2004－MPS－51　(4)
2004／9／13

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

行縄直人[†]　吉本潤一郎[††,†]
大羽成征[†]　石井信[†]

遺伝子発現ダイナミクスの解析のために，状態空間モデルに基く解析法が提案されている．従来の解析法では，状態変数のダイナミクスを仮定せず，また，システムノイズと観測ノイズを無視したモデルを仮定していたため，状態空間に含まれるダイナミクスを持たないノイズ成分を，状態変数として誤検出する可能性がある．本研究では，ノイズプロセスに白色ガウシアンを仮定した線形ダイナミカルシステムモデルを考え，変分ベイズ法による推定とモデル選択を行う．本手法を出芽酵母細胞周期に関する公開データセットに適用したところ，従来手法で選択されたモデルと比較し，よりシンプルで尤もらしいモデルが選択された．また，この結果得られたモデルパラメータは，生物学的な考察と良く一致した．人工データへの適用も行い，ノイズを含む時系列データに対する有効性が示された．

System Identification of Gene Expression Time-series based on a Linear Dynamical System Model with Variational Bayesian Estimation

Naoto Yukinawa[†] ,Jun-ichiro Yoshimoto[††,†] ,Shigeyuki Oba[†]
and Shin Ishii[†]

Several methods based on state space models have been proposed for analyzing dynamics of gene expression. Existing analysis methods can detect false noisy internal variables which seem to have no dynamics in state space because the methods don't assume any dynamics with system noise and observation noise. In this study, we propose a linear dynamical system model in which state variables and observation variables are generated by Gaussian white noise process and we provide a variational Bayes inference for the model. We first show effectiveness of our method for synthesized noisy time-series data set. We also apply our method to a published yeast cell-cycle gene expression data set, then show that our method could select simpler and more plausible model than existing method does. In addition, the resultant model parameters well match the biological considerations.

1. はじめに

本研究では，遺伝子発現プロファイルをもとに，多数の遺伝子の発現をコントロールする遺伝子発現制御因子の数を推定する問題を扱う．遺伝子発現の大域的な挙動については，転写制御因子や外的環境といったわずかな数の要因に支配されているという仮説に基づき，細胞状態の時間変化を観測した遺伝子発現プロファイルからのダイナミクスの解析のために，線形状態空間モデルをベースにした解析法が提案されている [1)~3)]．

† 奈良先端科学技術大学院大学情報科学研究科
Graduate School of Information Science, Nara Institute of Science and Technology
†† 独立行政法人科学技術振興機構
沖縄新大学院大学先行的研究事業
Initial Research Project, Okinawa Institute of Science and Technology, JST

状態空間モデルでは，観測系列は遺伝子発現量に対応し，非観測内部変数および遷移行列は，遺伝子発現を制御する因子を仮想している．このため，モデルの複雑さを規定する状態空間の次元数の最適決定が問題となる．従来手法では，しばしば状態空間でのダイナミクスと，システムノイズと観測ノイズを無視したモデルを仮定している．しかし，遺伝子発現プロファイルは高次元でありノイズが多く含まれるため，こうした簡略なモデルでは，ノイズを含みデータの生成過程にダイナミクスが想定されるデータに対して適切なモデルが選択ができない可能性がある．

本研究では，遺伝子制御系のモデルとして線形ダイナミカルシステム (Linear Dynamical System; LDS) モデルを仮定したシステム同定法を提案し，遺伝子発現レベルの時系列データから，動的に変化する発現制御因子の挙動と，遺伝子の特徴を同時に解析する手法

を提案する．

適用実験では，まず人工データからのモデルパラメータの推定を行い，ダイナミクスを持つ状態変数系列の次元を正しく捉えられることを示した．次に，出芽酵母の細胞周期に関する公開データ [9] を用いて，モデルとデータとの適合性，データ生成の内部状態に関して検討した．また，得られた観測行列と公開データに対する生物学的知識との関連づけを行い，遺伝子を特徴付ける情報が観測行列に抽出されている可能性があることを示した．

