NAIST-IS-DD0361037

博士論文

統計的手法による遺伝子発現情報からの細胞状態の同定に関する研究

行縄 直人

2006年 2月 2日

奈良先端科学技術大学院大学
情報科学研究科 情報生命科学専攻

本論文は奈良先端科学技術大学院大学情報科学研究科に
博士 (工学) 授与の要件として提出した博士論文である。

行縄 直人

審査委員：

石井 信 教授	（主指導教員）
小笠原 直毅 教授	（委員）
加藤 菊也 教授	（委員，大阪府立成人病センター研究所）
川端 猛 助教授	（委員）

統計的手法による遺伝子発現情報からの細胞状態の同定に関する研究*

行縄 直人

内容梗概

近年，マイクロアレイや定量的 PCR 法などの mRNA 定量化技術により，細胞サンプルにおける包括的な遺伝子発現情報を得ることが可能となり，細胞の状態と遺伝子発現を直接結び付けて解析する，いわゆるトランスクリプトーム解析が行われるようになった．本論文では，トランスクリプトーム解析における諸問題に対して，頑健な解析を行うための統計的手法に関して議論する．

まず，包括的 mRNA 測定技術の一つである，アダプタ付加競合 PCR (ATAC-PCR) 法により得られる蛍光量データの特徴とその補正法について報告する．これまで問題となっていたアダプタ長依存の測定バイアスの解明を主眼とし，ATAC-PCR 法で得られたデータの詳細な解析を行った．解析結果に基づき蛍光ピーク値に関する観測モデルの定式化を行ない，ノイズ項のパラメータの推定量の導出と，それらを用いたピーク値補正法を提案した．この手法を，アダプタノイズ解析のために特化した採取されたピークデータに適用し，アダプタ依存ノイズのパラメータを求め，次いで，実データに対しバイアス補正の適用を試み，その有効性を確認した．

次に，生きた細胞における遺伝子発現ダイナミクスの解析を目指し，遺伝子発現プロファイルの時系列に対する解析法について述べる．ここでは，状態空間モデルに基き，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ための新たな手

*奈良先端科学技術大学院大学 情報科学研究科 情報生命科学専攻 博士論文, NAIST-IS-DD0361037, 2006 年 2 月 2 日.

法を提案した．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，より単純かつ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．

最後に，遺伝子発現からの癌の病理診断を想定した，新たな多クラス識別法について述べる．本手法では，多クラス識別問題を一対一ペアや一対残りペアなどのラベルの任意の組み合わせから成る 2 値分類問題群に分解し，各問題での判別結果を統合することによって最適な識別結果を得る．各 2 値分類問題における真の分類確率がクラス所属確率をパラメータとした確率モデルによって生成されると考え，これを 2 値分類器によって得られた分類確率の推定値から推定する方法，さらに 2 値分類器の重みを推定する方法を導いた．本手法を人工データおよび甲状腺がん分類問題をはじめとした実データに適用し，従来のヒューリスティクスによる投票法と同等以上の性能を達成することを示した．さらに，この分野で提案されてきたいくつかの多クラス識別法との比較を行い，本手法の優位性および性質を明らかにした．

キーワード

遺伝子発現, アダプタ付加競合 PCR 法, 状態空間モデル, システム同定, 病理診断, 多クラスパターン識別

Studies on statistical methods for identification of cell states using gene expression information*

Naoto Yukinawa

Abstract

In order to understand biological activities of cells at molecular level, it is required to know various aspects of gene expression; the concentration, timing, conditions, and localization in subcellular organelle, in detail. Gene expression in living organisms is controlled within a complicated gene regulatory system consisting of many intermolecular interactions of various components such as nucleic acids, proteins or other small molecules. Even if we have individual concentration information of each biological component in cells, it is still difficult to understand its dynamics directly. At the present day, some analysis methods have been developed for this problem, which are based on statistics and/or information technology for gene expression profiles: transcriptome analysis – comprehensive measurement of expression levels of thousands of genes in cells using mRNA quantification technology. In this thesis, I validate robustness and effectiveness of statistical method dealing with three types of problems in transcriptome analysis.

First, I propose a bias-correcting method for adaptor-tagged competitive (ATAC)-PCR which is a PCR-based quantification technology for large-scale analysis. In this study, I evaluate adaptor-dependent bias under a setting of large-scale data production.

Next, I propose a linear dynamical system model in which state variables and observation variables are generated by Gaussian white noise process and provide a variational Bayes inference for the model for analyzing dynamics of gene expression. I

*Doctoral Dissertation, Department of Bioinformatics and Genomics, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD0361037, February 2, 2006.

first show effectiveness of our method when applied to a synthesized noisy time-series data set. I also applied our method to a published yeast cell-cycle gene expression data set, then our method could select a simpler and more plausible model than existing method did. In addition, the resultant model parameters well matched the biological considerations.

Third, for Multi-class cancer diagnosis based on gene expression profiling, I propose a novel framework for constructing a multi-class pattern classifier by optimally weighted aggregation of general binary classifiers including one-versus-the-rest, one-versus-one, and others. In principal our framework is independent of binary classification algorithms to use and gene selection methods. I apply the proposed method to various classification problems including a synthesized dataset, three gene expression datasets, and a modified gene expression dataset. The results demonstrate that our method can improve classification accuracy over the simple combinations of binary classifiers in most situations. Furthermore, I show that the performance of our method is better than or comparable to state-of-the-art multi-class predictors.

Keywords:

Gene expression, Adaptor-tagged competitive PCR, State space models, System identification, Pathological dignosis, Multi-class pattern classification

目 次

図目次

表 目 次

第1章　序論

1. 研究の背景

1.1 遺伝子発現解析

分化，増殖，生存および病理学的な状態など，細胞の多くのふるまいや状態は，それぞれに固有の遺伝子発現パターンを反映したものとなっている．このため，興味ある細胞の特定の状態における，特定の遺伝子の転写レベルの計測は，遺伝子機能解析に関する研究において中核をなしてきた [1]．遺伝子転写レベルの測定のために使われてきた代表的な手法としては，ノーザンブロッティング，*in situ* ハイブリダイゼーション法 [2]，RNAse プロテクション法 [3][4] および RT-PCR 法 [5] など (m)RNA 量測定法が挙げられる．基本的に，これらの測定法は調査対象となる遺伝子を絞り込んだ後で，それらの遺伝子発現を詳細に調べることを目的としている．

一方，1990 年代の中盤から後半にかけて，特定の条件における包括的な遺伝子発現を同時に測定することで，遺伝子あるいは遺伝子制御の時空間的な性質を調べようとする，遺伝子発現プロファイリング，トランスクリプトーム解析，あるいは単に遺伝子発現解析と呼ばれる新たな試みが登場した．その有効性は完全には検証されていないものの，現在は広く普及した方法論となっている．遺伝子発現解析において用いられている網羅的 RNA 測定法は，cDNA アレイ [6][7] やオリゴヌクレオチドアレイ [8][9][10] などの，いわゆる DNA マイクロアレイである．その他にも RT-PCR を基礎とした高効率な測定法 [11][12][13] などが用いられる．これらの測定法により得られた大規模データ，すなわち遺伝子発現プロファイルを用いることで，遺伝子間の相互作用や依存関係の定量的解釈による遺伝子の機能の解明の可能性が開けた．発現プロファイルに含まれる多量な情報からを活か

した解析を行うために，様々な統計的解析手法の適用が提案されてきた．この分野での古典的かつ代表的な解析手法としては，発現パターンからの遺伝子の機能分類を試みるクラスタ解析 [14] が挙げられる．また，組織，状態または環境の違いなど，条件が異なる細胞間の遺伝子発現パターンの違いを比較することで，細胞の状態と強く相関を持つ遺伝子の有意性検定を行う発現差解析法 [15][16] や発現パターンを基にした組織診断のための統計的学習法 [17] も代表的な解析手法であり，癌の表現型や薬剤の効果の解析など医学や製薬の分野で応用されている．さらに，野心的な研究として，細胞における遺伝子発現制御系のモデリングが挙げられる．これらでは，遺伝子発現プロファイルから直接遺伝子相互作用ネットワーク構造や簡略化された遺伝子制御ネットワークモデルを推定する試みがなされており，ベイジアンネットワークモデルをはじめとした様々な手法が提案されている [18][19][20].

1.2 遺伝子発現解析における諸問題

遺伝子発現解析は，データ解析の観点から見ると，大きく分けて

1. RNA 量の測定結果の発現情報への数値（データ）化
2. データに対する前処理
3. 目的に合わせた解析手法の適用

の 3 つのステージにより段階的に構成される [21]．このうち最終段階の 3 に該当するのは，前節で述べたような生物学的・医学的知見を得るための具体的な統計的解析である．この解析を成功させるためには，いかに高品質かつ再現性よく目的の RNA 量を定量化できているかが大前提となるため，1, 2 における処理も重要である．以下，それぞれの段階を具体的に見る．

1 の数値化処理では，RNA 測定法ごとに，その物理化学的特性を考慮した固有の解析が必要である．例えば，cDNA マイクロアレイでは，二色の蛍光色素の強度をスキャンしたマイクロアレイ画像からの 1) スポット領域を切り出し，2) 前景色と背景色の分割, 3) 各色のシグナル強度の比の計算という手順に従い数値化

を行う [22] が，各々のステップで精度を高めるための様々な工夫が提案されている．さらに，マイクロアレイ，色素，およびスキャナの性質等に起因する，蛍光強度の空間的な斑，各色間の蛍光強度のバイアス，および蛍光強度と発現比の非線形性などの問題を解消するための様々なノウハウが研究されており，たとえば，LOESS [23] による発現比の正規化法は広く用いられている．

2 の前処理は，1 で得られた発現情報のさらなる補正が主であり，発現プロファイル間のバイアスとバリアンスの正規化や，データ得られなかった遺伝子または値に信頼性のない部位に関する欠測値予測 [24][25] に関する研究がなされている．

3 の実際の解析手法の適用で問題となるのは，遺伝子発現プロファイル特有の次元の高さとノイズである．遺伝子発現プロファイルは，少ない場合で数百，多い場合 1 万を超える遺伝子に関する発現情報を含む．それに対して，ある実験の中で測定されるプロファイル（実験条件，時点，症例など）の数は最も多い場合でも 200 程度とはるかに少ない．このようにデータ数よりも説明変数が多い問題は，文書分類などの自然言語処理を除き，統計的学習でも新しい種類の問題である．さらに，観測データには信号成分とさほど変わらない大きさのノイズが含まれていることも問題である．こうした問題点のため，遺伝子発現解析は，データの多重性を考慮し，高次元のデータから重要な情報を余さず抽出し，また，ノイズにロバストな解析を行うための新たな統計的手法を必要とする．

以上のように，遺伝子発現解析の各段階は様々な問題を含んでおり，それぞれに対して多くの研究がなされてきた．だが，マイクロアレイのように成熟しつつある測定プラットフォームについては，数値化の精度向上のための多くの知識が集約されつつあるが，新たな測定法に関しては十分とは言えず，具体的な方法が確立されていないものもある．また，実際の解析手法の適用に関しては決定的な方法がなく，その目的に応じて，既存の手法を改善した枠組みや，新規な統計的解析手法の開発が進展中であるのが現状である．

2. 論文構成

本論文の構成を説明する．

第 2 章では，RNA 量の数値化法について焦点を当てる．ここでは，包括的 RNA 測定技術の一つである，アダプタ付加競合 PCR (ATAC-PCR) 法により得られる蛍光量データの特徴とその補正法について報告する．これまで問題となっていたアダプタ長依存の測定バイアスの解明を主眼とし，ATAC-PCR 法で得られたデータの詳細な解析を行った．解析結果に基づき蛍光ピーク値に関する観測モデルの定式化を行ない，ノイズ項のパラメータの推定量の導出と，それらを用いたピーク値補正法を提案した．この手法を，アダプタノイズ解析のために特化した採取されたピークデータに適用し，アダプタ依存ノイズのパラメータを求め，次いで，実データに対しバイアス補正の適用を試みた．また，最適な検量線の求め方に関しても調査を行った．

第 3 章では，遺伝子発現ダイナミクスのモデルに基いた，高次元かつノイズを含む遺伝子発現情報からの特徴抽出の 1 手法について述べる．遺伝子発現ダイナミクスの解析を目指し，遺伝子発現プロファイルの時系列に対する解析法について述べる．ここでは，状態空間モデルに基き，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ための新たな手法を提案した．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，より単純かつ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．

第 4 章では，遺伝子発現からの癌の病理診断を想定した，新たな多クラス識別法について述べる．本手法では，多クラス識別問題を一対一ペアや一対残りペアなどのラベルの任意の組み合わせから成る 2 値分類問題群に分解し，各問題での判別結果を統合することによって最適な識別結果を得る．各 2 値分類問題における真の分類確率がクラス所属確率をパラメータとした確率モデルによって生成されると考え，これを 2 値分類器によって得られた分類確率の推定値から推定する方法，さらに 2 値分類器の重みを推定する方法を導いた．本手法を人工データおよび甲状腺がん分類問題をはじめとした実データに適用し，従来のヒューリスティクスによる投票法と同等以上の性能を達成することを示した．さらに，この分野

で提案されてきたいくつかの多クラス識別法との比較を行い，本手法の優位性および性質を明らかにした．

第 5 章で以上の研究を総括する．

第2章　アダプタ付加競合 PCR 法のデータ補正法

1. 序論

1.1 ATAC-PCR 法

遺伝子発現解析において，mRNA の定量化の精度とロバストさを実現する polymerase chain reaction (PCR) 法に基く定量法は，ハイブリダイゼーションに基く定量法以上に多くの利点を有している．PCR 法ベースの mRNA 測定技術を定量的 PCR 法と呼ぶ．基本的な手法としては，現在，real-time reverse transcription (RT)-PCR [26][27] および競合 RT-PCR [28][29] の二つが存在する．これらの手法は，個々の遺伝子のより正確な定量化を目指し様々な改良がなされてきた．だが，そのデータの品質の高さは認められていたものの，大規模なデータ測定に用いた例は限られていた．

アダプタ付加競合 PCR (Adaptor-tagged competitive PCR; ATAC-PCR) 法 [11][30][31] は定量的 PCR 法の改良版の一つで，現状の大規模発現解析で用いられている mRNA 定量法の中では唯一 PCR ベースの手法である．ATAC-PCR 法の大きな特徴は，複数の cDNA サンプルに長さの異なるアダプタプライマーを付加することである．以下，アダプタプライマーを単にアダプタと呼ぶ．また，PCR ベースの手法であるため，マイクロアレイと比較してより少ない RNA 量から定量が行える．その有効性は，癌組織や神経変異性疾患の遺伝子診断による同定などで検証されてきた [32][33][34][35].

1.2 ATAC-PCR 法の特徴と問題点

ATAC-PCR 法が従来の競合 PCR 法と大きく異なるのは，ある cDNA サンプルの全ての遺伝子に対して共通なアダプタを用い 2 種類以上のサンプルを混在させて同時に増幅する点にあるが，これにより 2 種類以上の cDNA サンプルの増幅産物を電気泳動で分画できるようになるため，サンプルごとの cDNA 量を個別に測定することができる．cDNA 量は，蛍光標識した cDNA を DNA シーケンサーを用いて分画しながら読みとり，その蛍光量の観測ピーク値[1]として得る．オリジナルの ATAC-PCR 法では，2 つのアダプタを用いて測定対象サンプルとコントロールサンプルのピーク値を測定し，その比を発現比として得る方法をとっていたが [11]，現在の改良された方法では，7 種類のアダプタを用い，そのうち 3 アダプタを濃度が異なるように調整されたコントロールサンプルに割り当てる [31][32][33][34][35][36]．初期の 2 アダプタの場合では，コントロールサンプルが 1 サンプルのため，この増幅に失敗してしまうと，測定対象サンプルが増幅に成功しても発現比を計算することができない，いわゆる欠測が生じる．改良版では複数のコントロールサンプルを用いるために欠測が起こりにくく，より豊富な発現情報を算出することが可能である．さらに，複数のコントロールサンプルの増幅が成功した場合には，単に測定対象とコントロールのピーク値の比を取るだけでなく，複数のコントロールのピーク値を用いてキャリブレーションカーブを求めることで，より高い精度で発現比を計算することができる．しかし，これまでのところ複数のアダプタを用いることで生じうる増幅効率の変化は検討されておらず，全てのアダプタで増幅効率の変化は生じないという理想的な条件を仮定していた．

本研究では，大規模データ計測という設定のもとで，複数のアダプタの使用が cDNA の増幅結果に与える影響を評価するために，アダプタの違いによるピーク値に対するバイアスの性質を，仮定したピーク値生成の確率モデルを元に解析した．次いで，その結果を踏まえることで，アダプタ依存バイアスの補正できることを確認した．また，キャリブレーションのアルゴリズムと発現比の欠測率および精度の関係を調査した結果，従来の方法よりも欠測率が低く精度の高いキャリ

[1]こうした事実から，本論文では測定された cDNA 量を単純にピーク値と呼ぶ．

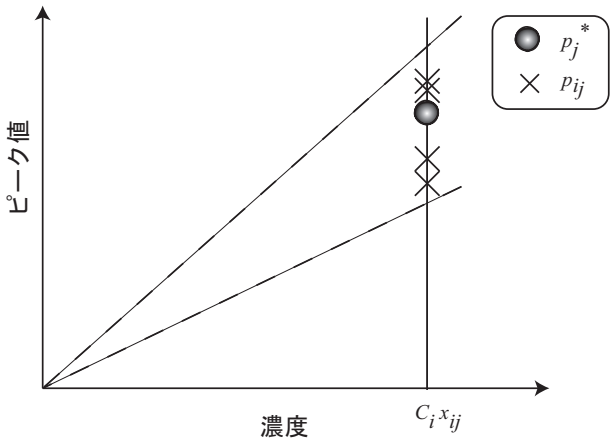

図 2.1 ピーク値生成の模式図

ブレーション法が存在することが分かった.

2. ピーク値生成のモデルとアダプタ依存ノイズ

2.1 ピーク値生成モデル

一回の ATAC-PCR 実験では，アダプタ数と遺伝子数この測定のセットのことを crew と呼ぶ．ある crew k で計測された遺伝子 j，アダプタ i におけるピーク値を p_{ij} とする．このとき p_{ij} は，遺伝子 j についての真のピーク値 p_j^* に遺伝子とアダプタに依存するノイズ項 ϵ_{ij} が加算された,

$$\frac{p_{ij}}{c_i} = \frac{p_j^*}{c_i}(1 + \epsilon_{ij}), \tag{2.1}$$

で表される．ここで，c_i はアダプタ i に対応するサンプルの体積比，x_j は遺伝子 j の真の濃度比である．この p_j^* に対応する真の濃度は $c_i x_j$ である (図 2.1).

ノイズ項 ϵ_{ij} は以下のように平均 μ_{ij}，分散 σ_{ij}^2 のガウシアンに従うものと考える．

$$\epsilon_{ij} = \frac{p_{ij}/c_i}{p_j^*/c_i} - 1 \sim \mathcal{N}(\mu_{ij}, \sigma_{ij}^2). \tag{2.2}$$

ここで μ_i はアダプタ i に依存するバイアス項，σ_i も同様にアダプタに依存する分散である．アダプタと遺伝子間の依存関係は完全に独立であるとみなせるため，これらは

$$\mu_i = \frac{\alpha_i c_i}{f(p_j^*)}, \tag{2.3}$$

$$\sigma_i = \frac{\beta_i c_i}{g(p_j^*)}, \tag{2.4}$$

$$\tag{2.5}$$

とかける．ここで，α_i, β_i はアダプタ i に依存する係数項，$f(\cdot), g(\cdot)$ は真のピーク値の関数である．ピーク値がバイアス項 μ_i と分散項 σ_i へ影響をあたえるとすると $f(p_j^*) = g(p_j^*) = p_j^*$，またピーク値の影響を無視できる場合には，$f(p_j^*) = g(p_j^*) = 1$ を想定する．本モデルでは前者の仮定を採用する（式 (2.2)）．

2.2 真のピーク値の推定量

現実のデータでは，ある遺伝子 j，アダプタ i に対応するピーク値の真の値 p_{ij}^* は観測できない．このため，ϵ_{ij} を求めるためには，真の値に対する推定値を用いる必要がある．ここでは，p_{ij}/c_i のアダプタに関する平均値

$$\frac{\bar{p}_{ij}}{c_i} = \frac{\bar{p}_j}{c_i} = \frac{1}{N_{\text{adaptor}}} \sum_{i=1}^{N_{\text{adaptor}}} \frac{p_{ij}}{c_i}, \tag{2.6}$$

を真の値の推定量として用いる．N_{adaptor} は ATAC-PCR で用いられるのアダプタ数を示す．

この値を用いて，ϵ_{ij} は以下のように計算される．

$$\epsilon_{ij} = \frac{p_{ij}/c_i}{\bar{p}_j/c_i} - 1. \tag{2.7}$$

2.3 ノイズ項のモデルパラメータの推定量

ここは，観測されたピーク値 p_{ij} からのノイズのモデルパラメータ (μ_i, σ_i^2) の推定について述べる．

ϵ_{ij} は以下の

$$p(\epsilon_{ij}|\mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(\epsilon_{ij} - \mu_i)^2\right), \tag{2.8}$$

で表される．このパラメータの最尤推定量 $\hat{\mu}_i, \hat{\sigma}_i^2$ は

$$\hat{\mu}_i = \frac{1}{N_{\text{gene}}} \sum_{j=1}^{N_{\text{gene}}} \epsilon_{ij}, \tag{2.9}$$

$$\hat{\sigma}_i^2 = \frac{1}{N_{\text{gene}}} \sum_{j=1}^{N_{\text{gene}}} (\epsilon_{ij} - \hat{\mu}_i)^2, \tag{2.10}$$

と求まる．ここで N_{gene} は一つの crew で計測された遺伝子数である．さらに式 (2.6) を用いて，ピーク値からのパラメータ推定量は結局

$$\hat{\mu}_i = \frac{1}{N_{\text{gene}}} \sum_{j=1}^{N_{\text{gene}}} \left(\frac{p_{ij}/c_i}{\bar{p}_{ij}/c_i} - 1\right), \tag{2.11}$$

$$\hat{\sigma}_i^2 = \frac{1}{N_{\text{gene}}} \sum_{j=1}^{N_{\text{gene}}} \left(\frac{p_{ij}/c_i}{\bar{p}_{ij}/c_i} - 1 - \hat{\mu}_i\right)^2, \tag{2.12}$$

