2000. University of Sao Paulo, Brazil: USP Press.

[13] P. T. Spellman and G. Sherlock et al. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. of the Cell*, 9:3273– 3297, 1998.