2. 線形ダイナミカルシステムモデル

2.1 遺伝子発現プロファイル

遺伝子発現プロファイルとは，さまざまな実験条件下での細胞サンプルにおける遺伝子の発現レベルを網羅的に測定したデータである．測定時点 t における発現プロファイルベクトル $\boldsymbol{y}_t$ を，

$$\boldsymbol{y}_t = (y_{t1}, \ldots, y_{tD})'; \quad t = 1, \ldots, T \tag{1}$$

で表す．ここで y_{tj} は測定時点 t, 遺伝子 j の発現レベル，D は測定対象となる遺伝子数，また T は測定時点数である．

2.2 線形ダイナミカルシステムの確率モデル

本研究で用いる LDS モデルは，離散時間で遷移する N 次元の非観測な内部状態変数の系列 $\boldsymbol{x}$ と，その線形変換により生成される D 次元の可観測な観測状態変数の系列 $\boldsymbol{y}$ の二つの状態系列について，以下のシステム方程式として定式化される．

$$\boldsymbol{x}_t = \boldsymbol{W}\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t; \quad t = 2, \ldots, T, \tag{2}$$

$$\boldsymbol{y}_t = \boldsymbol{V}\boldsymbol{x}_t + \boldsymbol{\eta}_t; \quad t = 1, \ldots, T, \tag{3}$$

$$\boldsymbol{x}_1 \sim \mathcal{N}_N(\boldsymbol{x}_1|\boldsymbol{\mu}_1, \sigma_1^2\boldsymbol{I}_N), \tag{4}$$

$$\boldsymbol{\epsilon}_t \sim \mathcal{N}_N(\boldsymbol{\epsilon}_t|\boldsymbol{0}_N, \sigma_\epsilon^2\boldsymbol{I}_N), \tag{5}$$

$$\boldsymbol{\eta}_t \sim \mathcal{N}_D(\boldsymbol{\eta}_t|\boldsymbol{0}_D, \sigma_\eta^2\boldsymbol{I}_D) \tag{6}$$

$\boldsymbol{\epsilon}_t \in \mathcal{R}^N$ および $\boldsymbol{\eta}_t \in \mathcal{R}^D$ はそれぞれ観測ノイズとシステムノイズである．これらのノイズは正規分布に従う．

$\boldsymbol{\mu}_1 \in \mathcal{R}^N$ は状態変数の初期値の平均値，$\boldsymbol{W} \in \mathcal{R}^{N\times N}$ は内部状態遷移行列（遷移行列）, $\boldsymbol{V} \in \mathcal{R}^{D\times N}$ は観測状態生成行列（観測行列）である．σ_1^2, σ_ϵ^2, σ_η^2 はそれぞれ $\boldsymbol{x}_1$, $\boldsymbol{\epsilon}_t$, $\boldsymbol{\eta}_t$ の分散である．$\theta \equiv \{\boldsymbol{\mu}_1, \sigma_1^2, \boldsymbol{W}, \sigma_\epsilon^2, \boldsymbol{V}, \sigma_\eta^2\}$, が，モデルパラメータのセットとなる．

以上をまとめると，システム方程式に対応する確率モデル

$$p(\boldsymbol{Y}_{1:T}, \boldsymbol{X}_{1:T}|\boldsymbol{\theta}) = \prod_{t=1}^{T} p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{\theta}) p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\theta}), \tag{7}$$

が得られる．ここで，$\boldsymbol{X}_{1:T} \equiv \{\boldsymbol{x}_t\}, \boldsymbol{Y}_{1:T} \equiv \{\boldsymbol{y}_t\}$ である．式 (7) をモデルパラメータ $\boldsymbol{\theta}$ の尤度関数と呼ぶ．

本研究では，モデルパラメータ $\boldsymbol{\theta}$ の事前分布として以下で与えられる共役事前分布を用いる．

$$p(\boldsymbol{\mu}) = \mathcal{N}_N\left(\boldsymbol{\mu}|\boldsymbol{0}, \gamma_0^{-1}\boldsymbol{I}_N\right), \tag{8}$$

$$p(\sigma_1^2) = \mathcal{G}\left(\sigma_1^{-2}|\gamma_0, \gamma_0\tau_{\mu 0}\right), \tag{9}$$

$$p(\boldsymbol{W}) = \prod_{i=1}^{N} \mathcal{N}_N\left(\boldsymbol{w}_i|\boldsymbol{0}_N, \gamma_0^{-1}\boldsymbol{I}_N\right), \tag{10}$$

$$p(\sigma_\epsilon^2) = \mathcal{G}\left(\sigma_\epsilon^{-2}|\gamma_\epsilon, \gamma_\epsilon\tau_{\mu_\epsilon}\right), \tag{11}$$

$$p(\boldsymbol{V}) = \prod_{j=1}^{D} \mathcal{N}_D\left(\boldsymbol{v}_j|\boldsymbol{0}_D, \gamma_0^{-1}\boldsymbol{I}_D\right), \tag{12}$$

$$p(\sigma_\tau^2) = \mathcal{G}\left(\sigma_\tau^{-2}|\gamma_\tau, \gamma_\tau\tau_{\mu_\tau}\right) \tag{13}$$