となる．

2.4 バイアスの補正

アダプタに依存するバイアス項 μ_i を用いて以下のように，バイアスの補正をおこなうことができる．アダプタ i，遺伝子 j のピーク値の補正後の値 $\tilde{p}_{ij}$ は次のように計算できる．

$$\begin{aligned} \tilde{p}_{ij} &= p_{ij} - \mu_i p_j^* \\ &= p_{ij} - \mu_i p_{ij}. \end{aligned} \tag{2.13}$$

3. 解析 1. アダプタ依存バイアスの推定と補正

本解析の基本的な目的は，補正のためのアダプタ依存のバイアスパラメータ μ_i の推定である．また，求めたパラメータ用いた未知 crew データのピーク値に対するバイアス補正を行ない，バイアス補正の効果の評価を行なう．

表 2.1 各 crew におけるサンプルの体積 (単位: μl)

adaptor	crew 1	crew 2	crew 3	crew 4	crew 5	crew 6	crew 7	crew 8	crew 9	crew 10
MB1	5	15	50	150	5	5	5	50	50	50
MB2	5	15	50	150	5	5	5	50	50	5
MB3	5	15	50	150	10	10	10	50	5	5
MB4	5	15	50	150	10	10	10	10	10	10
MB5	5	15	50	150	5	5	50	10	10	10
MB6	5	15	50	150	5	50	50	5	5	5
MB7	5	15	50	150	50	50	50	5	5	5
total volume	35	105	350	1050	90	135	180	180	135	90

3.1 データ概要

解析対象は全 10 crew の ATAC-PCR のピーク値データのセットである（表 4.3）. 各 crew につき，7 アダプタを用いて 384 個の各遺伝子について 7 サンプルのピーク値が得られている．これらのうち crew 4 に関しては，アダプタ依存のバイアス解析のために特化して測定したデータである．ある crew 内では全てのサンプルが同一体積比の同一サンプルが各アダプタに割り当てられている．4 つの crew の違いは，そのサンプルの体積にある．また，残り 6 crew はサンプルのアダプタに対する体積比の割り当てと総体積を変えて同一サンプルについて測定されたデータである．つまり，ある crew においてあるサンプルをコントロールとした場合の他のサンプルの遺伝子の対数発現比は，さまざまなノイズを完全に排除された理想的なコンディションでは全て $\log_2 1 = 0$ となる．

3.2 ノイズの分布とアダプタ依存バイアスの推定結果

crew 1 から crew 10 ついて，それぞれのノイズの分布を図 2.2 に示す．crew 1 から crew 4 について，ノイズのバイアスと分散を推定したものを，それぞれ表 2.2, 表 2.3 に示す． 本結果から，短いアダプタほどバイアスが大きく，長いものほど小さい傾向がみられる．これは，バイアスがアダプタ長に依存することを示唆している．

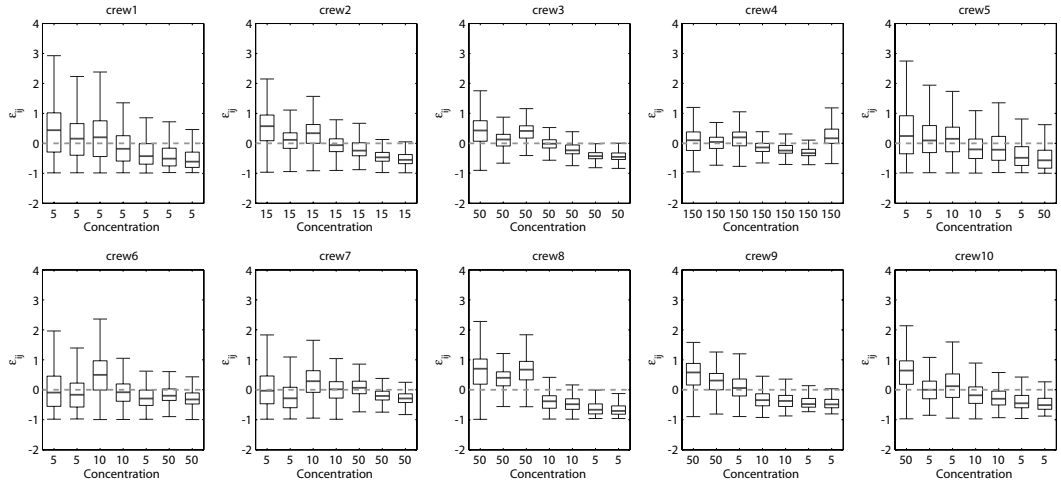

図 2.2 データセットにおけるピークノイズ ϵ_{ij} の分布

表 2.2 バイアスの推定値

crew ID	MB-1	MB-2	MB-3	MB-4	MB-5	MB-6	MB-7
crew 1	0.441	0.157	0.205	-0.181	-0.426	-0.511	-0.61
crew 2	0.572	0.116	0.339	-0.059	-0.240	-0.467	-0.549
crew 3	0.429	0.129	0.414	-0.012	-0.231	-0.426	-0.451
crew 4	0.103	0.045	0.199	-0.134	-0.239	-0.327	0.170

3.3 バイアス補正の適用

バイアス補正用のパラメータは，crew 2 と crew 3 のバイアスの平均値とした．これを，表 2.4 に示す．crew 2 と crew 3 の平均値を用いた理由としては，それらの crew のサンプルの体積が，それぞれ 15μl と 50μl であり，crew 4 から crew 10 までのサンプルの体積と概ね同じレベルであることと，2 つの crew 間のバイアスの誤差が小さいことの 2 点が挙げられる．

このパラメータを用いて，crew 5 から crew 10 までのデータについてバイアス補正を行なった．この結果得られた，各補正後のデータに関してノイズを再び求めた．その分布を図 2.3 に示す．

表 2.3 分散の推定値

crew ID	MB-1	MB-2	MB-3	MB-4	MB-5	MB-6	MB-7
crew 1	0.797	0.641	0.621	0.489	0.464	0.393	0.346
crew 2	0.526	0.249	0.306	0.230	0.197	0.117	0.197
crew 3	0.406	0.184	0.195	0.161	0.0944	0.0481	0.0905
crew 4	0.218	0.106	0.134	0.118	0.0771	0.0521	0.376

表 2.4 バイアス補正用パラメータ

MB-1	MB-2	MB-3	MB-4	MB-5	MB-6	MB-7
0.501	0.122	0.377	-0.0364	-0.236	-0.446	-0.5

4. 解析 2. キャリブレーション法の検討

本解析では，コントロールサンプルに対応するアダプタが異なる 6 つのデータセット (crew) に対する，バイアス補正およびキャリブレーションの性能の評価を行ない，最適なコントロールサンプルの取り方とピークデータの処理法についての知見を得ることが目的である．

4.1 キャリブレーション法

キャリブレーションとは，コントロールサンプルの体積 (mRNA 量に対応) と計測されたピーク値を用い検量線を求めることである．これにより，ターゲットサンプルの mRNA 量の推定が可能となる．本解析では，以下の 3 種類のキャリブレーション法を用いた．

K 法 原点および各コントロールサンプルを通過する区分線形関数によりキャリブレーションカーブを構成する．

L0 法 得られたコントロールサンプルを用いて最小二乗法により線形関数の傾きのみ決定し，切片に関しては原点を通るキャリブレーションカーブを構成

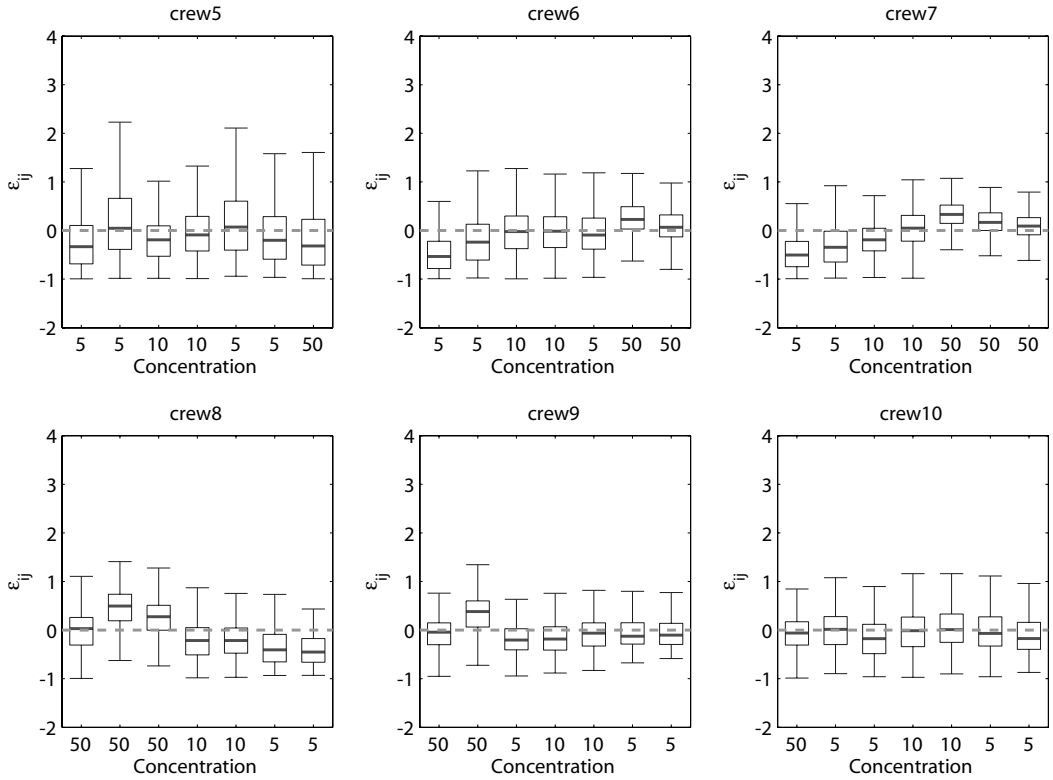

図 2.3 補正後データセットにおけるピークノイズ ϵ_{ij} の分布

する．

L 法 得られたコントロールサンプルを用いて最小二乗法により線形関数を求める．ただし，コントロールサンプルが一点しか得られなかった場合は，L0 同様に原点を通るキャリブレーションカーブを求める．

図 4.1 に各キャリブレーション法の模式図を示した．

これらの方法はいずれも，特定のコントロールサンプル間の体積とピーク値の関係が線形であるという仮定に基くものである．K 法は，仮に体積とピーク値の間に非線形性が認められた場合には有効に働くと考えられる．しかし， 最大で 2 つのコントロールサンプルしか用いないため，全コントロールサンプル間の非線形性を無視できる場合には，最大 3 つのコントロールサンプルを用いることの出来る L 法の方が有利であると考えられる．L0 法は L 法同様に全てのサンプルの

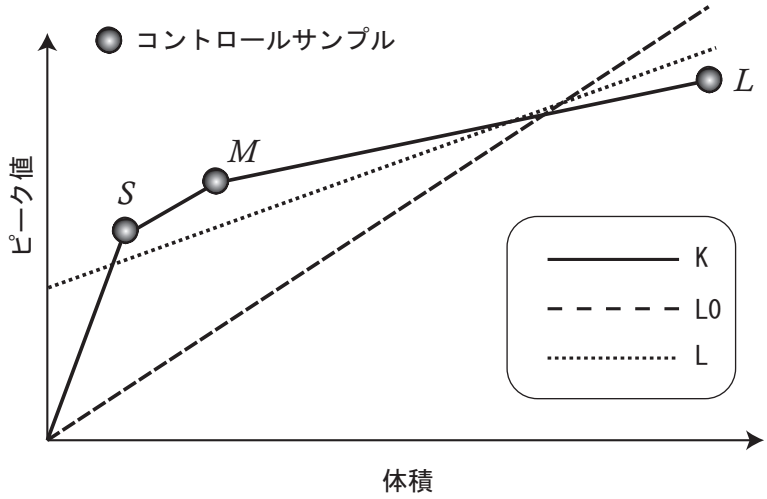

図 2.4 コントロールサンプルによるキャリブレーションの模式図．3 種類のキャリブレーションキャリブレーション法によって得られる検量線を示した．コントロールサンプルは体積の小さいものから順に *S*, *M*, *L* で示した．

線形性を仮定したものである．L 法では図 4.1 の例のように正の切片が得られた場合，それ以下のピーク値では体積の推定値が負となり欠測として扱うのに対し，L0 法では切片を 0 に補正することで欠測が抑えられる．ただし，欠測抑制の効果と同時に検量精度が犠牲となることもありうる．

4.2 コントロールパターン

最大で 3 つしか得ることの出来ないコントロールサンプルのピーク値は，キャリブレーションの精度に大きな影響を与える．コントロールパターンとは，真の mRNA 量に対応するコントロールサンプルの体積の大きさと，それらに対応するピーク値の大きさの順位の関係を示す．コントロールサンプルのピークが出ない場合も考慮し，全 16 パターンを表 2.5 のように定義した．

コントロールパターンを大別すると，コントロールサンプルの体積とピーク値の大小関係が一致する理想的なもの (コントロールパターン 1,7,9,11,13,14,15) と，体積とピーク値の大小関係が一致していないコントロールを含むものの二群とな

表 2.5 コントロールパターン

control pattern	magnitude relation of peak values
1	$S < M < L$
2	$S < L < M$
3	$M < S < L$
4	$M < L < S$
5	$L < S < M$
6	$L < M < S$
7	$S < M$
8	$M < S$
9	$S < L$
10	$L < S$
11	$M < L$
12	$L < M$
13	S
14	M
15	L
16	全て欠測

表 2.6 キャリブレーションに用いたコントロールに対応するアダプタ

	5 μl	10 μl	50 μl
crew 5	MB-2	MB-4	MB-7
crew 6	MB-5	MB-3	MB-7
crew 7	MB-2	MB-4	MB-6
crew 8	MB-7	MB-4	MB-1
crew 9	MB-3	MB-4	MB-1
crew 10	MB-2	MB-4	MB-1

る．また，それぞれで少なくとも 1 つ以上のコントロールサンプルに欠測が見られるものが存在する．理想的には，コントロールパターン 1 のみでキャリブレーションを行うのが望ましい．しかし，実際に計測されたデータには様々なコントロールパターンのものが存在するため，キャリブレーションの精度とデータの取得率を両立させたキャリブレーションが求められる．

4.3 キャリブレーション法の評価とコントロールの選択

解析対象

Crew

キャリブレーションに用いるデータセットは，表 4.3 における，crew 5 から crew 10 までの全 6 crew を用いる．

コントロールサンプル

本解析では，3 点のコントロールを用いて各キャリブレーションアルゴリズムでキャリブレーションカーブを計算し，これをもとに対数発現比を求める．全ての crew において，3 つのコントロールサンプルの体積はそれぞれ，5 μl, 15 μl, 50 μl と共通である．ただし，コントロールサンプルに対応するアダプタは，各 crew で異なるように設定した．以上のコントロールとアダプタの対応関係を 表 2.6 に示す．

ターゲットサンプル

ここで，ターゲットサンプルとは，crew における 7 点のアダプタに割り当てられているサンプルのうち，コントロールに用いる 3 サンプル以外の 4 サンプルを示す．つまり，コントロールサンプルのピーク値から計算したキャリブレーションカーブに対して，実際に発現量を求めるためのサンプルである．4 サンプルあるターゲットサンプルの体積は crew 毎に異なり，5μl, 10μl, 50μl のいずれかの組合せである (表 4.3).

手順

解析の手順を以下に示す.

1. 各 crew について，バイアス補正なし（未処理），バイアス補正ありのデータを用意することで，計 $6 \times 2 = 12$ 種類の crew を考える．
2. 各 crew のコントロールパターンの頻度を求める．
3. 各 crew に対して 3 通りのキャリブレーション法を適用し，ターゲットサンプルの対数発現比を計算する．これにより，$12 \times 3 = 36$ 通りのキャリブレーション結果が求まる．
4. 各結果の RMSE (root mean squared error) と，キャリブレーションされたサンプルの割合（成功率）を評価する．

解析の結果得られた RMSE とキャリブレーションの成功率をそれぞれ表 2.8 および 表 2.7 に示した．また，各 crew のコントロールタイプ別の解析結果の詳細は，付録 A に示した.

この結果から次のことが分かる．

- コントロールサンプルは，体積の大きさとアダプタの長さを対応させてとった場合に，RMSE が低くなる傾向にある．crew 5,6,7 がこのケースに該当するデータである．

表 2.7 crew 毎のキャリブレーションの成功率 (%)

crew ID	raw			bias corrected		
	K	L0	L	K	L0	L
crew 5	68.55	92.9	62.7	80.92	92.9	68.23
crew 6	89.19	96.94	68.29	94.86	96.94	86.07
crew 7	93.88	98.11	91.93	94.4	98.11	93.42
crew 8	92.32	94.27	91.47	88.8	94.27	89.26
crew 9	66.34	67.32	65.49	65.69	67.32	64.52
crew 10	71.61	72.72	69.01	70.7	72.72	62.11
all	80.32	87.04	74.82	82.56	87.04	77.27

- バイアス補正したデータと，補正しなかったデータに関するキャリブレーション結果の RMSE を比較した場合，補正後データのものの方が小さくなる傾向にある．
- キャリブレーション法の中では，RMSE での評価基準では，L 法がもっともその性能が優れている．ただし，欠測率に関しては K 法が有利である．

以上から，次のデータ処理がもっとも有効であると考えられる．まず，ピーク値に関してはバイアス補正を行なう．次いで，バイアス補正を行なったデータのキャリブレーションは基本的には L 法で行ない，それでとりこぼしたサンプルについては，K 法でカバーする方法が最良であると考えられる．

5. 結論

本研究では，ATAC-PCR 実験により得られたピーク値の生成モデルをを利用した，アダプタ依存バイアスに特化した補正手法を提案した．アダプタノイズ解析のために特化した採取されたピークデータを用いて，アダプタ依存ノイズのパラメータを求め，実データに対しバイアス補正を適用した結果，適切にバイアスを補正できることが示された．また，キャリブレーション法の違いによる発現比の

表 2.8 crew 毎の RMSE

crew ID	raw			bias corrected		
	K	L0	L	K	L0	L
crew 5	2.13	1.92	2.37	2.15	2.08	2.05
crew 6	1.73	1.77	1.69	2.1	2.27	1.56
crew 7	1.35	1.49	1.32	1.61	2.04	1.41
crew 8	1.62	2.51	1.39	1.41	1.71	1.36
crew 9	1.89	2.25	1.47	1.2	1.28	1.34
crew 10	1.85	2.45	1.42	1.58	1.53	1.69
all (crew 5-10)	1.75	2.07	1.61	1.73	1.89	1.57

精度・欠測率への影響を網羅的に調べることにより，バイアス補正によるキャリブレーションの成功率および精度の向上と適切なキャリブレーションの方法に関する知見が得られた.

第3章　線形ダイナミカルシステムモデルによる遺伝子発現時系列のシステム同定

1. 序論

1.1 問題設定

細胞の機能を分子レベルで理解するためには，どの遺伝子が，いつ，どんな条件で，どの細胞内小器官で，どれだけ発現しているのかを詳細に知る必要がある．しかし，生物における遺伝子の発現制御は，核酸や酵素から低分子にいたるまでの多くの構成要素の複雑な相互作用により形成される制御ネットワークを通して実現されている．そのため，たとえ細胞内の生体分子に関する濃度情報が詳細に得られたとしても，それらのダイナミクスを直接解釈するのは困難である．現在，この問題に対して mRNA 定量化技術を用いて細胞の状態と包括的な遺伝子発現量の関係を遺伝子発現プロファイルとして蓄積し，これらに対して統計学的，情報科学的な手法に基づく解析を行うアプローチが取られている [18].

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．細胞には数千から数万の遺伝子が含まれており，その個別の発現挙動は複雑な制御ネットワークによる．一方で，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おり，そのことが生物の恒常性維持のために重要である．こうした仮説に基づき，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 [37, 38, 39].

1.2 遺伝子発現時系列の状態空間モデル

遺伝子発現プロセスの状態空間モデルでは，観測系列は遺伝子発現量の時間変化に対応し，非観測な内部状態空間における状態変数および遷移行列は，細胞における潜在的な上位システムあるいは外部環境，すなわち，遺伝子発現を制御する因子を仮想している．遺伝子発現のダイナミクスの複雑さを知ることは，システム同定という工学的逆問題において重要のみならず，生物の恒常性維持と環境適応との競合のメカニズムを知る上で手がかりとなる．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，状態空間モデルにおいて，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．特に文献 [37] では，観測行列と内部状態変数の推定を因子分析の問題として定式化している．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を扱うには十分とはいえない．また，状態空間モデルにおけるシステムノイズは基底の滑らかさに貢献するため，大量の遺伝子の大域的な変化の特徴を捉えるためには重要な要素だと考えられる．

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，生きた細胞内で動的に変化する発現制御因子の挙動と，個々の遺伝子の特徴を同時に解析する手法を提案する．LDS モデルは，白色ガウスノイズをともなうガウス過程モデルの一つであり，時間変化する内部状態変数の系列から観測系列が生成されるものとする．

因子分析モデルや LDS の生成モデルは，内部状態変数とパラメータを持つ指数族に属するため，EM アルゴリズムによる最尤 (maximum likelihood; ML) 推定法を用いて推定することができる [40, 41]．ここで内部状態変数の次元数の同定が問題となるが，ML 推定法では複雑なモデルほど選択されやすいため，次元同定が困難である．この問題に対する一つの解決法としては，情報量基準を用いて

モデル選択を行う方法が挙げられる．たとえば，因子分析モデル [37] では次式で定義される Bayesian information criterion (BIC) [42] によるモデル選択を行っている．

$$BIC \equiv -2L + F \log_2 n.$$

ここで，L は推定されたモデルの対数尤度，F はモデルに含まれるパラメータ数，n は標本数である．BIC はモデルの自由度に対する罰則付きの負の対数尤度を表すため，この値が最小のモデルが，データを表現するのに適切なモデルとして選択される．これに対し，ベイズ推定法の一手法である，変分ベイズ (variational Bayes; VB) 法 [43] は，有効なパラメータ推定アルゴリズムであるとともに，特にデータ数が不十分な時などで情報量基準よりも有効なモデル選択法として用いることもできる．

本研究では，LDS モデルの 変分ベイズ推定法 [44][45] を用いて，ロバストなパラメータ推定と，システムの複雑さに関わる状態空間の次元の同定を行う．また，学習により得られた LDS モデルのパラメータのうち，特に観測行列に着目した．これは遺伝子の特徴を表す特徴ベクトルの集合と解釈できることから，遺伝子に関する既知の生物学的研究との比較を行うことで，手法の有効性を検討した．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内部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の各位相における遺伝子発現についての公開データ [46] を用いて，モデルパラメータ推定を試み，本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．通常，各遺伝子の発現レベルとして，測定対象サンプルとコントロールサンプルの対数発現比が用いられる．

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，

$$\boldsymbol{y}_t = (y_{t1}, \ldots, y_{tD})'; \quad t = 1, \ldots, T \tag{3.1}$$

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．

$$\begin{aligned}
\boldsymbol{x}_t &= \boldsymbol{W}\boldsymbol{x}_{t-1} + \boldsymbol{\epsilon}_t; \quad t = 2, \ldots, T, && (3.2)\\
\boldsymbol{y}_t &= \boldsymbol{V}\boldsymbol{x}_t + \boldsymbol{\eta}_t; \quad t = 1, \ldots, T, && (3.3)\\
\boldsymbol{x}_1 &\sim \mathcal{N}_N(\boldsymbol{x}_1 | \boldsymbol{\mu}_1, \sigma_1^2 \boldsymbol{I}_N), && (3.4)\\
\boldsymbol{\epsilon}_t &\sim \mathcal{N}_N(\boldsymbol{\epsilon}_t | \boldsymbol{0}_N, \sigma_\epsilon^2 \boldsymbol{I}_N), && (3.5)\\
\boldsymbol{\eta}_t &\sim \mathcal{N}_D(\boldsymbol{\eta}_t | \boldsymbol{0}_D, \sigma_\eta^2 \boldsymbol{I}_D) && (3.6)
\end{aligned}$$

$\boldsymbol{x}_1$ は $\boldsymbol{x}$ の初期値である．$\boldsymbol{\epsilon}_t \in \mathcal{R}^N$ および $\boldsymbol{\eta}_t \in \mathcal{R}^D$ はそれぞれシステムノイズと観測ノイズである．これらのノイズは正規分布に従うものと仮定する．ここで，システムノイズと観測ノイズを共に一般的な正規分布でモデル化するのは，ノイズモデルに関して先験的な知識がないこと，推定の対象となるパラメータ数を抑え

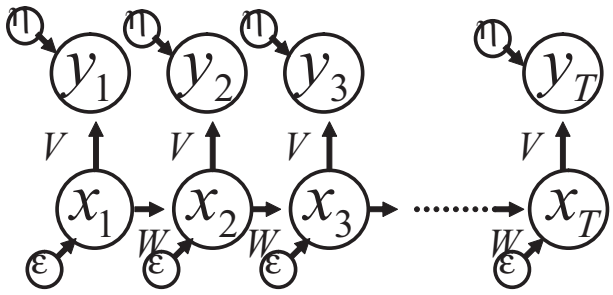

図 3.1 線形ダイナミカルシステムモデルの模式図

て推定をロバストにすること，および，状態変数の推定がロバストに行えるようにするためである．なお，

$$\mathcal{N}_p(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{S}) \equiv (2\pi)^{\frac{-p}{2}}|\boldsymbol{S}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{S}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

は，平均 $\boldsymbol{\mu}$，共分散行列 $\boldsymbol{S}$ の p 次元正規分布の確率密度関数である．

$\boldsymbol{\mu}_1 \in \mathcal{R}^N$ 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．$\sigma_1^2, \sigma_\epsilon^2, \sigma_\eta^2$ はそれぞれ $\boldsymbol{x}_1, \boldsymbol{\epsilon}_t, \boldsymbol{\eta}_t$ の分散である．$\boldsymbol{\theta} \equiv \{\boldsymbol{\mu}_1, \sigma_1^2, \boldsymbol{W}, \sigma_\epsilon^2, \boldsymbol{V}, \sigma_\eta^2\}$, が，モデルパラメータのセットとなる．モデルの模式図を図 3.1 に示した．

状態変数 $\boldsymbol{x}_t$ と観測変数 $\boldsymbol{y}_t$ に関するシステム方程式 (3.2), (3.3) と白色ガウシアンノイズの仮定 (3.4), (3.5), (3.6) より，以下の確率モデルが導かれる．

$$\begin{aligned} &p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{\theta}) \\ &= \begin{cases} \mathcal{N}_N(\boldsymbol{x}_1|\boldsymbol{\mu}_1, \sigma_1^2\boldsymbol{I}_N) & t = 1, \\ \mathcal{N}_N(\boldsymbol{x}_t|\boldsymbol{W}\boldsymbol{x}_{t-1}, \sigma_\epsilon^2\boldsymbol{I}_N) & t = 2, \ldots, T, \end{cases} \end{aligned}$$

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\theta}) = \mathcal{N}_D\left(\boldsymbol{y}_t|\boldsymbol{V}\boldsymbol{x}_t, \sigma_\eta^2\boldsymbol{I}_D\right)$$

以上をまとめると，完全データ $\boldsymbol{X}_{1:T} \equiv \{\boldsymbol{x}_t\}, \boldsymbol{Y}_{1:T} \equiv \{\boldsymbol{y}_t\}$ に対するモデルパラメータ $\boldsymbol{\theta}$ の尤度関数

$$p(\boldsymbol{Y}_{1:T}, \boldsymbol{X}_{1:T}|\boldsymbol{\theta}) = \prod_{t=1}^{T} p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{\theta})p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{\theta}) \tag{3.7}$$

が得られる．

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仮定した．

$$
\begin{aligned}
p(\boldsymbol{\mu}) &= \mathcal{N}_N\left(\boldsymbol{\mu}|\mathbf{0}, \gamma_0^{-1}\boldsymbol{I}_N\right), && (3.8)\\
p(\sigma_1^2) &= \mathcal{G}\left(\sigma_1^{-2}|\gamma_0, \gamma_0\tau_{\mu 0}\right), && (3.9)\\
p(\boldsymbol{W}) &= \prod_{i=1}^{N} \mathcal{N}_N\left(\boldsymbol{w}_i|\mathbf{0}_N, \gamma_0^{-1}\boldsymbol{I}_N\right), && (3.10)\\
p(\sigma_\epsilon^2) &= \mathcal{G}\left(\sigma_\epsilon^{-2}|\gamma_\epsilon, \gamma_\epsilon\tau_{\mu_\epsilon}\right), && (3.11)
\end{aligned}
$$

$$
\begin{aligned}
p(\boldsymbol{V}) &= \prod_{j=1}^{D} \mathcal{N}_D\left(\boldsymbol{v}_j|\mathbf{0}_D, \gamma_0^{-1}\boldsymbol{I}_D\right), && (3.12)\\
p(\sigma_\tau^2) &= \mathcal{G}\left(\sigma_\tau^{-2}|\gamma_\tau, \gamma_\tau\tau_{\mu_\tau}\right) && (3.13)
\end{aligned}
$$

ただし，$\mathcal{G}(\sigma^{-2}|\gamma, \gamma\tau)$ は，

$$
\mathcal{G}(\sigma^{-2}|\gamma, \gamma\tau) \equiv \frac{(\gamma\tau)^\gamma (\sigma^{-2})^{\gamma-1}}{\Gamma(\gamma)} \exp[-\gamma\tau\sigma^{-2}],
$$

で定義されるガンマ分布である．$\boldsymbol{w}_i$ 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．無情報事前分布を実現するために，$\gamma_0 = 0.0001$, $\gamma_{\epsilon 0} = \gamma_{\eta 0} = 0.01$, $\tau_{\epsilon 0} = \tau_{\epsilon 0} = 0.01$ を用いた．また後の適用実験では，$\tau_{\mu 0}$ を遺伝子ごとの分散の 1 遺伝子当たりの平均値とした．

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 $\boldsymbol{v}_i \in \mathcal{R}^{1\times N}, i = 1, \ldots, D$ は内部状態変数 $\boldsymbol{x}_t$ 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 変分ベイズ推定

観測変数の系列 $\boldsymbol{Y}$ が与えられた時，未知変量に関する事後分布 $p(\boldsymbol{X}, \boldsymbol{\theta}|\boldsymbol{Y})$ を求めることがベイズ推定の目的である．この事後分布はベイズの定理により，以下で与えられる．

$$
\begin{aligned}
p(\boldsymbol{X}, \boldsymbol{\theta}|\boldsymbol{Y}) &= \frac{p(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{Y})}{p(\boldsymbol{Y})}, && (3.14)\\
p(\boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{Y}) &= p(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta})p(\boldsymbol{\theta}), && (3.15)\\
p(\boldsymbol{Y}) &= \int p(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{\theta})p(\theta)d\boldsymbol{\theta}d\boldsymbol{X} && (3.16)
\end{aligned}
$$

正規化項 $P(\boldsymbol{Y})$ は，周辺化尤度と呼ばれ，LDS モデルにおける内部状態変数の次元 N に対する尤度を表すため，モデル選択のための指標として用いることができる [47]．LDS モデルは内部状態変数についてはガウス過程モデルの一種であるが，パラメータと内部状態変数の事後分布および周辺化尤度を解析的に求めることは困難である．このため，本研究では変分ベイズ法を用い，事後分布および周辺化尤度の近似計算を行う．

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．

$$
\begin{aligned}
\ln p(\boldsymbol{Y}) &\equiv \ln \int p(\boldsymbol{Y}, \boldsymbol{X}|\boldsymbol{\theta})p(\theta)d\boldsymbol{\theta}d\boldsymbol{X}\\
&\geq \int q(\boldsymbol{\theta}, \boldsymbol{X}) \ln \frac{p(\boldsymbol{Y}, \boldsymbol{X}|\theta)p(\boldsymbol{\theta})}{q(\boldsymbol{\theta}, \boldsymbol{X})} d\boldsymbol{\theta}d\boldsymbol{X}\\
&= \log p(\boldsymbol{Y}) - \mathrm{KL}(q(\boldsymbol{X}, \boldsymbol{\theta}) \| p(\boldsymbol{X}, \boldsymbol{\theta}|\boldsymbol{Y}))\\
&\equiv \mathcal{F}[q(\boldsymbol{\theta}, \boldsymbol{X})] && (3.17)
\end{aligned}
$$

ここで，$\mathrm{KL}(\cdot\|\cdot)$ は二つの分布間の Kullback-Leibler 情報量であり，$q(\boldsymbol{\theta}, \boldsymbol{X}) = p(\boldsymbol{X}, \boldsymbol{\theta}|\boldsymbol{Y})$ で最小値 0 となる．

$\mathcal{F}[q(\boldsymbol{\theta}, \boldsymbol{X})]$ の最大化は，独立分解近似

$$\begin{aligned} q(\boldsymbol{\theta}, \boldsymbol{X}) &= q(\boldsymbol{\theta})q(\boldsymbol{X}), \\ q(\boldsymbol{X}) &= \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}), \\ q(\boldsymbol{x}_1|\boldsymbol{x}_0) &\equiv q(\boldsymbol{x}_1) \end{aligned} \tag{3.18}$$

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 [47].

自由エネルギーは，前述の BIC と比較した場合，指数分布族で与えられる確率モデルに対して，より有効なモデル選択基準である．LDS モデルでは事後分布が単峰に近い可能性があるため，変分ベイズ法はサンプリング手法に迫る性能を，大幅に少ない計算コストで実現できると考えられる [48].

3. 関連研究

3.1 状態空間モデルの先行研究

状態空間モデルを用いた発現プロファイルからのシステム同定を目指した先行研究では，内部状態変数のダイナミクスとノイズの過程を無視した簡略なモデルを想定し，特異値分解 [38] や因子分析に対する EM アルゴリズム [37] により状態変数とパラメータを求める手法が提案されているが，現状で発現制御因子（内部状態）数の推定まで踏み込んでいるのは Wu らの研究 [37] のみである．Wu らは，因子分析モデルにおける自由度を，最尤推定の結果から得られる BIC を用いて決定することにより，発現制御因子数の推定を試みた.

因子分析モデル

Wu らの因子分析モデルは，

$$\boldsymbol{Y} = \boldsymbol{V}\boldsymbol{X} \tag{3.19}$$

で定義される．ここで，$\boldsymbol{Y}$ は T 点の発現プロファイルをまとめた $D \times T$ の遺伝子発現行列，$\boldsymbol{V}$ は LDS モデルと同様の $D \times N$ の観測行列，$\boldsymbol{X}$ は T 点の状態変数ベクトルからなる $N \times T$ の状態変数行列であり，これらがモデルパラメータとなる．

パラメータ推定では，与えられたデータ $\boldsymbol{Y}$ に対し，因子分析モデルの EM アルゴリズムによる推定 [49] により，因子得点行列と因子負荷行列に対応する $\boldsymbol{V}$ と $\boldsymbol{X}$ を求める．

3.2 その他の関連研究

状態空間モデル以外の遺伝子発現時系列データの解析法では，S-systems による非線形微分方程式モデルや [50, 51] ブーリアンネットワークモデル [52, 53] に基いた手法が代表的なものとして挙げられる．S-systems によるモデル化では，mRNA やタンパク質など生体分子の濃度変化のダイナミクスをそれぞれ微分方程式により記述し，連立微分方程式を構成する．そして，各種数理最適化手法によりデータから係数を同定する．ブーリアンネットワークモデルでは，各遺伝子の発現に応じて状態を二値化し，その制御規則をブール関数で表す．データからその状態遷移規則を学習することで，遺伝子制御ネットワークの構造推定を行うことができる．ブーリアンネットワークモデルと線形計画法を組み合わせて，微分方程式モデルの係数を最適化する手法も提案されている [54].

我々が提案する LDS モデル を含む状態空間モデルと，これらのモデルが大きく異なる部分は，前者が遺伝子間の相互作用を陽に仮定せず，全 D 個の遺伝子に共通する N 個の遺伝子制御因子により発現が駆動されるとするのに対し，後者では，個々の遺伝子が個別に相互作用を持つことで発現が制御されるとする点にある．このため，状態空間モデルでは $N = D$ という特殊なケースを除くと，遺伝子間の相互作用を直接扱うことができない．だが，その代わりに，観測ベクトル v_i を，内部状態変数のダイナミクスのモデルに応じて低次元に射影された遺伝子発現ベクトルとみなし，遺伝子間の類似度を評価できる．これが状態空間モデルならではの特徴である．特に，LDS モデルでは，連続的に時間変化する遺伝子制御因子のみを抽出するために，従来の因子分析では規定されていなかった線形の内部状態変数のダイナミクスを考えているため，大域的な遺伝子を制御する滑ら

かな基底と，それに対応した遺伝子の情報の両者の抽出が期待できる．

4. 実験と結果

4.1 実験 1. 人工データによる評価

状態空間に含まれるダイナミクスを持つ成分を抽出できるかどうかを評価するために，人工データを用いて提案手法の性能評価を行った．

人工データ

まず，状態遷移行列

$$\boldsymbol{W} = \begin{bmatrix} 0.9071 & 0.7655 & -0.2499 \\ -0.3238 & 0.7116 & -0.2128 \\ 0.6780 & 0.0002 & -0.2133 \end{bmatrix}$$

を持つ LDS モデル ($N = 3$) を用いて，15 時点 ($T = 15$) の内部状態変数系列を生成した．これらの内部状態変数系列に加え，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，合計 5 つの内部状態変数系列を得た（図 3.2 左）．ここで，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（図 3.2 右）．観測ノイズの標準偏差 σ_η は 0.05 とした．

システム同定

生成したデータ $\boldsymbol{Y}_{1:15}$ を用いて，LDS モデルを用いたパラメータ推定，およびモデル選択を行った．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．

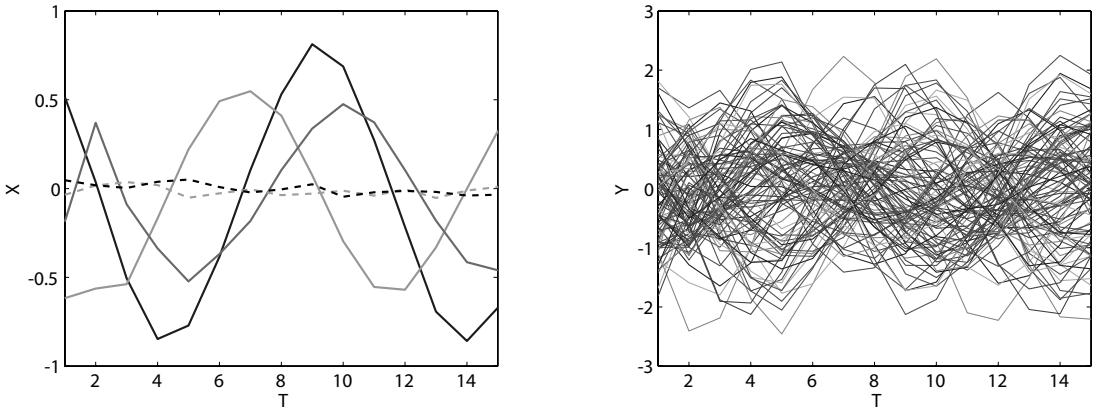

図 3.2 人工時系列データ．左図は 5 つの内部状態変数の時系列，右図は内部状態変数と観測行列から生成された 100 サンプルの観測系列を示す．

図 3.3 は，LDS モデルおよび因子分析モデルのそれぞれのモデル選択基準である自由エネルギーおよび BIC をプロットしたものである．自由エネルギー最大化の観点からモデル選択を行うと，データ $\boldsymbol{Y}_{1:15}$ について，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った．図 3.4 は，人工データに関する LDS モデルの $N = 2, \ldots, 9$ におけるシステムノイズと観測ノイズの標準偏差の推定値を示す．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ている．これは，LDS で選択されたモデルの内部状態変数の次元と一致しており，本 LDS モデルと変分ベイズによる推定では，システムノイズを最小化する方向でモデル選択が行われたと考えられる．

以上より，LDS モデルでは，そのダイナミクスに従う内部状態変数の成分の数 $N = 3$ を自動的に検出できることが示された．これに対し，因子分析モデルでは，ダイナミクスを持つ成分のみならず，無情報な成分も検出してしまう結果となった．

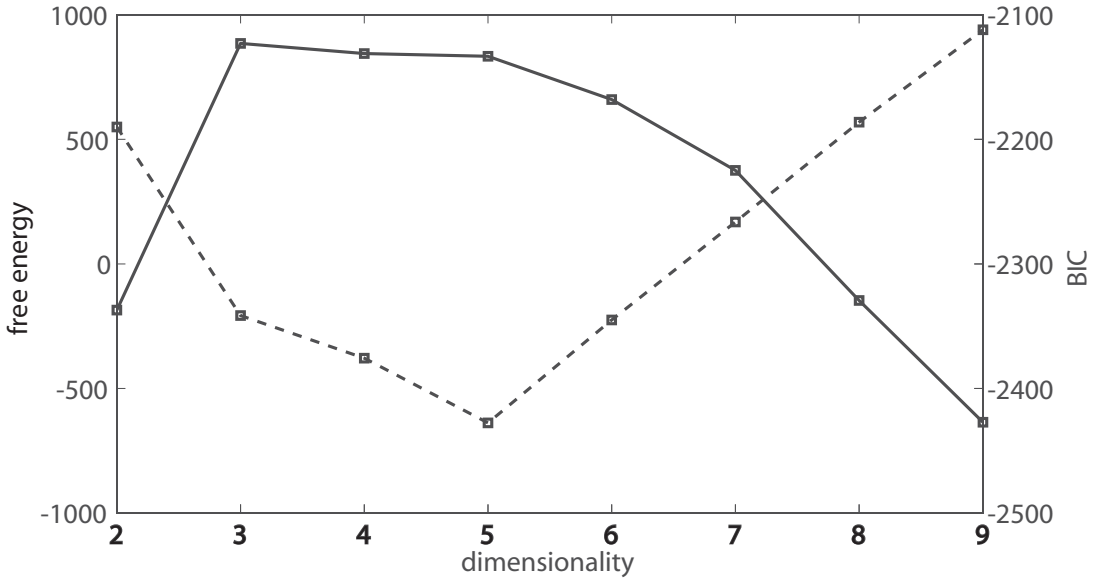

図 3.3 人工データに関する LDS と因子分析モデルにおけるモデル選択基準の比較．内部状態空間の次元 $N = 2, \ldots, 9$ に対する提案モデルでの自由エネルギー（実線）および因子分析モデルでの BIC （破線）を示した．

4.2 実験 2. 酵母遺伝子発現プロファイルに対する適用

出芽酵母遺伝子発現プロファイル

提案手法の実データに対する適用性を評価するため，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 [46] 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て得た対数発現比である．本データは，http://cellcycle-www.stanford.edu/ [1] から入手可能である．

本データは， 1) 包括的な遺伝子発現の観点から細胞内の現象を明らかにする目的で計測されたものである，2) 実験結果に基いた 800 個の遺伝子の機能分類情報が提供されているため，LDS モデルによる特徴抽出結果との対応付けが可能で

[1] Stanford 大学 Yeast Cell Cycle Analysis Project.

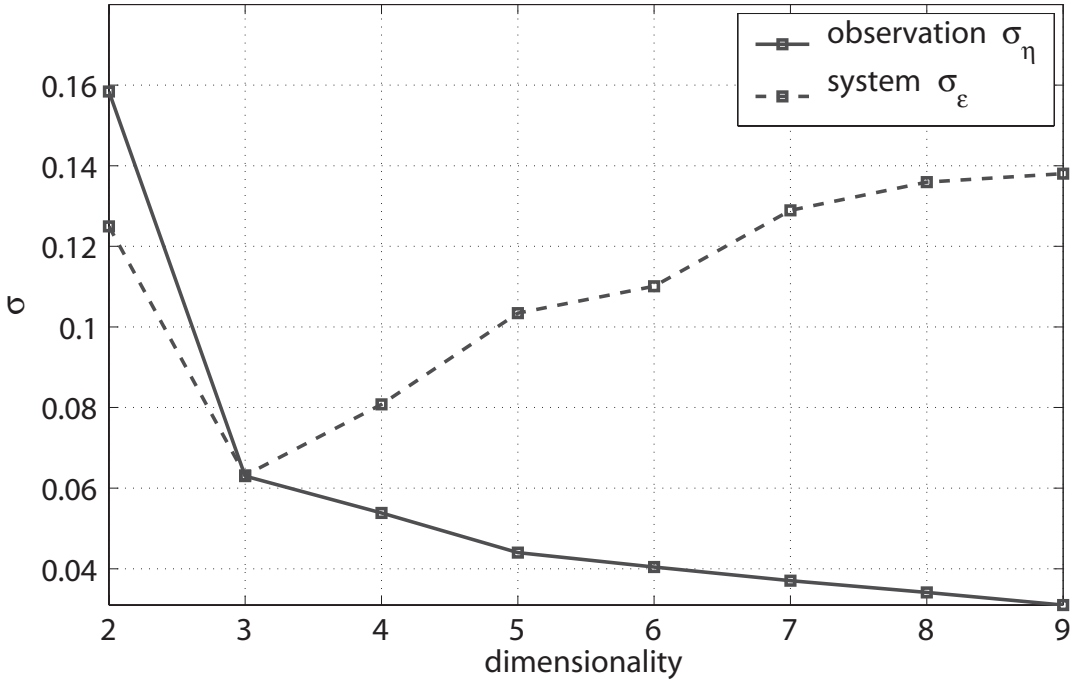

図 3.4 人工データに対するシステムノイズと観測ノイズの標準偏差の推定値．実線と破線はそれぞれ，システムノイズの標準偏差 σ_ϵ と観測ノイズの標準偏差 σ_η をである．

ある，3) 時系列データの形式をとり時点数も十分である，4) 従来手法の適用実験 [37] でも用いられた，といった性質を持ち， LDS モデルの評価に適していると考え，評価に用いることにした．

まず，前処理としてデータセットから，Spellman らによって同定された 800 個の遺伝子と，等間隔 (10 分間隔) で測定された 19 時点のサンプルのデータ （800 遺伝子 ×19 時点）を選択した．ついで，このデータに含まれる 1023 個 (5.3%) の欠測値について，Bayesian PCA アルゴリズム [25] により補完を行った．さらに，800 個の遺伝子から，ランダムに 200 遺伝子を抽出し学習データを構成した (図 3.5).