$\mathcal{G}(\sigma^{-2}|\gamma, \gamma\tau)$ は，ガンマ分布である．$\boldsymbol{w}_i$ と $\boldsymbol{v}_i$ はそれぞれ，$\boldsymbol{W}$ の第 i 行ベクトルと $\boldsymbol{V}$ の第 j 行ベクトルを示す．

2.3 観測行列の性質

観測行列 $\boldsymbol{V}$ は $D \times N$ 行列である．各行ベクトル $\boldsymbol{v}_i \in \mathcal{R}^{1\times N}, i = 1, \ldots, D$ は内部状態 $\boldsymbol{x}_t$ から観測変数 y_{ti} への写像を規定し，大域的因子に対する遺伝子 i の応答特性を示すものである．この性質から，$\boldsymbol{v}_i$ を遺伝子 i に対する特徴量と見なすことができ，観測ベクトルと呼ぶ．

2.4 変分ベイズ法

変分ベイズ推定では，内部状態変数 $\boldsymbol{X}$ およびパラメータ $\boldsymbol{\theta}$ の事後分布 $p(\boldsymbol{X}, \boldsymbol{\theta}|\boldsymbol{Y})$ を近似するための試験事後分布 $q(\boldsymbol{X}, \boldsymbol{\theta}) \approx p(\boldsymbol{\theta}, \boldsymbol{X}|\boldsymbol{Y})$ を用意し，以下で定義される対数周辺化尤度 $\ln p(\boldsymbol{Y})$ の下界である自由エネルギー (variational free energy) $\mathcal{F}[q(\boldsymbol{\theta}, \boldsymbol{X})]$ を，試験事後分布に関して変分法的に最大化することでベイズ推定を実現する．

$$\begin{aligned} \ln p(\boldsymbol{Y}) &\equiv \ln \int p(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{\theta}) p(\theta) d\boldsymbol{\theta} d\boldsymbol{X} \\ &\geq \int q(\boldsymbol{\theta}, \boldsymbol{X}) \ln \frac{p(\boldsymbol{Y}, \boldsymbol{X}|\theta) p(\boldsymbol{\theta})}{q(\boldsymbol{\theta}, \boldsymbol{X})} d\boldsymbol{\theta} d\boldsymbol{X} \\ &\equiv \mathcal{F}[q(\boldsymbol{\theta}, \boldsymbol{X})] \end{aligned} \tag{14}$$

$\mathcal{F}[q(\boldsymbol{\theta}, \boldsymbol{X})]$ の最大化は，独立分解近似 $q(\boldsymbol{\theta}, \boldsymbol{X}) = q(\boldsymbol{\theta})q(\boldsymbol{X})$ のもとで，$q(\boldsymbol{X})$ に関する最大化，$q(\boldsymbol{\theta})$ に関する最大化を交互に繰り返す変分法的 EM (VB-EM) アルゴリズムによって行うことができ，収束性が保証されている．また，自由エネルギーの最大値は対数周辺化尤度の近似値となっているため，パラメータ数の異なるモデル間での，モデル選択基準となり得る [10]．

3. 適用実験

3.1 人工データによる評価

まず，$N = 3$ の LDS モデルを用いて，15 時点の内部状態変数系列を生成した．これらに加え，ダイナミクスを仮定しない無情報な 2 つの内部状態変数系列を，区間 $[-0.053, 0.053]$ の一様乱数より生成することで，5 つの内部状態変数系列を得た．システムノイズの標準偏差 σ_ϵ は 0.02 とした．次に，得られた内部状態変数系列に対し，$\mathcal{N}_{100}(0, \boldsymbol{I}_{100})$ に従って生成した観測行列 $\boldsymbol{V}$ を用いて，100 サンプルの観測状態系列 $\boldsymbol{Y}_{1:15}$ を生成した．観測ノイズの標準偏差 σ_η は 0.05 とした．

データに対し，$N = 1$ から $N = 10$ までの 10 個の LDS モデルを用意し，推定の結果最大の自由エネルギーが得られたモデルを最適なモデルと決定した．比較のため，従来手法である，因子分析と BIC によるモデル選択も同様に行った．