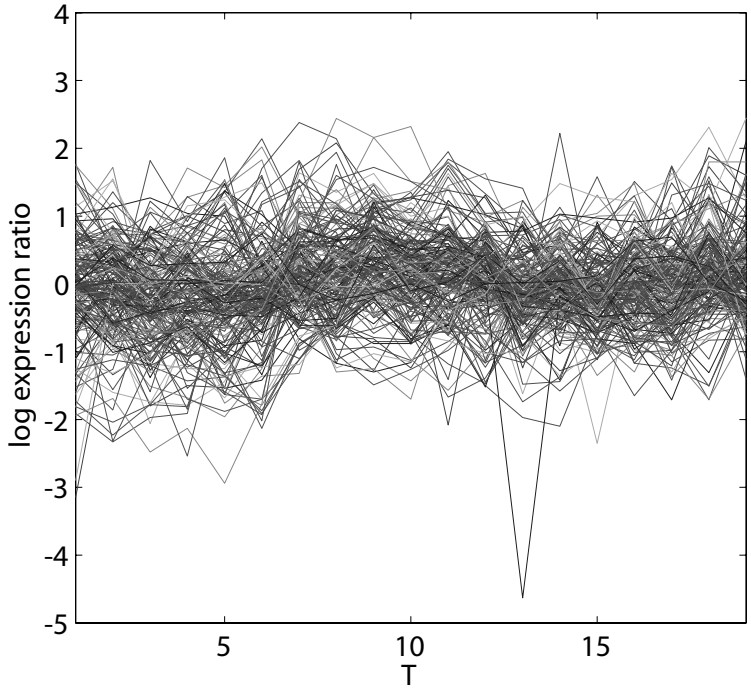

図 3.5 出芽酵母 200 遺伝子の 19 点の発現時系列データ．横軸は各測定点，縦軸は対数発現比を示す．

システム同定

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．事後分布は単峰に近いものと予想されるが，アルゴリズムが局所最適解に収束する可能性もあるため，推定の初期値を変えつつ 20 試行繰り返し，20 試行中で最大の自由エネルギーを実現したモデルを採用した．また，因子分析モデルに対する EM アルゴリズム（最尤推定）によるシステム同定を行い比較した．因子分析モデルでは解が一意に求まるため，各モデルに対し 1 試行のみ推定を行った．

図 3.6 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の最大値と，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，自由エネルギー最大化の観点から評価すると，LDS モデルでは最適な

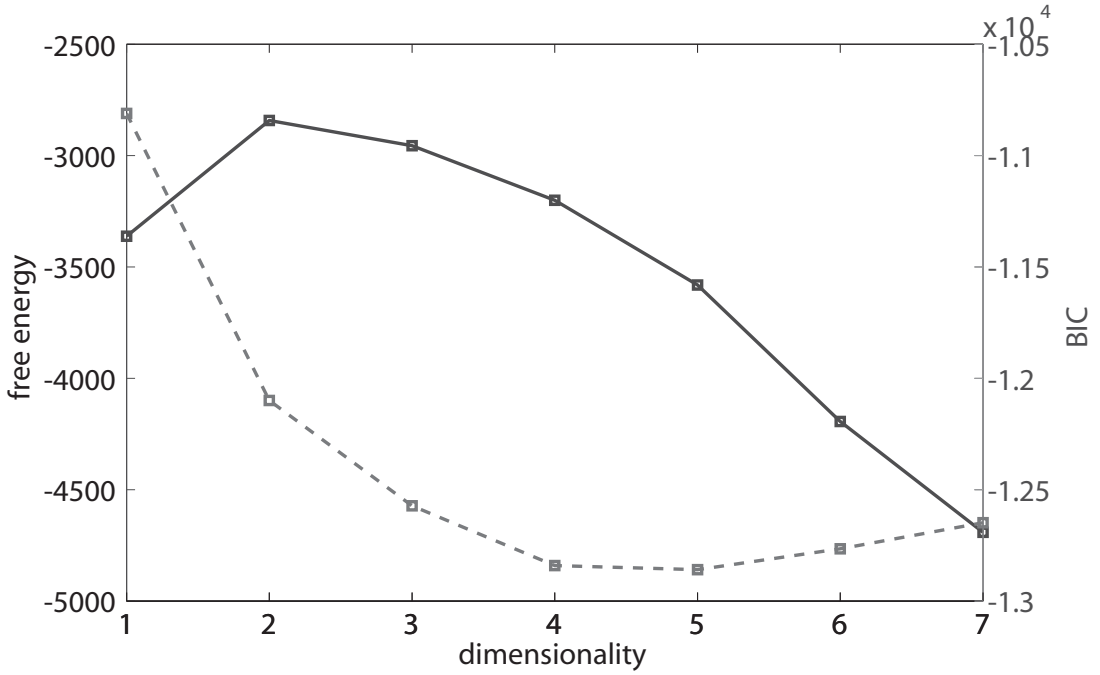

図 3.6 出芽酵母データに関する LDS と因子分析モデルにおけるモデル選択基準の比較．内部状態空間の次元 $N = 1, \ldots, 7$ に対する提案モデルでの自由エネルギー（実線）および因子分析モデルでの BIC （破線）を示した．

状態空間の次元数は $N = 2$ であるといえる．図には示していないが，状態変数の次元が 7 より大きいモデルでも，自由エネルギーは単調減少の傾向が見られた．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．

選択された $N = 1, \ldots, 7$ の LDS モデルに関する，システムノイズおよび観測ノイズの標準偏差の推定値を図 3.7 に示す．やはり $N = 2$ のモデルにおいてシステムノイズが最小となっていることが分かる．

図 3.8 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．

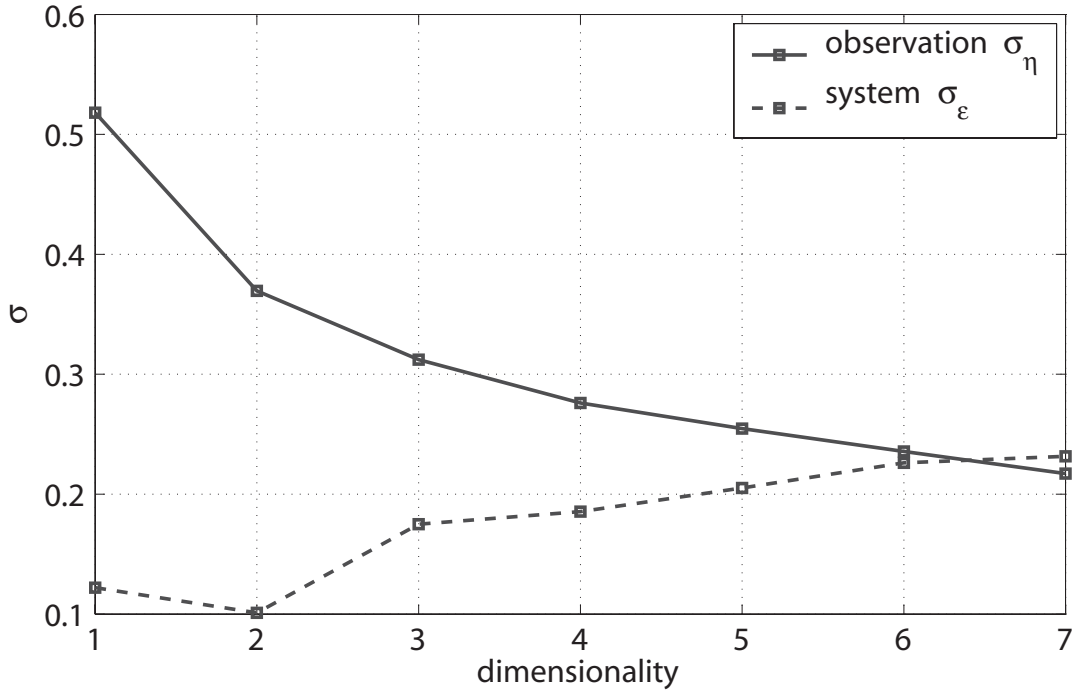

図 3.7 酵母データに関する LDS モデル ($N = 1, \ldots, 7$) のシステムノイズと観測ノイズの標準偏差の推定値．実線と破線はそれぞれ，システムノイズの標準偏差 σ_ϵ と観測ノイズの標準偏差 σ_η を示す．

このことは，図 3.7 において，システムノイズ分散が $N = 4$ や $N = 5$ で $N = 2$ よりも大きくなっているということに対応すると考えられる．あらゆるモデル中で自由エネルギーが最大となった $N = 2$ では，ちょうどフーリエ基底に対応するような，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．

観測行列から得られる生物学的知見

図 7 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．また，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

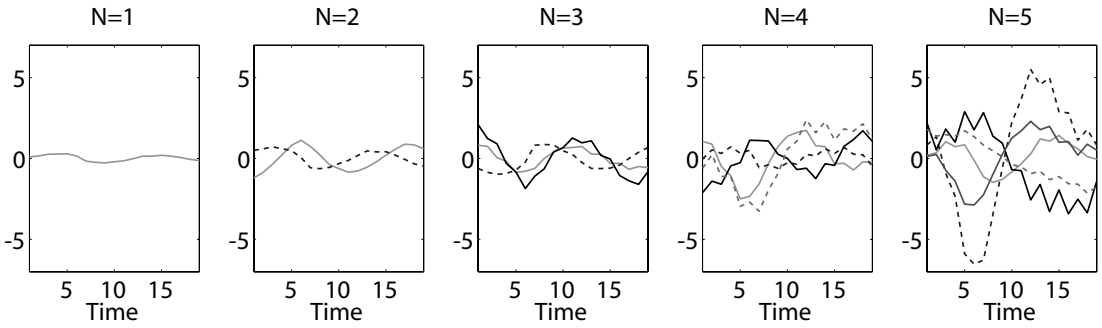

図 3.8 自由エネルギーが最大となったモデルでの内部状態変数 $\boldsymbol{x}$ の時系列．各列はモデルの内部状態変数の次元に対応する．

ズを示す．Spellman らは，細胞周期における機能が既知である 93 個の遺伝子の発現プロファイルを元に，細胞周期に関与すると考えられる 800 個の遺伝子を同定し，各遺伝子に対して，G_1/S, S, S/G_2, G_2/M, M/G_1 の 5 つに分割できる細胞周期のなかでいつ活性されるのかを，既知遺伝子との時系列の類似性に基づき分類を行った．図中のシンボルは，この分類結果に対応している．観測ベクトルの空間である 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．

5. 議論

体細胞分裂は，多細胞生物にみられるもっとも基本的な周期的かつ自律的な生理現象である．細胞が分裂し遺伝情報を複製した後，再び分裂するまでの過程を細胞周期というが，これは巨視的な観点から明確に 4 つの段階にわけることができる．分子レベルで見た場合も，各段階で発現する遺伝子は特異的であり，各遺伝子の発現量は細胞周期の中で常に動的に変動している．Spellman ら [46] は，細胞周期における遺伝子発現変化の周期性を仮定し，解析のための遺伝子発現のダイナミクスのモデルとして，周期変動する二種類の基底，サイン波とコサイン波の線形和すなわち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，我々はLDSモデルと変分ベイズ推定を組み合わせることにより，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）を自動的に構成した．

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を，我々が用いたものと同様のデータへ適用した結果では，我々の結論とは異なる状態空間次元$N = 5$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，彼らのモデルが，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．

我々のLDSモデルは，状態変数のダイナミクスとシステムノイズと観測ノイズを組み込んだ確率モデルとなっているため，データのノイズに対して比較的ロバストに，時間変化する状態変数の基底を抽出できると考えられる．マイクロアレイ実験などから得られたノイズのある時系列データを解析するためには，この性質は大きな利点としてはたらくと思われる．

6. 結論と今後の予定

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，

1. 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
2. システムに定常性を要求すること
3. 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

3 つが主に考えられる．特に第 3 の簡略化は，生物の環境への適応性を議論するためには問題であり，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，細胞の機能と発現制御モデルを結び付け，生物学的事実を明らかにするための，より具体的な系への適用を行いたい．現時点では，モデルの適用が，発現プロファイルの特徴抽出に留まっている．本手法を，機能は明確であるがその仕組みが明らかにされていないような細胞の活動における発現プロファイルに適用して，新たな事実を発見し，解析手法としての有効性も同時に示すことが必要である．そのために，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．

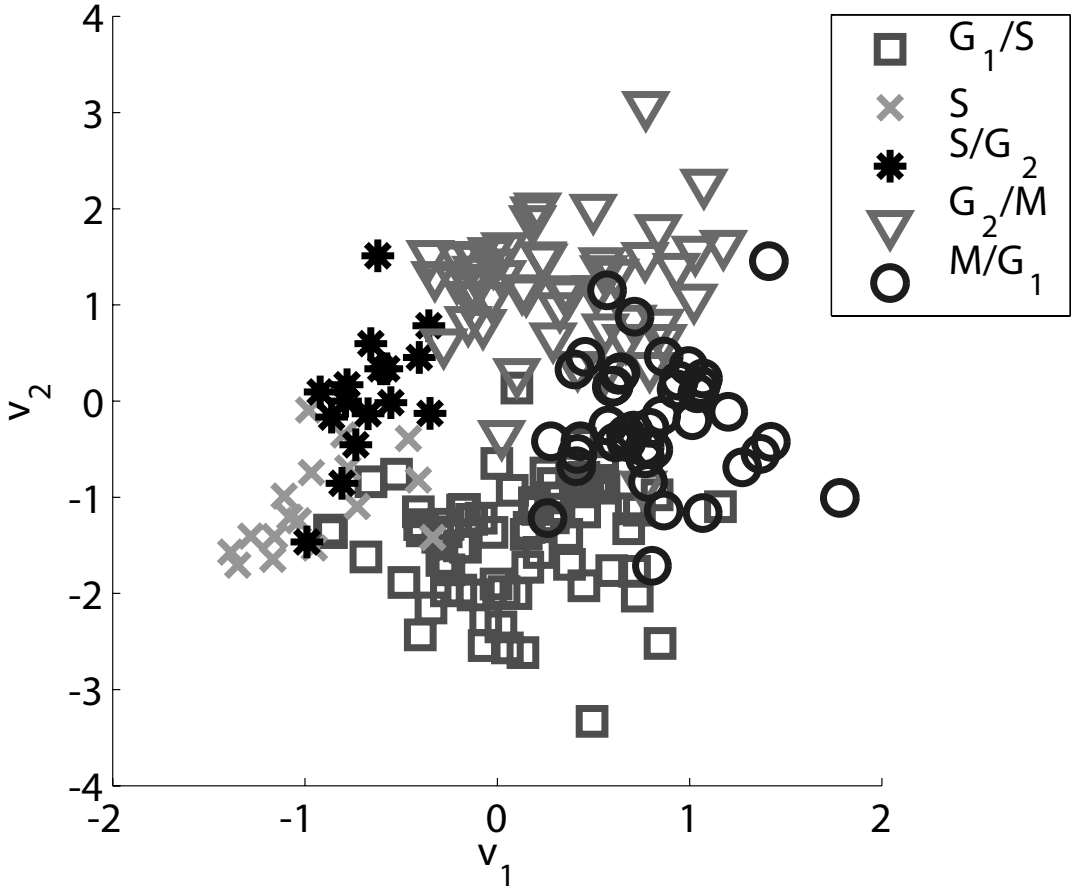

図 3.9 自由エネルギー最大の $N = 2$ の LDS モデルの推定結果から得られた $\boldsymbol{V}$ の横ベクトルの散布図．各シンボルは，Spellman らによって同定された，体細胞分裂の過程において遺伝子が高いレベルで発現するフェーズを示す．

第4章　二値分類器の最適な組み合わせによる遺伝子発現プロファイルからの癌サブクラス識別法

1. 序論

1.1 遺伝子発現プロファイルを用いた腫瘍分類

ゲノムワイドな遺伝子発現情報の新しく重要な利用法の一つとなっているのが癌の病理診断である．数千から数万の遺伝子発現情報は，組織が示す特異的な表現型と関わりを持つ大量の分子生物学的マーカーとして利用することができる．こうした遺伝子発現プロファイルを利用したアプローチは，従来の組織病理学的な癌診断法で問題となっていた，病理組織学者ごとの診断結果の揺れや，形態的な類似性により悪性組織と正常組織の鑑別が困難な場合などへの解決法の一つとして期待されている．近年，高次元な遺伝子発現データを用いて診断法を構築するための手段として，遺伝子発現プロファイルを症例組織サンプル固有のパターンベクトルとみなし，教師あり学習アルゴリズムを適用するいくつかの研究報告なされている．代表的なものを挙げると，2 種類の急性白血病の weighted voting アルゴリズムによる識別 [17] ，4 種類の small round blue cell tumors (SRBCTs) の人工ニューラルネットワークによる分類 [55]，そして，multi-class support vector machine (SVM) による 14 種類の成人悪性腫瘍の診断 [56] がある．これらの研究は，異なる組織に由来する腫瘍に関しては，分類アルゴリズムを用いてよく識別

することができるという事実を明らかにした．この理由は主として，ある組織の遺伝子発現プロファイルが，別の種の組織のそれとは大きく異なることによる．一方で，例えば家族性乳癌 [57] など，同一組織を起源とした複数の種類の腫瘍の識別は，異なる癌の表現型間に遺伝子発現パターンの類似性があるために難しい問題となっており，決定的な手法はいまだ存在しない．今後，より増加するであろう遺伝子発現情報を用いて有効に組織病理学的な診断を行うためには，こうした困難な状況に対処しなければならず，より高精度な多クラスパターン分類法が求められている．本研究では，こうした多クラス識別問題に対する新たな教師あり学習法を提案する．

1.2 腫瘍分類問題の定式化

発現プロファイルから腫瘍クラスを識別するパターン識別問題は，以下の様に定式化される [58]．遺伝子数 p，サンプル数 n の遺伝子発現データは，$n \times p$ 行列 $\boldsymbol{X} = (x_{ij})$ と表すことができる．ここで，x_{ij} は，サンプル i における j 番遺伝子の発現量を表している．サンプル i がある腫瘍クラスに属していると分かる場合，サンプルの遺伝子発現プロファイル $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ と，クラスラベル y_i が同時に観測される．腫瘍クラス数が K 個の場合，クラスラベル y_i のとりうる値は，1 から K までの整数となる．ここで，クラス k に属するサンプルの数を n_k とする．K 個の腫瘍クラスに対する識別器は，遺伝子発現プロファイルの空間 $\mathcal{X}$ を，K 個の不連続な空間 $A_1, \ldots, A_K$ に分割する．$\boldsymbol{x} = (x_1, \ldots, x_p) \in A_k$ ならば，$\boldsymbol{x}$ はクラス k に属していると予測される．

識別器は，これまでに観測された，クラスが既知であるサンプルに基づいて構成される．識別器を構成するプロセスを学習と呼ぶ．学習に用いるサンプルを学習データと呼ぶ．n_L 個の学習データ $\mathcal{L} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_{n_L}, y_{n_L})\}$ から構成された識別器による，サンプル $\boldsymbol{x}$ に対する予測ラベルを $C(\boldsymbol{x}, \mathcal{L})$ と表す．

学習データに含まれず，かつ腫瘍ラベルが既知であるサンプル $\tilde{\boldsymbol{x}}$ を考える．真の腫瘍ラベル ($\tilde{y}$ と表す) と識別器により予測されたラベル $C(\boldsymbol{x}_i, \mathcal{L})$ を比較することで，識別器の誤り率を推定することができる．識別器には多種多様なものが考えられるが，以上のようにして誤り率を評価することによって，サンプルの識別

に有効な識別器を選ぶことができる．

1.3 教師あり多クラスパターン識別法

機械学習の分野においてこれまでに研究されてきたパターン分類器には，2 値分類問題と多クラス分類問題の区別なく同様に使うことが可能であるものと，基本的には二値分類器専用として開発されたものとがある．前者としては K 最近傍法 [59]，shrunken centroids algorithm [60]，各クラスに属する入力の分布を多次元正規分布などのパラメトリック確率モデルで表現する方法 [61] などがあり，中でも Naive Bayes 法 [17] は簡単であり多用されている．後者は margin classifiers と呼ばれ，SVM [62][63], AdaBoost [64] が代表例である．これらは，2 つのクラス間の決定境界のマージンを制御することによってクラス分類の汎化性能を高めるアプローチであるため，多クラス分類問題に応用するためには，目的関数を多値に拡張する方法 [65][66]，あるいは，多値分類問題を複数の 2 値分類問題に分割して，後に統合するというヒューリスティクスが用いられてきた．特に，多クラス分類に拡張された SVM を multi-class SVM (MC-SVM) と呼ぶ．後者で通常用いられるのは，各クラスについて one-versus-the-rest (1R) の 2 クラス問題を解く分類器を M 個作り，M 個で投票をして推定クラスを決定する方法 [62], one-versus-one (11) を全て網羅したうえで，投票をして推定クラスを決定する方法である [67][68]．また 1R と 11 との両方を用いて結果を示している研究もある [56][69]．これらは理論的な根拠のないヒューリスティクスながら，SVM のような二値分類器自体が強力に働く問題であれば前者のどれよりも高い性能を示すことがある．しかし，11 と 1R のどちらの組合せがどういう場合に良いか，など具体的な組合せ方については不明確なままだった．例えば [70] は，公開遺伝子発現データセットをプラットフォームとして，MC-SVM として 1R， 11 の単純投票型 (引き分けはランダム選択) 及び，エラー訂正符号方式 (ECOC) [71] のバリエーションである random coding [72]，exhaustive coding [71] を比較し，加えて，Naive Bayes 法， K 最近傍法，決定木 J4.8 を比較した．彼らが示した結果によれば，ほとんどの場合で MC-SVM が最高の性能を示したが，様々な二値分類器の集合の作り方のうちでどれが良いかは問題依存であり，常に最高性能を示すアル

ゴリズムは無かった．また，同様の研究として [73] では，各種公開遺伝子発現量データセットに対して，各種 MC-SVM およびそれ以外の多クラス識別法を網羅的に適用することで性能評価を行い，その結果では，Weston and Watkins の方法 [65]，Crammer and Singer の方法 [66]，および，1R による単純投票を用いた MC-SVM が多くの場合で優れた性能を示したが，どの MC-SVM アルゴリズムが良いかは問題依存であった．

本研究では，3 値以上の教師付き分類問題に対して 11， 1R のみならず，任意の 2 値分類器 (SVM に限らない) の出力を組み合わせた場合にも適切な最終決定を得るための統計的枠組みを考える．まず，与えられた 2 値分類結果と真の 2 値分類結果との適合性を向上するという枠組みで，真の多値分類結果に対する事後確率最大化推定を行う MAP 法を提案する．これは，上記の [67] の拡張であり，[74] と等価である．さらに，この拡張として，任意の 2 値分類器に対して重みを持たせ，かつそれを統計的推定の枠組みでデータから決定する WMAP 法を提案する．これによって，複数の発現量データからの症例分類問題において，多値分類の正解率を向上できることを示す．

2. 統計的推定による 2 値分類器の組合わせ

2.1 2 値分類器の組合わせの確率モデル

D 次元のデータ点 $\boldsymbol{x} \in \mathcal{R}^D$ がクラスラベル $i \in C$ を持つものとする．ただし $|C| = M$ である．ラベルを指標する M 項変数 $\boldsymbol{t}$ を以下のように用意する．

$$
\begin{aligned}
\boldsymbol{t} &= \{t_i\}_{i \in C} \\
t_i &= \begin{cases} 1 & \text{if } \boldsymbol{x} \text{ belongs to class } i \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

ラベル集合 C の重複しない任意の 2 つの部分集合について，それらを分ける 2 値分類器を考えることができる．すなわち，クラスラベルのべき集合 $2^C = \{\{1\}, \{2\}, \ldots, \{1,2\}, \ldots, C, \emptyset\}$ から，意味のない部分集合 $C, \emptyset$ を除いたべき集合 $\tilde{2}^C \equiv 2^C - \{C, \emptyset\}$ を考え，$\tilde{2}^C$ から重複のないクラスラベルの部分集合 $l, m \in \tilde{2}^C, l \cap m = \emptyset$

を選択することで，クラス集合 l 対クラス集合 m の 2 値分類問題を定義することができる．この時，$[l|m]$ をターゲットと呼ぶ．全ての可能なターゲットの集合を B^{AA} とする．

$$\begin{aligned}[l|m] \quad \in B^{AA} \equiv \{&[1|2],[1|3],\ldots,[1|M],\ldots,[M-1|M],\\ &[12|3],[12|4],\ldots,[12|M],\cdots,\\ &[1\ldots M-1|M],\cdots\}.\end{aligned}$$

またその部分集合として，B^{11} すなわち $\#l = \#m = 1$ であるような $[l|m]$ の集合 (one versus one)，あるいは B^{1R} すなわち $\#l = 1, \#m = M-1$ であるような $[l|m]$ の集合 (one versus the rest) も考えられる．

$$\begin{aligned}B^{11} &\equiv \{[l,m] \,|\, l,m \in 2^C, \#l = 1, \#m = 1, l \cap m = \phi\},\\ B^{1R} &\equiv \bigcup_{j=1}^{M/2}\left(\bigcup_{i=1}^{m-1} B_{ji}\right).\end{aligned}$$

ここで $\#l$ および $\#m$ はクラス集合 l および m に含まれるクラス数を表す．

ここで考える問題は学習データセット $\boldsymbol{L} \equiv \{(\boldsymbol{x}^{(n)}, \boldsymbol{t}^{(n)})|n = 1:N\}$ について，ある任意のターゲット集合 B に属する 2 値分類器が与えられた状況で，新しいデータ点 $\boldsymbol{x}$ の所属クラスを求めることである．データ点 $\boldsymbol{x}$ が各クラスに確率的に所属すると仮定し，$\boldsymbol{x}$ のクラス $i \in C$ への所属確率のベクトル $\boldsymbol{p}(\boldsymbol{x})$ を以下で定義する．

$$\begin{aligned}&\boldsymbol{p}(\boldsymbol{x}) \equiv \{p_i(\boldsymbol{x})\}_{i\in C}\\ &p_i(\boldsymbol{x}) \geq 0, \quad \sum_{i\in C} p_i(\boldsymbol{x}) = 1.\end{aligned} \tag{4.1}$$

これを用いると，$\boldsymbol{x}$ のクラス部分集合 $l \in \tilde{2}^C$ への所属確率 $p_l(\boldsymbol{x})$ は $p_l(\boldsymbol{x}) = \sum_{i\in l} p_i(\boldsymbol{x})$ で表される．

ターゲット $[l|m]$ について，学習データセット $\boldsymbol{L}$ から得られる判別関数 $f^{\boldsymbol{L}}_{[l|m]}(\boldsymbol{x}) \in \mathcal{R}$ を考える．ここで，学習アルゴリズムは何でも良く，例えば SVM であるとしておく．このとき，判別関数値 $f^{\boldsymbol{L}}_{[l|m]}(\boldsymbol{x})$ を用いて，ターゲット $[l|m]$ に関するクラス部分集合 l への所属確率を $q_{[l|m]}(\boldsymbol{x}) \equiv Pr\left(c(\boldsymbol{t}) \in l | f^{\boldsymbol{L}}_{[l|m]}(\boldsymbol{x}), c(\boldsymbol{t}) \in l \cup m\right)$ で表す．$c(\boldsymbol{t})$ は $\boldsymbol{t}$ の指標するクラスである．また，$q_{[l|m]}(\boldsymbol{x})$ は，ラベルに関して以下の対称性を

満たすものとする．

$$q_{[l|m]}(\boldsymbol{x}) = 1 - q_{[m|l]}(\boldsymbol{x}).$$

本研究では，判別関数 $f^{\boldsymbol{L}}_{[l|m]}(\boldsymbol{x})$ から $q_{[l|m]}(\boldsymbol{x})$ への変換には，logistic regressor [75][76] に基づく方法と用いた．この詳細は付録 B に示した．ターゲット集合 $\boldsymbol{B}$ に関するクラス所属確率をまとめて，$\boldsymbol{q}(\boldsymbol{x}) \equiv \{q_{[l|m]}(\boldsymbol{x})\}_{[l|m]\in B}$ で表す．$\{\boldsymbol{q}(\boldsymbol{x}^{(n)})\}_{n=1:N}$ はデータセット $\boldsymbol{L}$，2 値分類学習器，判別関数値からクラス所属確率への変換法の 3 つから決まるものであり，以下ではデータとして見なしている点に注意する．

真の所属確率ベクトル $\boldsymbol{p}(\boldsymbol{x})$ が与えられたとき，ターゲット $[l|m]$ の各ラベル l, m への $\boldsymbol{x}$ の真の所属確率 $\pi_{[l|m]}(\boldsymbol{x})$ が以下で与えられるとする．

$$\pi_{[l|m]}(\boldsymbol{x}) \quad = \quad \frac{p_l(\boldsymbol{x})}{p_l(\boldsymbol{x}) + p_m(\boldsymbol{x})}.$$

本節では，単一のデータ点 $\boldsymbol{x}$ の所属確率ベクトル $\boldsymbol{p}(\boldsymbol{x})$ の推定問題を考えるものとして，以下では表記の簡単化のため $(\boldsymbol{x})$ を省略する．用いることのできるデータ $\boldsymbol{q}$ から $\boldsymbol{p}$ を求めたいが，両者は次元が違うため，$\boldsymbol{q}$ と同じ次元の変数 $\boldsymbol{\pi} \equiv \{\pi_{[l|m]}\}_{[l|m]\in B}$ を用意し，$\boldsymbol{q}, \boldsymbol{\pi}$ 間の以下の Kullback-Leibler (KL) ダイバージェンスを最小化するように，$\boldsymbol{p}$ を求める．

$$KL(\boldsymbol{q}; \boldsymbol{\pi}(\boldsymbol{p})) = \sum_{[l|m]\in B} \left\{ q_{[l|m]} \log \frac{q_{[l|m]}}{\pi_{[l|m]}} + (1 - q_{[l|m]}) \log \frac{1 - q_{[l|m]}}{1 - \pi_{[l|m]}} \right\}. \tag{4.2}$$

$\boldsymbol{p}$ の自然な分布は多項分布であるので，推定の正則化のために Dirichlet 事前分布を導入して，以下の目的関数の最大化問題として定式化する．

$$\begin{aligned} V(\boldsymbol{p}) \quad &= \quad \sum_{[l|m]\in B} \{q_{[l|m]} \log p_l + q_{[m|l]} \log p_m - \log(p_l + p_m)\} + \sum_{i\in C} \gamma_0 \log p_i + R \\ &= \quad \sum_{k\in B} a_k \log p_k + \sum_{i\in C} \gamma_0 \log p_i + R. \end{aligned} \tag{4.3}$$

ここで，γ_0 は Dirichlet 事前分布の強さを表すハイパーパラメータ，R は $\boldsymbol{q}$ のみに依存する定数である．また，a_k は次式で定義される．

$$a_k \equiv \sum_{[l|m]\in B, k=l} q_{[l|m]} + \sum_{[l|m]\in B, k=m} (1 - q_{[l|m]}) - \sum_{[l|m]\in B, k=l\cup m} 1. \tag{4.4}$$

目的関数 $V(\boldsymbol{p})$ を，式 (4.1) の制約の下で，$\boldsymbol{p}$ について最大化することにより，クラス所属確率の推定値 $\hat{\boldsymbol{p}}$ を得ることができる．これは，ラグランジュの未定係数法を用いた勾配法により実現できる．このように，任意のターゲット集合 B について，2 値分類器の判別関数値とそれからの所属確率を用いて，クラス所属確率を推定することができる．以下では，この手法を MAP (maximum *a posteriori*) 法と呼ぶ．

2.2 2 値分類器の重みの推定

前節では，単一のデータ点 $\boldsymbol{x}$ のクラス所属確率 $\boldsymbol{p}(\boldsymbol{x})$ の推定法として，MAP 法を導出した．しかし，ターゲット集合 B には，2 値分類器が学習しやすいものもしにくいものもあるにも関わらず，MAP 法では，その効果を無視して全ての分類器に一定の信頼度をおいていることになる．一方，信頼度を確率的に表現し，それを 2 値分類器の *a priori* な選択割合として導入する手法もあり得るが，信頼度の表現法に恣意性が残る．そこで，学習データセット $\boldsymbol{L}$ 全体に対する判別結果に基づき，適切な重み付けを行うための統計的方法を提案する．

ターゲット $[l|m] \in B$ に対する 2 値分類器について，信頼性に基づく選択確率 $w_{[l|m]} \geq 0$ を導入し，ある目的関数 (後述するが，学習データセット全体に対する判別ロスに関わるもの) を最大化するものとして，その選択確率を最適化する．すなわち，変数はターゲット集合 B に対する全ての重み $\boldsymbol{w} \equiv \{w_{[l|m]}\}_{[l|m]\in B}$ である．この重みは選択確率であるので，

$$w_{[l|m]} \geq 0, \sum_{[l|m]\in B} w_{[l|m]} = 1 \tag{4.5}$$

とする．この時，前節の KL ダイバージェンス (式 (4.2)) は以下のような重みつき KL ダイバージェンスとなる．

$$KL(\boldsymbol{q};\boldsymbol{\pi}(\boldsymbol{p})) = \sum_{[l|m]\in B} w_{[l|m]} \left\{ q_{[l|m]} \log \frac{q_{[l|m]}}{\pi_{[l|m]}} + (1-q_{[l|m]}) \log \frac{1-q_{[l|m]}}{1-\pi_{[l|m]}} \right\}. \tag{4.6}$$

これに対応して，式 (4.4) は，

$$a_k \equiv \sum_{[l|m]\in B, k=l} w_{[l|m]} q_{[l|m]} + \sum_{[l|m]\in B, k=m} w_{[l|m]} (1-q_{[l|m]}) - \sum_{[l|m]\in B, k=l\cup m} w_{[l|m]} \tag{4.7}$$

に変更される．データセット $\boldsymbol{L}$ 全体についてこれを行う必要があるので，目的関数

$$V(\{\boldsymbol{p}^{(n)}\}|\boldsymbol{w}) = \sum_{i=1}^{N}\sum_{k\in B} a_k^{(n)} \log p_k^{(n)} + \sum_{i=1}^{N}\sum_{i\in C} \gamma_0 \log p_i^{(n)} + R \tag{4.8}$$

を $\{\boldsymbol{p}^{(n)}\}$ について最適化することになる．ただし，$\{\boldsymbol{p}^{(n)}\}$ の各要素は互いに独立に最適化できるので，前節のアルゴリズムを N 回繰り返すことで推定可能である．また，R は $\{\boldsymbol{q}^{(n)}\}_{n=1:N}$ に依存した定数項である．

$\boldsymbol{w}$ は，クラス所属確率 $\boldsymbol{p}$ による判別の能力について，データセット $\boldsymbol{L}$ 全体について最適化するものとして決める．そのための効用関数 U を，クラス所属確率 $\boldsymbol{p}$ を用いた推定判別結果と真のクラスラベル $\boldsymbol{t}$ との一致度として定義する．

$$U \equiv U(\{\boldsymbol{p}^{(n)}\}, \{\boldsymbol{t}^{(n)}\}) = \sum_{n=1}^{N}\sum_{i\in C} t_i^{(n)} \mathrm{mx}(p_i^{(n)}). \tag{4.9}$$

ここで，$\mathrm{mx}(\boldsymbol{p}_i)$ は逆温度パラメータ β を持つ soft-max 関数である．

$$\mathrm{mx}(p_i) = \frac{\exp(\beta p_i)}{Z}, \quad Z = \sum_{i'\in C} \exp(\beta p_{i'}),$$

$\beta \to +\infty$ のとき，$\mathrm{mx}(p_i)$ は $\arg\max_i p_i$ のみ 1 とするものである．β はクラス所属確率 $\boldsymbol{p}$ を用いたクラス推定に対するノイズの大きさを制御するものであり，0 より十分に大きいものとして適当に設定する．

$\boldsymbol{w}$ の変化が $\boldsymbol{p}$ の最適化すべき目的関数 V を変えることに注意すると，ここで考える学習は，学習データセット $\boldsymbol{L} \equiv \{\boldsymbol{q}^{(n)}, \boldsymbol{t}^{(n)}\}_{n=1:N}$ について，

$$\tilde{\boldsymbol{w}} = \arg\max_{\boldsymbol{w}} U(\{\tilde{\boldsymbol{p}}(\boldsymbol{w})^{(n)}\}, \{\boldsymbol{t}^{(n)}\}) \quad \text{under the condition (4.1)} \tag{4.10}$$

$$\{\tilde{\boldsymbol{p}}^{(n)}\} = \arg\max_{\{\boldsymbol{p}^{(n)}\}} V(\{\boldsymbol{p}^{(n)}\}|\boldsymbol{w}) \quad \text{under the condition (4.5)} \tag{4.11}$$

を満たす $\tilde{\boldsymbol{w}}$ を推定することである．データセット $\boldsymbol{L}$ 全体に対して最適化された $\tilde{\boldsymbol{w}}$ を求め，それを用いて，新しいデータ点 $\boldsymbol{x}$ のクラス所属確率の推定値 $\boldsymbol{p}(\boldsymbol{x})$ を得ることで，適切な M 値判別を行うことができる．式 (4.10)(4.11) で定義される最適化の具体的な計算は，付録 C に示した．この手法を重み付き MAP (WMAP) 法と呼ぶことにする．

表 4.1 多クラス識別法の構成

	MAP-1R	MAP-11	MAP-AA	WMAP-1R	WMAP-11	WMAP-AA
B^+	B^{1R}	B^{11}	B^{AA}	B^{1R}	B^{11}	B^{AA}
重み学習	なし	なし	なし	あり	あり	あり

3. 実験と結果

3.1 実験 1. 人工データへの適用

まず，人工データセットへの適用を行い，我々の提案手法である MAP および WMAP のデモンストレーションを行う．多クラス識別法を構成するために，多クラス識別アルゴリズムとして MAP および WMAP の二種類，また，そこで用いるターゲット集合として B^{1R}, B^{11}, B^{AA} の三種類を用いる．これらの組み合わせから，合計 6 種類の多クラス識別法を準備しておく (表 4.1)．さらに，各多クラス識別法で用いる 2 値分類器は重み学習の効果を見るために単純な linear kernel SVM を用いた．SVM の実装には LIBSVM [77] を用いた．MAP 法における (ハイパー) パラメータは，あらかじめ $\gamma_0 = 2, \beta = 2000$ に設定した．

人工データセットとして 2 次元のデータ点に関する 3 クラス問題を想定し，以下の手順にしたがいデータの生成を行った．まず，各データ点 $\boldsymbol{x} = (x_1, x_2)$ を $[-2, 2]\times[-2, 2]$ の領域の 2 次元の一様乱数から生成した．次いで，3 つのクラス中心 $\boldsymbol{x}_{c_1} = (-\sqrt{2}, -\sqrt{2}), \boldsymbol{x}_{c_2} = (-\sqrt{2}, \sqrt{2}), \boldsymbol{x}_{c_1} = (\sqrt{2}, \sqrt{2}), \boldsymbol{x}_{c_3} = (\sqrt{2}, -\sqrt{2})\}$ との距離に基づき，$\arg\min_i \|\boldsymbol{x} - \boldsymbol{x}_{c_i}\|^2 - b_{c_i}$ を満たすものとして，クラスラベルを決定した．クラス c_1 は二つの中心を持つことに注意．ここで，$b_{c_1} = 2\log(0.35), b_{c_2} = 2\log(0.50), b_{c_3} = 2\log(0.20), b_{c_4} = 2\log(0.75)$ である．訓練データセットとして 400 点（c_1:200 点, c_2,c_3:100 点ずつ），テストデータセットとして 600 点（c_1:300 点, c_2,c_3:150 点ずつ）を生成した (図 左)．本データセットは明らかに線形判別が困難なターゲット [1|23] を含むため，ターゲット集合として B^{1R} を用いた場合に望ましくない結果となることが予想される．また，B^{1R} を包含している B^{AA} においても，適切な

表 4.2 人工データに対する適用結果．各数値は 5-fold cross-validation accuracy，括弧中の数値はその標準偏差を表す．

	MAP-1R	MAP-11	MAP-AA	WMAP-1R	WMAP-11	WMAP-AA
Training	0.5625 (0.0643)	0.8785 (0.0088)	0.8415 (0.0297)	0.8670 (0.0330)	0.8825 (0.0105)	**0.8925** (0.0173)
Test	0.5567 (0.0497)	0.8687 (0.0155)	0.8313 (0.0152)	0.8603 (0.0211)	0.8637 (0.0197)	**0.8783** (0.0130)

識別境界を得るためにはこのターゲットの信頼性を低く見積もる必要があると考えられる．

準備した 6 種類の識別法を人工データセットに適用し，訓練データセットおよびテストデータセットに対する正答率により性能評価を行った．データセットの生成は 5 回行い，そのおのおのに対して 6 種の識別法を適用して得られた正答率の平均と標準偏差を表 4.2 に示した．

全般的な性能に関しては，WMAP-AA により最良の正答率が得られていることが分かる．重み推定の効果は顕著であり，WMAP-11 と WMAP-AA はそれぞれ，重み推定なしの MAP-11，MAP-AA よりも高い正答率が得られている．WMAP-11 において，MAP-11 よりもテストデータの正答率が低下しているのは過学習によるものである可能性がある．ターゲットの違いが MAP 法の性能に与える影響については，データ依存であるといえる．ただし，MAP-1R は，有意に性能が悪い．これは，MAP-1R において一つのターゲットが性能に与える影響が大きいことに由来している．

ここで，WMAP-AA の適用例を取り上げ，重みの推定に関して評価する．この人工データセットは 線形判別器で分類不可能なターゲットを含むデータ構造となっているため，2 値分類器群の組み合わせにより最適な識別境界を得るためには，特定の解けない 2 値分類問題を無視すると同時に，正答率に貢献する 2 値分類器に関しては重きをおきながら統合することが必要となる．本実験の結果は，我々の WMAP-AA により，これが自動的に実現されていることを示す．図 4.1 は，人工データセットに対し WMAP-AA を適用した結果得られた識別境界と 2 値分類器の重みを表す．左図の実線はデータの散布図と WMAP-AA により得られた識別境界を示す．比較のため，MAP-AA による識別境界を点線で示した（実線より

やや原点側に位置する）．破線はベイズ最適な識別境界である．右図のマトリクスは行がターゲットのインデックス (全 25 ターゲット)，列がクラス (全 4 クラス) を表す．マトリクスの各行における色の濃淡で各ターゲットで対となる二値クラスを示す．白色はターゲットに含まれないクラスである．マトリクスの各行の右にある上下二本のバーは，上が各ターゲットの重み（最大値が 1 になるように正規化済み），下が各ターゲットの training accuracy を示す．本データではクラス c_1 が対角状に分布しているため，ターゲット [1|23]（図 2，4 行目に対応）に関する二値分類器の学習は困難であり，その判別性能はランダムな判別に近い．逆に，ターゲット [2|3]（図 2，3 行目に対応）は容易に判別が可能なクラス分布となっており，実際高い識別率が得られている．ところが，このターゲットに関する最適な識別境界では，c_1 を c_2 または c_3 にほぼランダムに分類することになるため，2 値分類性能が高いからといって信頼を寄せてしまうとトレードオフにより全体的な性能低下につながる．重みの推定結果は興味深く，ターゲット [1|23], [2|3] に関して 0 の値を与えている．[2|3] に関する分類は，他のターゲットにより間接に行っていることが考えられる．また，WMAP-AA で得られた識別境界と重み推定無しの MAP-AA のそれとを比較すると，WMAP-AA のものは，ロスを最小にする方向に適当な量だけマージンを広げた結果としてとらえることができる．