この結果，LDS モデルでは状態空間の次元 $N = 3$ が選択された．また，因子分析モデルでは，BIC より $N = 5$ のモデルが選択された．次に，LDS モデルに関するシステムノイズ分散と観測ノイズ分散の推定値に関して評価を行なった．モデルの複雑さが増加するほど，データに適合しやすくなるため，観測ノイズ分散は内部状態変数の次元に対して単調減少を示す．一方，システムノイズ分散は，$N = 3$ のモデルで最小値をとる形となった．

3.2 酵母遺伝子発現プロファイルに対する適用

公開遺伝子発現プロファイルデータに対する適用実験を行った．用いるデータは，Spellman らが文献 9) の実験において，出芽酵母 *cdc15-2* の変異株の細胞周期における 6177 遺伝子の発現量の 24 時点にわたる時間変化を cDNA マイクロアレイを用いて観測した対数発現比である．本データは，http://cellcycle-www.stanford.edu/ から入手可能である．このデータセットより，ランダムに 200 遺伝子を抽出し，200 遺伝子 ×19 時点の学習データを構成した．

内部状態変数の次元を $N = 1, \ldots, 10$ に設定した 10 個の LDS モデルを用意し，VB-EM アルゴリズムによるパラメータ推定を行った．また，因子分析モデルに対する EM アルゴリズムによるシステム同定を行い比較した．

図 1 は，内部状態変数の次元 $N = 1, \ldots, 7$ に対する，LDS モデルにおける自由エネルギーと，因子分析モデルにおける BIC を示すプロットである．これより，LDS モデルでは最適な状態空間の次元数は $N = 2$ であるといえる．一方，因子分析モデルでは $N = 5$ のモデルが選択されている．

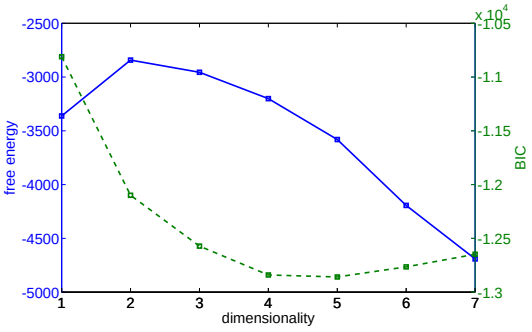

図 1 モデルの内部状態空間の次元に対する提案モデルでの自由エネルギーと 因子分析モデルでの BIC．実線が自由エネルギー，破線が BIC を示す．

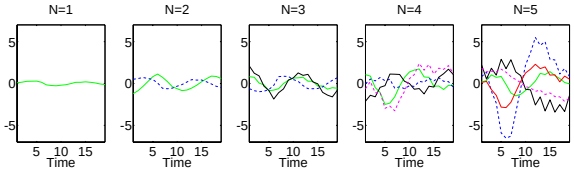

図 2 自由エネルギーが最大となったモデルでの内部状態変数 $\boldsymbol{x}$ の時系列．各列はモデルの内部状態変数の次元に対応する．

図 2 は，自由エネルギーが最大となった $N = 1, \ldots, 5$ のモデルの内部状態変数の変動を，推定したパラメータから再現したものである． $N = 1$ のモデルでは発現プロファイルの変動を表すには十分ではないと考えられる．また，$N = 4$ や $N = 5$ のモデルでは，ある状態変数の変動が，他の状態変数のものの定数倍，もしくは，状態変数の変動同士の重ね合わせ表現されるような，冗長性が観察される．全モデル中で 自由エネルギーが最大の $N = 2$ では，位相が異なりながら周期的挙動を示す二種類の変動が抽出されている．

図 3 は，自由エネルギーが最大となった $N = 2$ のモデルにおける $\boldsymbol{V}$ の推定値における観測ベクトル $\boldsymbol{v}_i, i = 1, \ldots, D$ を，二次元の要素空間にプロットしたものである．図中の各点が 1 遺伝子に対応する．図中のシンボルは，Spellman らのフェーズの分類結果に対応している． 観測ベクトルの v_1–v_2 空間において，時計回りの回転方向に 5 つの細胞周期フェーズに分類された遺伝子が並んでいることから，LDS モデルが Spellman らが遺伝子を分類した際の特徴空間を自動的に構成していることが分かる．