3.2 実験 2. 各種遺伝子発現量に基づく腫瘍分類問題への適用

実問題への適用として，4 つの遺伝子発現プロファイルデータに基く癌分類問題へ我々の手法の適用を行った．用いたデータセットは以下の 4 種類である．

Thyroid cancer データセット

甲状腺癌 168 サンプル，2,000 遺伝子に関して，ATAC-PCR [11] 法により計測されたオリジナルの遺伝子発現プロファイルである．現在，甲状腺癌の分類は主に微細針吸引生検によって行われているが，摘出の際に組織構造が崩壊しやすく，このことが鑑別診断を極めて困難にしている [78][79][80]．こうした事情により，

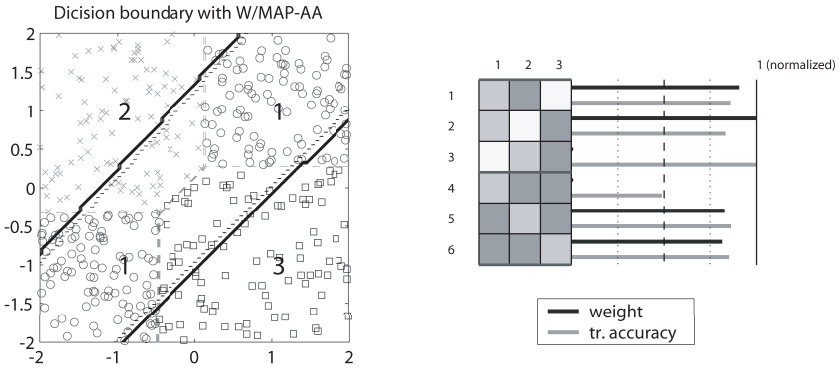

図 4.1 人工データセットに対する WMAP-AA の適用結果．左図の実線は WMAP-AA により得られた識別境界を示す．比較のため，MAP-AA による識別境界を点線で示した（実線よりやや原点側に位置する）．破線はベイズ最適な識別境界である．右図は，推定された重みと 2 値分類器の訓練データに対する正答率をターゲット毎に表したものである．マス目の行はターゲット，列はクラスに対応し，各ターゲットにおける 2 値分類の対象クラスを色の濃淡で表現している．白色のマス目は未使用クラスである．各行の 2 本のバーは，上段が推定された重み，下段が 2 値分類器の訓練セットに対する正答率を表し，分かり易さのため，それぞれを最大値が 1 となるように正規化した．

遺伝子発現プロファイルからの診断が期待されてきた．本データセットでは 4 種類の病理組織の分類を目的としており，サンプル構成は，58 (follicular adenoma; FA), 28 (follicular carcinoma; FC), 40 (normal; N), 42 (papillary adenocarcinoma; PC) となっている．

Esophageal cancer データセット

本データセットは，日本人食道癌患者のサンプルから ATAC-PCR 法によって計測された遺伝子発現プロファイルであり [81][82]，甲状腺癌同様にオリジナルのデータセットである．日本における食道癌は扁平上皮癌がほとんどである一方で，米国やヨーロッパでは腺癌，いわゆる Barret tumor が多いことに注意されたい．本データセットでの目的は，低分化型扁平上皮癌，中分化型扁平上皮癌，お

表 4.3 4 種類の癌分類問題

データセット名	サンプル数	観測クラス数	遺伝子数
Thyroid cancer	168	4	2,000
Esophageal cancer	141	3	1,763
SRBCT	83	4	2,308
Leukemia	72	3	11,225

よび，高分化型扁平上皮癌の 3 つの組織病理学的タイプの分類である．サンプル構成は，順に 14, 97, 30 である．病理学者にとって，この鑑別診断は容易ではなく，病理学者により診断結果に食い違いの出ることが多い．

SRBCT データセット [55]

Small round blue cell tumors (SRBCTs) の 83 サンプル，2,308 遺伝子からなる遺伝子発現プロファイルであり，4 種のクラスラベルを含む．ラベルごとのサンプル構成は，29 (Ewing family of tumors; EWS), 25 (rhabdomyosarcoma; RMS), 11 (Burkitt lymphoma; BL), 18 (neuroblastoma; NB) である．

Leukemia データセット [83]

72 サンプル，11,225 遺伝子からなる急性白血病に関する遺伝子発現プロファイルであり，3 種類のクラスラベルを含む．ラベルごとのサンプル構成は，28 (acute myeloid leukemia; AML), 24 (acute lymphoblastic leukemia; ALL), 29 (mixed lineage leukemia; MLL) となっている．

以上のデータセットの情報を 4.3 にまとめた．

多クラス識別法は，実験 1 で用いたものと同じ 6 種類を準備しておく (表 4.1)．各多クラス識別法で用いる 2 値分類器として linear kernel を用いた SVM を準備した．Linear kernel SVM は，一般に，高次元低サンプル数という特徴を持つ遺伝子発現プロファイルを用いた癌分類において十分な性能が得られる [73]．MAP 法お

よび WMAP 法のハイパーパラメータは，thyroid cancer データセットと SRBCT データセットに関しては $\gamma = 2, \beta = 2000$，esophageal データセット と leukemia データセットに関しては $\gamma = 2, \beta = 1500$ にあらかじめ設定した．さらに，提案手法の性能を従来の手法と比較検討するために，shrunken centroid algorithm (SC) [60], MC-SVM として Weston and Watkins (WW) のアルゴリズム [65] および Crammer and Singer (CS) のアルゴリズム [66] の 2 種類，合計 3 つの先端の多クラス識別アルゴリズムを準備した．SC では，shrinkage parameter Δ を 0 から 0.25 刻みで 6 まで設定した 25 通りの識別器を準備し最適値を選ぶ．二種類の MC-SVM では，linear kernel を用いた．

それぞれのデータセットに対し，5-fold cross-validation (CV) により各データセットに対し各多クラス識別法を適用し，training accuracy と test accuracy を用いて評価を行った．その結果得られた各 accuracy の平均値と標準偏差を表 4.4 に示した．各数値が accuracy の平均値，括弧内の数値が標準偏差を表す．

提案した 6 種類の多クラス識別法の結果に関して比較すると，ターゲットセット B^{AA} が他のものよりも一貫して優れた性能をもたらしていることが分かる．Thyroid cancer と esophageal データセットでは最高の test CV accuracy を示し (それぞれ 0.7744, 0.7026)，また，SRBCT および leukemia データセットでは最高値と同等の性能を示している (それぞれ 1.0, 0.9692)．Esophageal を除く 3 つのデータセットでは，training CV accuracy が 6 種全ての多クラス識別法で上限の 1.0 に達している．SRBCT データセットは，test CV accuracy ですら 1.0 に達しているため，発現情報による分類が易しすぎるものであると考えられる．Thyroid と leukemia データセットでは，test CV accuracy が 1.0 に達していないことから，training CV accuracy が 1.0 となっているのは過学習であると考えられる．これら 3 種類のデータセットでは，重みパラメータの変化により training accuracy を改善させる余地がないため (こうした状況を飽和と呼ぶ)，WMAP 法の利点が活かせず，その結果，WMAP 法 の CV accuracy は MAP 法のものと全く同一となっている．一方，esophageal データセットでは，training CV accuracy が上限まで達しておらず飽和していないため，WMAP 法の training CV accuracy と test CV accuracy が MAP 法の性能と比べて改善されており，WMAP 法の有効性が分かる．ここで，ある CV

表 4.4 多クラス識別法の性能比較

	MAP-1R	MAP-11	MAP-AA	WMAP-1R	WMAP-11	WMAP-AA
Thyroid Cancer						
Training	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
Test	0.7619 (0.0649)	0.7624 (0.0723)	**0.7744** (0.0743)	0.7619 (0.0649)	0.7624 (0.0723)	**0.7744** (0.0743)
Esophageal cancer						
Training	0.8207 (0.1088)	0.9007 (0.0034)	0.9007 (0.0034)	0.9007 (0.0034)	0.9167 (0.0368)	0.9007 (0.0034)
Test	0.6954 (0.0259)	0.6883 (0.0759)	0.6957 (0.0497)	0.6954 (0.0259)	0.6883 (0.0759)	**0.7026** (0.0511)
SRBCT						
Training	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
Test	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	0.0000 (0.000)	1.0000 (0.0000)
Leukemia						
Training	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
Test	**0.9846** (0.0334)	0.9559 (0.0678)	0.9692 (0.0688)	**0.9846** (0.0688)	0.9559 (0.0678)	0.9692 (0.0688)

	SC	MC-SVM (WW)	MC-SVM (CS)
Thyroid Cancer	($\Delta = 0.50$)		
Training	0.8871 (0.0221)	1.0000 (0.0000)	1.0000 (0.0000)
Test	0.7437 (0.0345)	0.7682 (0.0686)	0.7619 (0.0679)
Esophageal cancer	($\Delta = 0.00$)		
Training	0.9115 (0.0246)	1.0000 (0.0000)	1.0000 (0.0000)
Test	0.6745 (0.0512)	0.6806 (0.0557)	0.6727 (0.0647)
SRBCT	($\Delta = 1.75$)		
Training	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
Test	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)
Leukemia	($\Delta = 0.00$)		
Training	0.9723 (0.0196)	1.0000 (0.0000)	1.0000 (0.0000)
Test	0.8903 (0.0572)	**0.9846** (0.0344)	0.9692 (0.0688)

のステップにおける WMAP-AA での重みの推定における効用関数の変化を図 4.2 に示す．図の横軸を勾配法のステップ数とし，効用関数を training set のサンプル数で割った値，training accuracy, および test accuracy をプロットした．飽和していない場合，効用関数の値の上昇し，それに追従して，training accuracy および test accuracy が上昇している様子が分かる．

既存の最新の多クラス識別法と比較した場合，提案した MAP, WMAP 法は同等かそれ以上の性能を持つことが示された．

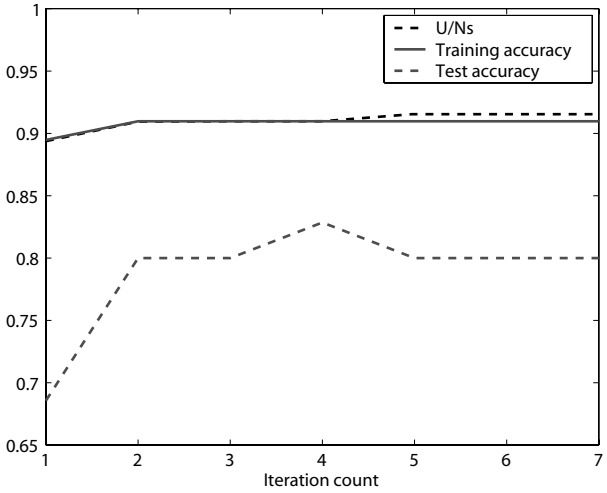

図 4.2 重み推定における効用関数の値の変化

3.3 実験 3. 識別に寄与する遺伝子が少ない分類問題に対する適用

実験 2 の結果が示す通り，遺伝子発現プロファイルからの分類が容易な構造を持つデータセットやサンプル数が少ない場合には training CV accuracy の飽和が起こる．こうした状況では，効用関数の二値分類器に対する重みによる調整ができないため，WMAP 法が本来持ちうる性能を評価することが出来ない．だが，以下の理由により，組織病理学的に意義のある分類困難な問題では飽和は生じないと考えられる．

- 遺伝子発現プロファイルにより容易に分類可能な癌サブクラスは，そもそも他の臨床的指標によっても容易に分類される傾向がある．たとえば，実験 2 でデータセットとして用いた SRBCTs は異なる組織に由来したものであり，遺伝子発現情報同様に顕微鏡観察を用いても鑑別診断が可能である．
- 診断が困難な問題では，医療生物学者はより多数のサンプルに関する遺伝子発現プロファイルを収集する．

本実験では，SRBCT データセットを元に分類困難な状況を人工的に作り出し，そこでの MAP/WMAP 法の性能評価を行った．

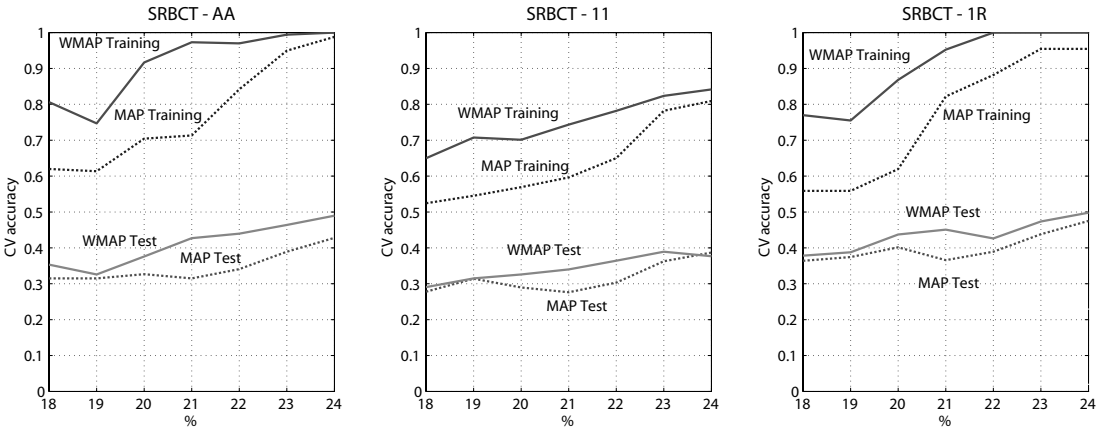

図 4.3 縮小データセットに対する MAP/WMAP 法の 5-fold cross-validation accuracy．各図の横軸は，相関比下位遺伝子数の割合 (%)，縦軸は，cross-validation accuracy である．

分類困難な問題を準備するため，SRBCT データセットから情報のある遺伝子を次の方法に従い削減した．まず，学習データセットに含まれる遺伝子発現量プロファイルとラベル情報を用いて，各遺伝子ごとに相関比 η^2 を計算する．

$$\eta^2 = \frac{\sum_{c \in C} n_c (\bar{x}_c - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}, \quad 0 \leq \eta^2 \leq 1.$$

ここで，x_j はサンプル j の発現量，$\bar{x}_c$ はクラス c に属するサンプルに関する発現量の平均値，$\bar{x}$ は全サンプルに関する発現量の平均値を示す．次いで，相関比の大きさで下位にランクされる，つまり，より情報の少ない遺伝子を r % 分だけ選択しデータセットを再構成する．このように再構成されたデータセットを縮小データセットと呼ぶことにする．縮小データセットに対し，MAP/WMAP 法とターゲットセット B^{1R}, B^{11} および B^{AA} の組み合わせによる 6 種類の多クラス構成法を適用し，5-fold CV により評価した．

結果を図 4.3 に示す．各図の横軸は学習に用いた相関比下位の遺伝子数の割合 (%) を示す．その範囲は training accuracy の飽和が生じない $18 \leq r \leq 24$ と設定した．ここで作成した縮小データセットには，オリジナルのデータセットと比較してより少ない情報を持つ遺伝子しか含まれていないが，それでも提案手法がチャ

表 4.5 縮小データセット ($r = 23\%$) に対する 5-fold cross-validation の適用結果. 各数値が cross-validation accuracy，また，括弧中の数値はその標準偏差を表す.

SRBCT reduced	AA	1R	11
Training			
MAP	0.9495 (0.0603)	0.9545 (0.1016)	0.7820 (0.1864)
WMAP	0.9941 (0.0132)	1.0000 (0.0000)	0.8236 (0.1309)
Test			
MAP	0.3893 (0.0954)	0.4380 (0.0960)	0.3627 (0.0636)
WMAP	0.4638 (0.1709)	0.4735 (0.0886)	0.3893 (0.0751)

ンスレベル以上の test CV accuracy を達成していることが分かる．また，これらの状況では，WMAP 法の性能が MAP 法のものよりも常に上回っている．表 4.5 は，$r = 23\%$ における trainig CV accuracy と test CV accuracy およびそれらの標準偏差を示す．数値が CV accuracy，括弧内の数値は CV accuracy の標準偏差である．

6 種類の多クラス識別法の中で，最良の test CV accuracy を示したのは WMAP-1R (0.4735)，次点は WMAP-AA (0.4638) であり，これらは同等と言える．MAP-AA の性能が，最悪の MAP-11 の 0.3627 に次ぐ 0.3893 であることを考えると，WMAP-AA は MAP-AA の重みパラメータを適切に更新し，顕著に性能を改善させることができていることが分かる．

B^{AA} には B^{11} と B^{1R} が含まれているため，最適な重みでの WMAP-AA の性能は，WMAP-11 と WMAP-1R のうちどちらかより優れたほうのものと少なくとも同等となりうる．ノイズや限られた数のサンプルなどデータによっては必ずしも最適な重みが獲得できるとは限らないが，今回の実験で WMAP-AA で重みが非常に良く推定できることが分かった．以上から，WMAP-AA を用いることで，一貫して最良の正答率が得られると期待できる．

4. 議論

4.1 教師あり学習アルゴリズムとしての MAP 法 と WMAP 法

本章で提案した MAP 法は，これまでの単純投票法で良く使われていたターゲット集合である B^{1R}, B^{11} だけでなく，任意のターゲット集合を組み合わせるための統計的手法である．単純投票法は，二値分類器の出力 $\boldsymbol{q}_{[l|m]}(x)$ を各クラスラベル集合 l, m について加算するアルゴリズムであるが，加算されるのは $l \in C$ または $m \in C$，つまり l または m がクラスラベルである場合についてのみであり，複数のクラスラベルを含む場合の二値分類器の出力は破棄され利用できない．例えば，B^{1R} では one-versus-the-rest の the-rest に関する投票は全て無視されることになる．一方，MAP 法では，KL divergence 最小化に基づき $q_{[l|m]}$ を各クラスラベルにうまく配分することができるために，どのようなクラスラベル集合に関する二値分類器の出力も利用可能である．過去の研究では，単純投票に基づく MC-SVM では B^{1R} が良い場合が多いことが示されていたが [56][73]，今回の実験結果では B^{AA} が多くの場合で他のターゲット集合と同等かそれ以上の正答率を示したことを踏まえると，B^{1R}, B^{11} では十分ではなく，その他の情報を利用することで改善の余地があることが示唆される．

MAP 法の統計的モデルは，B^{11} に含まれる二値分類確率の推定値をペアワイズに組み合わせる方法 [67] を，あらゆるラベルの組み合わせを用いることが可能なように拡張したモデルである．また，ECOC の観点から [67] の拡張を行った [74] とは数学的には等価である．[74] では，B^{1R}, B^{11} および B^{AA} からランダムに選択したターゲット集合に関する評価を行っているが，B^{AA} そのものは扱っていない．また，どのターゲットを組み合わせて多クラス分類法を構築すれば最適であるかに関しては未解決であった．WMAP 法での 2 値分類器出力に対する重みの導入とその推定法は，この問題に対する一つの解である．

WMAP 法では，全ターゲットを含む B^{AA} を用いた時にもっとも性能を発揮することが示された．B^{AA} を使わない場合には，どのターゲット集合が良いかは問題依存であるといえる．重みの学習では，基本的にどのターゲット集合を用いたとしても test accuracy に改善が見られるため，計算量との兼ね合いを考えて経験的

に決めることが良い．しかし，実験 3 の WMAP-1R の結果が示すように，B^{1R}, B^{11} など含まれるターゲットが少数である場合に，学習すべき重みのパラメータ数が少ないために過学習が起きる可能性がある．これは，学習データセットに多数のサンプルを用いることで回避できる可能性があるが，大量のサンプルを準備することが困難であることが多い遺伝子発現データの場合には工夫が必要となる．この問題に対処するための一つの手段として，leave-two-out (LTO) [84] などクロスバリデーションを階層的に行うアプローチが有効であると考えられる．

WMAP 法では，重み学習開始時点で training accuracy が上限の 1 に達している場合には，重みの更新がほとんど起こらないために WMAP の利点が生かせないことがある．これは，重みの学習に対して二値分類器の構成に用いたデータと同一のものを使っているためである．この傾向は，二値分類器が強力であればあるほど強くなると考えられる．この問題に対処する方法として，学習データを二値分類器構成用と重み学習用の 2 つに分割することで，重みの更新が可能となり，また，同時に過学習も防げると考えられる．ただし，この場合はサンプル数が多く必要となる．将来的に大量のサンプルが得られる状況において，本手法は真価を発揮すると考えられる．

4.2 2 値分類器の重みの組織病理学的な解釈

遺伝子発現解析の立場から見ると推定された重みは有用な情報を与えると考えられる．データからロスに基づき決定された 2 値分類器に対する重みは，多値分類問題が持つ複雑で階層的なベイズ決定境界を構成する際に，どのような 2 値分類器が部品として参加すべきかであるかの情報を表現している．この情報は，特に，高次元データの多値分類問題などにおいて，データの構造を幾何学的に理解する際に有用であると考えられる．実験 1 では人工データの解析により，クラスラベルの分布と推定された重みに関連が見られることを示したが，遺伝子発現解析においても，クラスラベルの生物学的意味を考える上でも知見を与えることが期待される．これは，識別器における特徴（遺伝子）選択と並び重要な情報であると考えられる．具体的な対応が得られるかについては今後の課題としたい．

5. 結論と今後の予定

本研究で提案した MAP 法と WMAP 法は，同一組織を起源とする類似した癌のサブタイプを鑑別するための方法として，新たな可能性を開いた．もちろん，臨床面での有効性をより深く検証するためには本研究で行った適用実験だけでは十分とは言えない．まず，第一の課題として，本手法を様々な癌分類問題やより多くのサンプルを含むデータへの適用を行い性能を検証ことが挙げられる．近年，NCBI の Gene Expression Omunibus (GEO) [85] をはじめとして，癌などの疾病を含むあらゆる遺伝子発現プロファイルを公開・閲覧するための環境が急速に整いつつあり，データの入手性が大幅に向上している．今後，これらの公開リソースを利用することで，性能評価を検証していきたい．

第二の課題は，多クラス識別法の改善である．関連研究として，Gharahmani らの Bayesian classifier combination [86] は，野心的な確率モデルを提案している．彼らは二値分類器に限らず，任意の多クラス識別器ユニットを前提として複数用意したうえで，それらが学習データに対して示した分類実績から，混同行列のベイズ推定を行い，これに基いて最適な最終出力を得る．単純な投票によるモデル出力の平均とは異なり，確率モデルに基くモデル混合を行っているという点で，彼等のモデルは我々のモデルと目的が似ており，重要な関連研究として挙げられる．彼らはさらに各学習データ点に対してそれの分類が (1) 困難である (0) 簡単である，という二値の隠れ変数を設定するモデル，ユニット間の相関モデルなどを加えて導入しており，いくつかの参考問題によっては効果を上げている．前者については，我々のモデルとも親和性の高い工夫であるので，このアイディアを含めた場合の性能評価も近い将来の課題のひとつとして挙げておく．