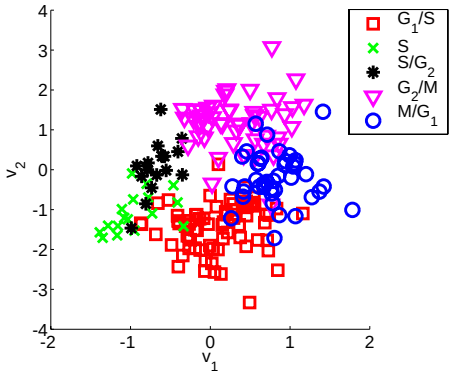

図 3 $N = 2$ の LDS モデルの推定結果から得られた $\boldsymbol{V}$ の横ベクトルの散布図．各シンボルは，Spellman らによって同定された，体細胞分裂の過程において遺伝子が高いレベルで発現するフェーズを示す．

4. 議　　論

Spellman ら [9] は，細胞周期における遺伝子発現変化の周期性を仮定し，解析のためにフーリエ基底を採用している．このモデルでは，位相と振動数がシステムを規定するパラメータであり，それらは LDS モデルにおける状態遷移行列と状態変数の初期値に対応する．また，2 つの線形和の重みパラメータは，LDS モデルの観測ベクトルに対応する．今回，我々の手法により，Spellman らの解析モデルにおいて仮定されているものと等価な $N = 2$ の基底（内部状態時系列）を自動的に構成することができた．

一方，ノイズと状態変数のダイナミクスを陽に仮定しない因子分析モデルでは，状態空間次元 $N = 5$ のモデルが選ばれた．人工データの解析結果を踏まえると，これは，因子分析モデルが，状態空間に含まれる意味のないノイズ成分を別の因子として捉えてしまったことが一因であると考えられる．

5. おわりに

我々の手法の一番の強みは，定常的な過程にあると考えられる現象から観測されたダイナミクスを持つ時系列データに対して，最適な基底を自動的に求めることができることにある．これは，生物のような自律的に恒常的活動を刻むシステムの背後にある要因を探ることを可能にする道具となりうる．

一方で，本手法の欠点としては，一般的に構成要素の因果関係が非線形であると考えられている生物のシステムに線形性の仮定を行っていること，システムに定常性を要求すること，また，内部状態変数に対する外部因子の入力を省いた状態空間モデルとなっていることの 3 つが主に考えられる．将来的には，これらの簡略化を除去した手法を提案したい．また，より広範なデータに対して本手法を適用し，その有効性を検討する予定である．

参　考　文　献

1) Wu, F. X., Zhang, W. J. and Kusalik, A. J.: Modeling gene expression from microarray expression data with state-space equations, *Pacific Symposium on Biocomputing*, Vol. 9, pp. 581–592 (2004).

2) Dewey, T. G. and Galas, D. J.: Dynamic models of gene expression and classification, *Functional & Integrative Genomics*, Vol. 1, pp. 269–278 (2001).

3) Holter, N. S., Maritan, A., Cieplak, M., Fedoroff, N. and Banavar, J. R.: Dynamic modeling of gene expression data, *PNAS*, Vol. 98, pp. 1693–1698 (2001).

4) Dempster, A.P., Laird, N.M. and Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm, *Journal of Royal Statistical Society B*, Vol. 39, pp. 1–38 (1977).

5) Roweis, S. and Gharahmani, Z.: A unifying review of Linear Gaussian models, *Neural Computation*, Vol. 11, pp. 305–345 (1999).

6) Attias, H.: Inferring parameters and structure of latent variable models by variational Bayes, *Proceedings of 15th Conference on Uncertainty in Artificial Inteligence*, pp. 21–30 (1999).

7) Gharahmani, Z. and Beal, M. J.: Propagation algorithms for variational Bayesian learning, *Advances in Neural Information Processing Systems 13*, pp. 507–513 (2001).

8) Yoshimoto, J., Ishii, S. and Sato, M.: System identification based on on-line variational Bayes method and its application to reinforcement learning, *Artificial Neural Networks and Neural Information Processing - ICANN/ICONIP 2003*, pp. 123–131 (2003). Lecture Notes in Computer Science 2714.

9) Spellman, P. T., Sherlock, G., Zang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B.: Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Molecular Biology of the Cell*, Vol. 9, pp. 3273–3297 (1998).

10) J. Yoshimoto, S. I. and Sato, M.: Hierarchical model selection for NGnet based on variational Bayes inference, *Artificial Neural Networks - ICANN 2002*, pp. 661–666 (2002). Lecture Notes in Computer Science 2415.