第5章　結言

本論文では，遺伝子発現解析における 3 つの問題に関する統計的解析手法について議論した．

第 2 章では，遺伝子発現解析の基礎となる RNA 測定法を対象として，ATAC-PCR 法の精度向上の方法について扱った．網羅的な蛍光ピークデータ解析によりアダプタ長とピーク値のバイアスの間に相関があることを突き止め，メディアン補正によるピーク値のバイアス補正法を提案した．詳細な物理化学的な特性が未知であっても，データの背後にあるモデルを考え，それに基く数理的なフィルターを当てはめることでデータの精度を高めることが出来ることを示した．また，3 種類の検量線の中で，精度に対応する RMSE，欠測率の二つの指標から最適なキャリブレーション法について検討を行った．メディアン補正データと最適なキャリブレーションにより，従来の方法よりも精度よくより多くのデータを取得することが出来ることが示された．

第 3 章では，遺伝子発現プロファイルの時系列，つまり高次元かつノイズを多く含む時系列データから，時間的なダイナミクスを含む特徴を抽出する手法を提案した．従来の因子分析モデルを拡張した線形ダイナミカルシステムモデルとその変分ベイズ法による推定法を導き，人工データおよび実データに適用した．その結果，滑らかな時間変動を持つ因子を抽出し，さらに，その観測行列に各遺伝子（観測変数）の縮約された情報も含まれることを示した．

第 4 章では，遺伝子発現プロファイルからの癌分類問題を扱った．このために，多クラス分類問題を 2 値分類問題群に分解し，各問題での判別結果を確率モデルに基き統合することによって最適な識別結果を得る新たな多クラス識別法を提案した．さらに，2 値分類器に対する重みを導入することで，最適な 2 値分類問題のセットを自動的に決定する枠組みを提案した．本手法は，特に表現型に関して寄

与する遺伝子が少なく識別器の学習が非常に困難な場合でも，確実に正答率を向上させる出来た．このため，今後現れるであろう，発現プロファイルからの分類が困難な癌分類問題においても有効なアプリケーションとなることが期待される．

以上，いずれの問題においても，遺伝子発現情報の背後にある数理的モデルを仮定した統計的モデリングの枠組みが有効に働くことが示された．

謝辞

まず，本研究を遂行するにあたり，実質的なご指導をしていただいた大羽成征先生，また，指導教官として，様々なご指導，ご支援をしていただいた石井信先生に深くお礼申し上げます．ご両名の導きなくして，本論文の完遂は実現しなかったと思います．また，吉本潤一郎先生（沖縄大学院大学兼務）ならびに竹之内高志博士には，本研究での手法の導出および実装に関する様々なご協力をいただきました．ここに感謝申し上げます．

大阪府立成人病センター研究所の加藤菊也先生には，本研究で用いた様々なデータのご提供，解析結果に議論，投稿論文に関するご指導など，共同研究という形で多岐にわたるサポートしていただきました．ここに深くお礼申し上げます．また，加藤研究室の松尾浩子博士，小泉恭子博士，谷口一也さんからはデータのご提供および様々な情報をご教授を承り，大変お世話になりました．

小笠原直毅先生には，研究発表のたびに数多くの示唆に富んだご指摘をいただき，研究を進める上で多くのヒントを得ることができました．ありがとうございました．様々な形で研究活動と研究生活の両面に渡り様々な面でご支援していただいた，論理生命学分野のスタッフ，学生の皆様に深く感謝を申し上げます．

最後に，論文審査をこころよく引き受けてくださった石井信先生，小笠原直毅先生，加藤菊也先生，川端猛先生の皆様のご厚意に深く感謝申し上げます．

付録

A. バイアス補正と検量精度

crew 5 から crew 10 までの全 6 crew それぞれのキャリブレーションの成功率を表 5.1 から表 5.6 に，キャリブレーションによって得られた対数発現比の RMSE を表 5.7 から表 5.12 に示す.

表 5.1 キャリブレーションの成功率 (%) crew5

control type	raw				bias corrected			
	K	L0	L	frequecy	K	L0	L	frequecy
1	97.57	97.57	71.89	48.18	96.7	96.7	77.42	59.11
2	23.44	90.62	40.62	8.333	33	92	43	6.51
3	94.32	94.32	57.95	17.19	94.59	94.59	59.8	19.27
4	NaN	96.88	56.25	8.333	NaN	93.42	55.26	4.948
5	NaN	95	60	5.208	NaN	93.75	75	1.042
6	NaN	84.78	58.7	5.99	NaN	80.56	66.67	2.344
7	NaN	100	100	0.2604	NaN	100	100	0.2604
8	NaN	NaN	NaN	0	NaN	NaN	NaN	0
9	50	50	25	0.2604	50	50	25	0.2604
10	NaN	75	50	0.2604	NaN	75	50	0.2604
11	67.86	67.86	53.57	3.646	67.65	67.65	48.53	4.427
12	50	68.75	68.75	1.042	NaN	75	75	0.2604
13	NaN	NaN	NaN	0	NaN	NaN	NaN	0
14	NaN	NaN	NaN	0	NaN	NaN	NaN	0
15	33.33	33.33	33.33	0.7812	33.33	33.33	33.33	0.7812
16	NaN	NaN	NaN	0.5208	NaN	NaN	NaN	0.5208
all (1-16)	68.55	92.9	62.7	100	80.92	92.9	68.23	100

表 5.2 キャリブレーションの成功率 (%) crew6

control type	raw				bias corrected			
	K	L0	L	frequecy	K	L0	L	frequecy
1	98.32	98.32	71.06	77.6	98.22	98.22	87.46	76.82
2	36.76	97.79	31.62	8.854	50	100	25	0.5208
3	91.67	91.67	83.97	10.16	95.07	95.07	87.5	19.79
4	NaN	100	75	0.7812	NaN	100	58.33	0.7812
5	NaN	100	75	0.2604	NaN	100	75	0.2604
6	NaN	100	37.5	0.5208	NaN	NaN	NaN	0
7	NaN	NaN	NaN	0	NaN	NaN	NaN	0
8	NaN	NaN	NaN	0	NaN	NaN	NaN	0
9	50	50	25	0.5208	50	50	25	0.5208
10	NaN	25	25	0.2604	NaN	25	25	0.2604
11	NaN	NaN	NaN	0	NaN	NaN	NaN	0
12	NaN	100	75	0.5208	NaN	100	62.5	0.5208
13	NaN	NaN	NaN	0	NaN	NaN	NaN	0
14	NaN	75	75	0.2604	NaN	75	75	0.2604
15	25	25	25	0.2604	25	25	25	0.2604
16	NaN	NaN	NaN	0	NaN	NaN	NaN	0
all (1-16)	89.19	96.94	68.29	100	94.86	96.94	86.07	100

表 5.3 キャリブレーションの成功率 (%) crew7

control type	raw				bias corrected			
	K	L0	L	frequecy	K	L0	L	frequecy
1	99.02	99.02	95.52	79.95	99.08	99.08	96.23	84.64
2	28.57	96.43	28.57	1.823	10.71	92.86	35.71	1.823
3	98.26	98.26	88.95	11.2	94.83	94.83	83.62	7.552
4	NaN	87.5	66.67	1.562	NaN	100	93.75	1.042
5	NaN	87.5	50	0.5208	NaN	NaN	NaN	0
6	NaN	75	62.5	0.5208	NaN	75	62.5	0.5208
7	NaN	NaN	NaN	0	NaN	NaN	NaN	0
8	NaN	75	66.67	0.7812	NaN	75	66.67	0.7812
9	NaN	NaN	NaN	0	NaN	NaN	NaN	0
10	NaN	NaN	NaN	0	NaN	NaN	NaN	0
11	94.23	94.23	80.77	3.385	94.23	94.23	88.46	3.385
12	NaN	NaN	NaN	0	NaN	NaN	NaN	0
13	NaN	75	75	0.2604	NaN	75	75	0.2604
14	NaN	NaN	NaN	0	NaN	NaN	NaN	0
15	NaN	NaN	NaN	0	NaN	NaN	NaN	0
16	NaN	NaN	NaN	0	NaN	NaN	NaN	0
all (1-16)	93.88	98.11	91.93	100	94.4	98.11	93.42	100

表 5.4 キャリブレーションの成功率 (%) crew8

control type	raw				bias corrected			
	K	L0	L	frequecy	K	L0	L	frequecy
1	97.13	97.13	96.11	77.08	97.23	97.23	95.8	72.92
2	14.29	96.43	64.29	1.823	22.92	97.92	64.58	3.125
3	89.52	89.52	83.06	8.073	91.91	91.91	82.35	8.854
4	NaN	NaN	NaN	0	NaN	85	45	1.302
5	NaN	NaN	NaN	0	NaN	87.5	87.5	0.5208
6	NaN	NaN	NaN	0	NaN	75	25	0.2604
7	NaN	NaN	NaN	0	NaN	NaN	NaN	0
8	NaN	NaN	NaN	0	NaN	NaN	NaN	0
9	78.57	78.57	67.86	1.823	75	75	65	1.302
10	NaN	100	75	0.2604	NaN	91.67	58.33	0.7812
11	82.35	82.35	77.21	8.854	82.35	82.35	72.79	8.854
12	43.75	62.5	43.75	1.042	NaN	62.5	43.75	1.042
13	NaN	NaN	NaN	0	NaN	NaN	NaN	0
14	NaN	NaN	NaN	0	NaN	NaN	NaN	0
15	75	75	75	1.042	75	75	75	1.042
16	NaN	NaN	NaN	0	NaN	NaN	NaN	0
all (1-16)	92.32	94.27	91.47	100	88.8	94.27	89.26	100

表 5.5 キャリブレーションの成功率 (%) crew9

control type	raw				bias corrected			
	K	L0	L	frequecy	K	L0	L	frequecy
1	74.58	74.58	73.45	46.09	75.42	75.42	72.88	61.46
2	NaN	75	50	0.2604	15	60	45	1.302
3	76.37	76.37	73.9	23.7	75.93	75.93	75	7.031
4	NaN	50	25	0.5208	NaN	58.33	33.33	0.7812
5	NaN	NaN	NaN	0	NaN	NaN	NaN	0
6	NaN	NaN	NaN	0	NaN	NaN	NaN	0
7	NaN	25	25	0.2604	NaN	25	25	0.2604
8	NaN	50	50	0.2604	NaN	50	50	0.2604
9	58.7	58.7	54.35	5.99	58.7	58.7	57.61	5.99
10	NaN	75	75	0.2604	NaN	75	75	0.2604
11	55.77	55.77	54.81	13.54	56.5	56.5	52.5	13.02
12	NaN	25	NaN	0.5208	12.5	31.25	12.5	1.042
13	NaN	NaN	NaN	0.2604	NaN	NaN	NaN	0.2604
14	NaN	NaN	NaN	0	NaN	NaN	NaN	0
15	37.07	37.07	37.07	7.552	37.07	37.07	37.07	7.552
16	NaN	NaN	NaN	0.7812	NaN	NaN	NaN	0.7812
all (1-16)	66.34	67.32	65.49	100	65.69	67.32	64.52	100

表 5.6 キャリブレーションの成功率 (%) crew10

control type	raw				bias corrected			
	K	L0	L	frequecy	K	L0	L	frequecy
1	83.77	83.77	81.25	59.38	83.9	83.9	72.99	61.46
2	12.5	37.5	25	0.5208	25	56.25	18.75	1.042
3	78.33	78.33	72.5	15.62	77.55	77.55	63.78	12.76
4	NaN	75	75	0.2604	NaN	62.5	62.5	0.5208
5	NaN	NaN	NaN	0	NaN	NaN	NaN	0
6	NaN	37.5	37.5	0.5208	NaN	37.5	37.5	0.5208
7	NaN	NaN	NaN	0	NaN	NaN	NaN	0
8	NaN	NaN	NaN	0	NaN	NaN	NaN	0
9	55	55	47.5	2.604	55	55	45	2.604
10	NaN	NaN	NaN	0.2604	NaN	NaN	NaN	0.2604
11	57.65	57.65	50	12.76	56.52	56.52	45.65	11.98
12	NaN	33.33	25	0.7812	NaN	54.17	41.67	1.562
13	NaN	12.5	12.5	0.5208	NaN	12.5	12.5	0.5208
14	NaN	33.33	33.33	0.7812	NaN	33.33	33.33	0.7812
15	17.65	17.65	17.65	4.427	17.65	17.65	17.65	4.427
16	NaN	NaN	NaN	1.562	NaN	NaN	NaN	1.562
all (1-16)	71.61	72.72	69.01	100	70.7	72.72	62.11	100

表 5.7 コントロールタイプと RMSE crew5

control type	raw			bias corrected		
	K	L0	L	K	L0	L
1	1.85	1.75	1.62	1.85	2.09	1.61
2	2.62	1.86	4.27	2.21	1.71	3.44
3	2.71	1.8	1.92	2.76	1.95	2.18
4	NaN	2.1	4.27	NaN	1.92	4.52
5	NaN	2.13	3.04	NaN	1.73	2.66
6	NaN	2.52	2.5	NaN	2.37	2.42
7	NaN	2.57	0.75	NaN	2.79	0.777
8	NaN	NaN	NaN	NaN	NaN	NaN
9	1.88	1.67	3.11	1.74	0.639	0.54
10	NaN	5.35	2.1	NaN	4.51	2.87
11	1.98	2.21	1.65	2.82	2.88	1.04
12	2.54	2.6	4.44	NaN	2.6	1.88
13	NaN	NaN	NaN	NaN	NaN	NaN
14	NaN	NaN	NaN	NaN	NaN	NaN
15	4.24	4.24	4.24	3.83	3.83	3.83
16	NaN	NaN	NaN	NaN	NaN	NaN
all (1-16)	2.13	1.92	2.37	2.15	2.08	2.05

表 5.8 コントロールタイプと RMSE crew6

control type	raw			bias corrected		
	K	L0	L	K	L0	L
1	1.6	1.62	1.67	1.83	2.09	1.58
2	1.34	1.89	2.25	5.52	3.71	0.653
3	2.54	2.33	1.43	2.85	2.78	1.42
4	NaN	1.48	2.95	NaN	2.43	2.58
5	NaN	3.69	1.59	NaN	3.56	1.61
6	NaN	3.14	2.63	NaN	NaN	NaN
7	NaN	NaN	NaN	NaN	NaN	NaN
8	NaN	NaN	NaN	NaN	NaN	NaN
9	3.7	2.53	2.54	4.14	3.12	2.45
10	NaN	3.11	0.544	NaN	3.11	0.392
11	NaN	NaN	NaN	NaN	NaN	NaN
12	NaN	3.43	1.48	NaN	3.01	1.7
13	NaN	NaN	NaN	NaN	NaN	NaN
14	NaN	1.92	1.92	NaN	1.25	1.25
15	2.94	2.94	2.94	2.17	2.17	2.17
16	NaN	NaN	NaN	NaN	NaN	NaN
all (1-16)	1.73	1.77	1.69	2.1	2.27	1.56

表 5.9 コントロールタイプと RMSE crew7

control type	raw			bias corrected		
	K	L0	L	K	L0	L
1	1.16	1.32	1.21	1.47	1.95	1.37
2	1.74	2.33	3.27	1.43	2.55	1.97
3	2.12	1.51	1.39	2.47	1.99	1.38
4	NaN	2.86	2.71	NaN	2.58	1.74
5	NaN	1.84	1.54	NaN	NaN	NaN
6	NaN	3	1.33	NaN	2.56	1.38
7	NaN	NaN	NaN	NaN	NaN	NaN
8	NaN	3.94	2.8	NaN	3.68	3.06
9	NaN	NaN	NaN	NaN	NaN	NaN
10	NaN	NaN	NaN	NaN	NaN	NaN
11	2.12	2.31	1.71	2.46	2.96	1.57
12	NaN	NaN	NaN	NaN	NaN	NaN
13	NaN	3.97	3.97	NaN	4.03	4.03
14	NaN	NaN	NaN	NaN	NaN	NaN
15	NaN	NaN	NaN	NaN	NaN	NaN
16	NaN	NaN	NaN	NaN	NaN	NaN
all (1-16)	1.35	1.49	1.32	1.61	2.04	1.41

表 5.10 コントロールタイプと RMSE crew8

control type	raw			bias corrected		
	K	L0	L	K	L0	L
1	1.38	2.5	1.27	1.16	1.58	1.22
2	3.72	2.33	2	1.85	1.85	3.01
3	2.6	2.53	1.73	2.28	1.86	1.4
4	NaN	NaN	NaN	NaN	1.28	3.16
5	NaN	NaN	NaN	NaN	1.84	1.27
6	NaN	NaN	NaN	NaN	6.3	0.814
7	NaN	NaN	NaN	NaN	NaN	NaN
8	NaN	NaN	NaN	NaN	NaN	NaN
9	2.77	2.52	2.36	2.27	1.9	2.19
10	NaN	2.04	1.54	NaN	2.87	1.45
11	1.94	2.7	1.35	1.85	2.02	1.26
12	1.74	2.3	3.75	NaN	2.24	1.78
13	NaN	NaN	NaN	NaN	NaN	NaN
14	NaN	NaN	NaN	NaN	NaN	NaN
15	2.83	2.83	2.83	2.93	2.93	2.93
16	NaN	NaN	NaN	NaN	NaN	NaN
all (1-16)	1.62	2.51	1.39	1.41	1.71	1.36

表 5.11 コントロールタイプと RMSE crew9

control type	raw			bias corrected		
	K	L0	L	K	L0	L
1	1.75	2.34	1.34	1.05	1.19	1.1
2	NaN	0.288	2.82	3.09	1.67	4.84
3	2.2	2.23	1.43	1.17	1.07	0.964
4	NaN	0.674	2.51	NaN	1.32	4.77
5	NaN	NaN	NaN	NaN	NaN	NaN
6	NaN	NaN	NaN	NaN	NaN	NaN
7	NaN	1.59	1.52	NaN	1.23	1.25
8	NaN	2.08	3.17	NaN	1.64	4.03
9	2.1	2.23	2.05	1.54	1.49	1.7
10	NaN	0.806	1.85	NaN	1.03	1.75
11	1.35	1.98	1.34	1.26	1.35	1.18
12	NaN	0.57	NaN	1.75	1.28	4.01
13	NaN	NaN	NaN	NaN	NaN	NaN
14	NaN	NaN	NaN	NaN	NaN	NaN
15	2.19	2.19	2.19	2.15	2.15	2.15
16	NaN	NaN	NaN	NaN	NaN	NaN
all (1-16)	1.89	2.25	1.47	1.2	1.28	1.34

表 5.12 コントロールタイプと RMSE crew10

control type	raw			bias corrected		
	K	L0	L	K	L0	L
1	1.73	2.53	1.32	1.46	1.52	1.75
2	0.571	1.47	1.94	2.49	1.25	2.14
3	2.39	2.38	1.66	2.12	1.44	1.25
4	NaN	0.828	2.35	NaN	2.62	4.21
5	NaN	NaN	NaN	NaN	NaN	NaN
6	NaN	0.492	2.19	NaN	1.16	1.6
7	NaN	NaN	NaN	NaN	NaN	NaN
8	NaN	NaN	NaN	NaN	NaN	NaN
9	2.02	2.54	1.69	1.99	1.84	1.78
10	NaN	NaN	NaN	NaN	NaN	NaN
11	1.56	2.09	1.31	1.4	1.31	0.958
12	NaN	2.62	2.41	NaN	2.28	2.88
13	NaN	1.56	1.56	NaN	1.06	1.06
14	NaN	3.1	3.1	NaN	3.48	3.48
15	1.92	1.92	1.92	1.55	1.55	1.55
16	NaN	NaN	NaN	NaN	NaN	NaN
all (1-16)	1.85	2.45	1.42	1.58	1.53	1.69

B. 2 値分類器の判別関数値から確率値への変換

学習データセット $\boldsymbol{L} \equiv \{(\boldsymbol{x}^{(n)}, \boldsymbol{t}^{(n)})|n = 1 : N\}$ を用いてターゲット $[l|m]$ に関する 2 値分類器の判別関数 $f^{\boldsymbol{L}}_{[l|m]}(\boldsymbol{x}) \in \mathcal{R}$ が得られたとする．このとき，データ $\boldsymbol{x}$ のクラスラベルが l に含まれる確率 $q_{[l|m]}(\boldsymbol{x}) \equiv Pr\left(c(\boldsymbol{t}) \in l | f^{\boldsymbol{L}}_{[l|m]}(\boldsymbol{x}) c(\boldsymbol{t}) \in l \cup m\right)$ を，以下の logistic regression モデルで表す．

$$q_{[l|m]}(\boldsymbol{x}) = \frac{1}{1 + \exp(A_{[l|m]} f^{\boldsymbol{L}}_{[l|m]}(\boldsymbol{x}_i) + B_{[l|m]})}. \tag{5.1}$$

パラメータ $A_{[l|m]}, B_{[l|m]}$ は，2 値分類器の学習に用いたデータを用い，以下の対数尤度 L を最大化することにより最尤推定する．

$$L \equiv \sum_n \left\{ s^{(n)} \log(q_{[l|m]}(\boldsymbol{x}^{(n)})) + (1 - s^{(n)}) \log(1 - q_{[l|m]}(\boldsymbol{x}^{(n)})) \right\}. \tag{5.2}$$

ここで $s^{(n)}$ は

$$s^{(n)} = \begin{cases} 1 & \text{if } c(\boldsymbol{t}^{(n)}) \in l \\ 0 & \text{if } c(\boldsymbol{t}^{(n)}) \in m, \end{cases} \tag{5.3}$$

とする．L の最適化には勾配法を用いた．

C. WMAP 法の導出

式 (4.11) の最適化は MAP 法で可能であるが、式 (4.10) の最適化は、U が $\tilde{\boldsymbol{p}} \equiv \{\tilde{\boldsymbol{p}}^{(n)}\}$ を通じて間接的に $\boldsymbol{w}$ に依存しているため、少々の工夫が必要である。そこで、以下のように $\boldsymbol{w}, \boldsymbol{p} \equiv \{\boldsymbol{p}^{(n)}\}$ の関数 $f(\boldsymbol{w}, \boldsymbol{p})$ を定義する。

$$f(\boldsymbol{w}, \boldsymbol{p}) \equiv \frac{\partial}{\partial \boldsymbol{p}} \tilde{V}(\boldsymbol{p}|\boldsymbol{w}).$$

$\tilde{V}$ は V にラグランジュ未定係数項が付加されたものである。式 (4.11) を満たす $\tilde{\boldsymbol{p}}(\boldsymbol{w})$ について、

$$f(\boldsymbol{w}, \tilde{\boldsymbol{p}}) = 0$$

が成立する。このとき、$\boldsymbol{w}$ が微少変動 $d\boldsymbol{w}$ した際の、式 (4.11) の解 $\tilde{\boldsymbol{p}} + d\boldsymbol{p}$ について、

$$\begin{aligned} f(\boldsymbol{w} + d\boldsymbol{w}, \tilde{\boldsymbol{p}} + d\boldsymbol{p}) &= f(\boldsymbol{w}, \tilde{\boldsymbol{p}}) + \frac{\partial}{\partial \boldsymbol{w}} f(\boldsymbol{w} + \theta d\boldsymbol{w}, \tilde{\boldsymbol{p}} + \theta d\boldsymbol{p}) d\boldsymbol{w} + \frac{\partial}{\partial \tilde{\boldsymbol{p}}} f(\boldsymbol{w} + \theta d\boldsymbol{w}, \tilde{\boldsymbol{p}} + \theta d\boldsymbol{p}) d\boldsymbol{p} \\ &= \sum_{[l|m] \in B} \frac{\partial f}{\partial w_{[l|m]}} dw_{[l|m]} + \sum_{i \in C} \sum_{n=1}^{N} \frac{\partial f}{\partial p_i^{(n)}} dp_i^{(n)} \\ &= 0 \end{aligned} \tag{5.4}$$

が成立する。ここで $0 < \theta < 1$ である。

表記の分かりやすさのために、ここで行列 $\boldsymbol{A} = \{a(\mu, \nu)\}$ を導入する。ただしインデクス μ は $\boldsymbol{p}$ の各成分 $p_i^{(n)}$ に対応し、ν は 各ターゲット $[l|m] \in B$ に対応する。行列 $\boldsymbol{A}$ の各要素は以下で定義される。

$$a([l|m], (i, n)) = \frac{\partial^2 \tilde{V}}{\partial w_{[l|m]} \partial p_i^{(n)}}.$$

また正方行列 $\boldsymbol{H} = \{h(\mu, \mu')\}$ を導入する。ただしインデクス μ, μ' はそれぞれ $\boldsymbol{p}$ の各成分 $p_i^{(n)}$ に対応し、各要素は以下で定義される。

$$h((i, n), (i', n')) = \frac{\partial^2 \tilde{V}}{\partial p_i^{(n)} \partial p_i'^{(n')}}.$$

これらを用いると、陰関数で表現された条件 (式 (5.4)) は

$$\boldsymbol{A}d\boldsymbol{w} + \boldsymbol{H}d\boldsymbol{p} = 0$$

と書くことができ、$d\boldsymbol{w} \to \boldsymbol{0}$ のとき

$$\left\{\frac{dp_i^{(n)}}{dw_{[l|m]}}\right\} \equiv \frac{d\boldsymbol{p}}{d\boldsymbol{w}} = -\boldsymbol{H}^{-1}\boldsymbol{A}$$

となる。この微分を用いると、

$$\frac{\partial U}{\partial \boldsymbol{w}} = \frac{\partial \tilde{\boldsymbol{p}}}{\partial \boldsymbol{w}}\frac{\partial U}{\partial \tilde{\boldsymbol{p}}} = -\boldsymbol{H}^{-1}\boldsymbol{A}\frac{\partial U}{\partial \boldsymbol{p}} \tag{5.5}$$

となる。$\partial U/\partial \boldsymbol{p}$ の各成分は

$$\frac{\partial U}{\partial p_i^{(n)}} = \left(1 - \frac{\exp(\beta p_i^{(n)})}{\sum_{i' \in C}\exp(\beta p_{i'}^{(n)})}\right)\frac{\exp(\beta p_i^{(n)})}{\sum_{i' \in C}\exp(\beta p_{i'}^{(n)})}\beta t_i^{(n)}$$

のように書き下せるため、式 (4.10) の最適化は勾配 (式 (5.5)) に基づき行うことができる。

参考文献

[1] P. L. Zamorano, V. B. Mahesh, and D. W. Brann. Quantitative RT-PCR for neuroendocrine studies. A minireview. *Neuroendocrinology*, Vol. 63, No. 5, pp. 397–407, May 1996.

[2] R. M. Parker and N. M. Barnes. mRNA: detection by in Situ and northern hybridization. *Methods Mol Biol*, Vol. 106, pp. 247–283, 1999.

[3] Y. Hod. A simplified ribonuclease protection assay. *Biotechniques*, Vol. 13, No. 6, pp. 852–854, Dec 1992.

[4] C. F. Saccomanno, M. Bordonaro, J. S. Chen, and J. L. Nordstrom. A faster ribonuclease protection assay. *Biotechniques*, Vol. 13, No. 6, pp. 846–850, Dec 1992.

[5] J. H. Weis, S. S. Tan, B. K. Martin, and C. T. Wittwer. Detection of rare mRNAs via quantitative RT-PCR. *Trends Genet*, Vol. 8, No. 8, pp. 263–264, Aug 1992.

[6] J. L. DeRisi, V. R. Iyer, and P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, Vol. 278, No. 5338, pp. 680–686, Oct 1997.

[7] D. A. Lashkari, J. L. DeRisi, J. H. McCusker, A. F. Namath, C. Gentile, S. Y. Hwang, P. O. Brown, and R. W. Davis. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc Natl Acad Sci U S A*, Vol. 94, No. 24, pp. 13057–13062, Nov 1997.

[8] R. J. Lipshutz, D. Morris, M. Chee, E. Hubbell, M. J. Kozal, N. Shah, N. Shen, R. Yang, and S. P. Fodor. Using oligonucleotide probe arrays to access genetic diversity. *Biotechniques*, Vol. 19, No. 3, pp. 442–447, Sep 1995.

[9] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi andH. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotechnol*, Vol. 14, No. 13, pp. 1675–1680, Dec 1996.

[10] R. J. Lipshutz, S. P. Fodor, T. R. Gingeras, and D. J. Lockhart. High density synthetic oligonucleotide arrays. *Nat Genet*, Vol. 21, No. 1 Suppl, pp. 20–24, Jan 1999.

[11] K. Kato. Adaptor-tagged competitive PCR: a novel method for measuring relative gene expression. *Nucleic Acids Res*, Vol. 25, No. 22, pp. 4694–4696, Nov 1997.

[12] S. Su, R. G. Vivier, M. C. Dickson, N. Thomas, M. K. Kendrick, N. M. Williamson, J. G. Anson, J. G. Houston, and F. F. Craig. High-throughput RT-PCR analysis of multiple transcripts using a microplate RNA isolation procedure. *Biotechniques*, Vol. 22, No. 6, pp. 1107–1113, Jun 1997.

[13] H. Tian, L. Cao, Y. Tan, S. Williams, L. Chen, T. Matray, A. Chenna, S. Moore, V. Hernandez, V. Xiao, M. Tang, and S. Singh. Multiplex mRNA assay using electrophoretic tags for high-throughput gene expression analysis. *Nucleic Acids Res*, Vol. 32, No. 16, p. e126, 2004. Evaluation Studies.

[14] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, Vol. 95, No. 25, pp. 14863–14868, Dec 1998.

[15] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*, Vol. 98, No. 9, pp. 5116–5121, Apr 2001.

[16] J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A*, Vol. 100, No. 16, pp. 9440–9445, Aug 2003.

[17] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M Gaasenbeek, J. P. Mesirov, H Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, Vol. 286, No. 5439, pp. 531–537, Oct 1999.

[18] H. de Jong. Modeling and simulation of genetic regulatory system: a literature review. *Journal of Computational Biology*, Vol. 9, pp. 67–103, 2002.

[19] N. Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, Vol. 303, No. 5659, pp. 799–805, Feb 2004.

[20] K. Basso, A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, and A. Califano. Reverse engineering of regulatory networks in human B cells. *Nat Genet*, Vol. 37, No. 4, pp. 382–390, Apr 2005.

[21] J. Quackenbush. Computational analysis of microarray data. *Nature Review Genetics*, Vol. 2, pp. 418–427, jun 2001.

[22] A. N. Jain N, T. A. Tokuyasu, A. M. Snijders, R. Segraves, D. G. Albertson, and D. Pinkel. Fully automatic quantification of microarray image data. *Genome Res*, Vol. 12, No. 2, pp. 325–332, Feb 2002.

[23] W. S. Cleveland and S. J. Devlin. An Approach to Fitting Analysis by Local Fitting. *J Am Stat Assoc*, Vol. 83, pp. 596–610, 1988.

[24] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for DNA microarrays. *Bioinformatics*, Vol. 17, No. 6, pp. 520–525, Jun 2001. Evaluation Studies.

[25] S. Oba, M. Sato, I. Takemasa, M. Monden, K. Matsubara, and S. Ishii. A bayesian missing value estimation method. *Bioinformatics*, Vol. 19, pp. 2088–2096, 2003.

[26] U. E. Gibson, C. A. Heid, and P.M. Williams. A novel method for real-time quantitative rt-pcr. *Genome Res*, Vol. 6, pp. 995–1001, 1996.

[27] C. A. Heid, J. Stevens, K. J. Livak, and P. M. Williams. Real time quantitative pcr. *Genome Res*, Vol. 6, pp. 986–994, 1996.

[28] M. Becker-Andre and K. Hahlbrock. Absolute mrna quantification using the polymerase chain reaction (pcr): a novel approach by a pcr aided transcript titration assay (patty). *Nucleic Acids Res*, Vol. 22, pp. 9437–9446, 1989.

[29] *PCR Protocols: A Guide to Methods and Applications*. Academic Press, 1990.

[30] S. Kawamoto, T. Ohnishi, H. Kita, O. Chisaka, and K. Okubo. Expression profiling by iaflp: a pcr-based method for genome-wide gene expression profiling. *Genome Res*, Vol. 9, pp. 1305–1312, 1999.

[31] S. Saito, R. Matoba, and K. Kato. Adapter-tagged competitive pcr (atac-pcr): a high-throughput quantitative pcr method for microarray validation. *Methods*, Vol. 31, pp. 326–331, 2003.

[32] K. Iwao, R. Matoba, N. Ueno, A. Ando, Y. Miyoshi, K. Matsubara, S. Noguchi, and K. Kato. Molecular classification of primary breast tumors possessing distinct prognostic properties. *Hum Mol Genet*, Vol. 11, pp. 199–206, 2002.

[33] H. Kita, J. Carmichael, J. Swartz, S. Muro, A. Wyttenbach, K. Matsubara, D.C. Rubinsztein, and K. Kato. Modulation of polyglutamine-induced cell death by genes identified by expression profiling. *Hum Mol Genet*, Vol. 11, pp. 2279–2287, 2002.

[34] S. Muro, I. Takemasa, S. Oba, R. Matoba, N. Ueno, C. Maruyama, R. Yamashita, M. Sekimoto, H. Yamamoto, S. Nakamori, M. Monden, S. Ishii, and K. Kato.

Identification of expressed genes linked to malignancy of human colorectal carcinoma by parametric clustering of quantitative expression data. *Genome Biol*, Vol. 4, pp. 1–10, 2003.

[35] Y. Kurokawa, R. Matoba, I. Takemasa, S. Nakamori, M. Tsujie, H. Nagano, K. Dono, K. Umeshita, M. Sakon, N. Ueno, H. Kita, S. Oba, S. Ishii, K. Kato, and M. Monden. Molecular features of non-b, non-c hepatocellular carcinoma: a pcr-array gene expression profiling study. *J Hepatol*, Vol. 39, pp. 1004–1012, 2003.

[36] R. Matoba, K. Kato, C. Kurooka, C. Maruyama, Y. Sakakibara, and K. Matsubara. Correlation between gene function and developmental expression pattern in the mouse cerebellum. *Eur J Neurosci*, Vol. 12, pp. 1357–1371, 2000.

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

[38] T. G. Dewey and D. J. Galas. Dynamic models of gene expression and classification. *Functional & Integrative Genomics*, Vol. 1, pp. 269–278, 2001.

[39] N. S. Holter, A. Maritan, M. Cieplak, N.V. Fedoroff, and J. R. Banavar. Dynamic modeling of gene expression data. *Proceedings of National Academy of Sciences of the United States of America*, Vol. 98, pp. 1693–1698, 2001.

[40] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of Royal Statistical Society B*, Vol. 39, pp. 1–38, 1977.

[41] S. Roweis and Z. Gharahmani. A unifying review of linear gaussian models. *Neural Computation*, Vol. 11, pp. 305–345, 1999.

[42] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, Vol. 6, pp. 461–464, 1978.

[43] H. Attias. Inferring parameters and structure of latent variable models by variational bayes. In *Proceedings of 15th Conference on Uncertainty in Artificial Inteligence*, pp. 21–30, 1999.

[44] Z. Gharahmani and M. J. Beal. Propagation algorithms for variational bayesian learning. In *Advances in Neural Information Processing Systems 13*, pp. 507–513, 2001.

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

[46] P. T. Spellman, G. Sherlock, M. Q. Zang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, and B. Futcher. 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.

[47] J. Yoshimoto, S. Ishii, and M. Sato. Hierarchical model selection for ngnet based on variational bayes inference. In *Artificial Neural Networks - ICANN 2002*, Lecture Notes in Computer Science 2415, pp. 661–666, 2002.

[48] M. T. Beal and Z. Ghahramani. The variational bayesian em algorithm for imcomplete data: with application to scoring graphical model structures. *Bayesian Statistics*, Vol. 7, pp. 453–464, 2003.

[49] D. Rubin and D. Thayer. Em algorithms for ml factor analysis. *Psychometrika*, Vol. 47, pp. 69–76, 1982.

[50] D. H. Arvine and M. A. Savageau. Efficient solution of nonlinear ordinary differential equations expressed in s-system canonical form. *SIAM Journal on Numerical Analysis*, Vol. 27, pp. 704–735, 1998.

[51] D. Tominaga and M. Okamoto. Design of canonical model describing complex nonlinear dynamics. In *Proceedings of IFAC International Conference*, CAB7, pp. 85–90, 1998.

[52] S. Liang, S. Fuhrman, and R. Somogyi. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. In *Pacific Symposium on Biocomputing*, Vol. 3, pp. 18–29, 1998.

[53] T. Akutsu, S. Miyano, and S. Kuhara. Identification of genetic networks from a small number of gene expression patterns under the boolean network model. In *Pacific Symposium on Biocomputing*, Vol. 4, pp. 17–28, 1999.

[54] T. Akutsu, S. Miyano, and S. Kuhara. Inferring qualitative relations in genetic networks and metabolic pathways. *Bioinformatics*, Vol. 16, pp. 727–734, 2000.

[55] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med*, Vol. 7, No. 6, pp. 673–679, Jun 2001.

[56] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.H Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, T. Poggio, W. Gerald, M Loda, E. S. Lander, and T. R. Golub. Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A*, Vol. 98, No. 26, pp. 15149–15154, Dec 2001.

[57] I. Hedenfalk, M. Ringnerand A. Ben-Dor, Z. Yakhini, Y. Chen, G. Chebil, R. Ach, N. Loman, H. Olsson, P. Meltzer, A. Borg, and J. Trent. Molecular classification of familial non-BRCA1/BRCA2 breast cancer. *Proc Natl Acad Sci U S A*, Vol. 100, No. 5, pp. 2532–2537, Mar 2003.

[58] S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classfication of tumors using gene expression data. *J Am Stat Assoc*, Vol. 97, No. 457, pp. 77–87, mar 2002.

[59] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley Interscience, 2000.

[60] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A*, Vol. 99, No. 10, pp. 6567–6572, May 2002.

[61] S. Oba, M. Sato, and S. Ishii. Variational Bayes method for Mixture of Principal Component Analyzers. In *Proceeding for 7th International Conference on Neural Information Processing (ICONIP2000)*, Vol. 2, pp. 1416–1421, 2000.

[62] B. Schölkopf and C. Burges and V. Vapnik. Extracting support data for a given task. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, pp. 252–257, 1995.

[63] V. Vapnik. *Statistical Learning Theory*. Wiley, NY, 1998.

[64] Y. Freund and R.E. Schapire. Experiments with a new boosting algorithm. In *International Conference on Machine Learning (ICML)*, pp. 148–156, 1996.

[65] J. Weston and C. Watkins. Multi-class support vector machine. Technical report, University of London, 1998.

[66] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, Vol. 2, pp. 265–292, 2001.

[67] T. Hastie and R. Tibshirani. Classification by pairwise coupling. In *Advances in Neural Information Processing Systems (NIPS)*, Vol. 10, pp. 507–513, 1998.

[68] B. Schölkopf and C. Burges and A. Smola. *Advances in Kernel Methods Support Vector Learning*. The MIT Press, 1999.

[69] D. Tax and R. P. W. Duin. Using two-class classifiers for multi-class classification. In *Proceedings 16th International Conference on Pattern Recognition (ICPR)*, Vol. 2, pp. 124–127, 2002.

[70] T. Li, C. Zhang, and M. Ogihara. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, Vol. 20, No. 15, pp. 2429–2437, Oct 2004.

[71] T. G. Dietterich and G. Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, Vol. 2, pp. 263–286, 1995.

[72] E. L. Allwein, R. E. Schapire, and Y. Singer. Reducing multiclass to binary: A unifying approach for margin classifiers. In *Proc. 17th International Conf. on Machine Learning*, pp. 9–16. Morgan Kaufmann, San Francisco, CA, 2000.

[73] A. Statnikov, C. F. Aliferis, I. Tsamardinos, D. Hardin, and S. Levy. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, Vol. 21, No. 5, pp. 631–643, 2005.

[74] B. Zadrozny. Reducing multiclass to binary by coupling probability estimates. In *Advances in Neural Information Processing Systems (NIPS)*, Vol. 14, pp. 1041–1048, 2001.

[75] J. Anderson. Logistic discrimination. *Biometrika*, Vol. 59, pp. 19–35, 1972.

[76] J. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In A. J. Smola, P. Bartlett, B. Schoelkopf, and D. Schuurmans, editors, *Advances in Large Margin Classiers*, pp. 61–74, 2000.

[77] C. Chang and C. Lin. *LIBSVM: a library for support vector machines*, 2001.

[78] E. Saxen, K. Franssila, O. Bjarnason, T. Normann, and N. Ringertz. Observer variation in histologic classification of thyroid cancer. *Acta Pathol Microbiol Scand [A]*, Vol. 86A, No. 6, pp. 483–486, Nov 1978.

[79] A. S. Fassina, M. C. Montesco, V. Ninfo, P. Denti, and G. Masarotto. Histological evaluation of thyroid carcinomas: reproducibility of the "WHO" classification. *Tumori*, Vol. 79, No. 5, pp. 314–320, Oct 1993.

[80] Z. W. Baloch, S. Fleisher, V. A. LiVolsi, and P. K. Gupta. Diagnosis of "follicular neoplasm": a gray zone in thyroid fine-needle aspiration cytology. *Diagn Cytopathol*, Vol. 26, No. 1, pp. 41–44, Jan 2002.

[81] K. Taniguchi, T. Takano, A. Miyauchi, K. Koizumi, Y. Ito, Y. Takamura, M. Ishitobi, Y. Miyoshi, T. Taguchi, Y. Tamaki, K. Kato, and S. Noguchi. Differentiation of Follicular Thyroid Adenoma from Carcinoma by Means of Gene Expression Profiling with Adapter-Tagged Competitive Polymerase Chain Reaction. *Oncology*, Vol. 69, No. 5, pp. 428–435, Nov 2005.

[82] K. Kato, R. Yamashita, R. Matoba, M. Monden, S. Noguchi, T. Takagi, and K. Nakai. Cancer gene expression database (CGED): a database for gene expression profiling with accompanying clinical information of human cancer tissues. *Nucleic Acids Res*, Vol. 33, No. Database issue, pp. 533–536, Jan 2005.

[83] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet*, Vol. 30, No. 1, pp. 41–47, Jan 2002.

[84] M. Ohira, S. Oba, Y. Nakamura, E. Isogai, S. Kaneko, A. Nakagawa, T. Hirata, H. Kubo, T. Goto, S. Yamada, Y. Yoshida, M. Fuchioka, S. Ishii, and A. Nakagawara. Expression profiling using a tumor-specific cDNA microarray predicts the prognosis of intermediate risk neuroblastomas. *Cancer Cell*, Vol. 7, No. 4, pp. 337–350, Apr 2005.

[85] T. Barrett, T. O. Suzek, D. B. Troup, S. E. Wilhite, W. Ngau, P. Ledoux, D. Rudnev, A. E. Lash, W. Fujibuchi, and R. Edgar. NCBI GEO: mining millions of

expression profiles–database and tools. *Nucleic Acids Res*, Vol. 33, No. Database issue, pp. 562–566, Jan 2005.

[86] Z. Ghahramani and H. Kim. Bayesian combination of classifier. Technical report, Gatsby Computational Neuroscience Unit University College London, 2003.

業績リスト

学術論文

1. H. Kita-Matsuo, N. Yukinawa, R. Matoba, S. Saito, S. Oba, S. Ishii, and K. Kato. Adaptor-tagged competitive polymerase chain reaction: amplification bias and quantified gene expression levels. *Analytical Biochemistry* 339, pp.15-28, 2005.

2. 行縄 直人, 吉本 潤一郎, 大羽 成征, 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, 情報処理学会論文誌: 数理モデル化と応用, **46**(10), pp.57-65, 2005.

3. N. Yukinawa, S. Oba, K. Kato, K. Taniguchi, K. Iwao-Koizumi, Y. Tamaki, S. Noguchi. and S. Ishii. A multi-class predictor based on a probabilistic model: application to gene expression profiling-based diagnosis of thyroid tumors. *BMC Genomics* (in revision).

国際会議

1. N. Yukinawa, J. Yoshimoto, S. Oba, and S. Ishii. Modeling gene expression dynamics based on a linear dynamical system model. *International Symposium on Nonlinear Theory and its Applications (NOLTA)*, pp.577-580, 2004.

2. N. Yukinawa, S. Oba, K. Kato, and S. Ishii. Multi-class pattern classification based on a probabilistic model of combining binary classifiers. *International Conference on Artificial Neural Networks (ICANN 2005)*, Lecture Notes in Computer Science, 3697, pp.337-342, 2005.

その他の業績

1. 行縄 直人, 吉本 潤一郎, 大羽 成征, 石井 信. 線形ダイナミカルシステムモデルによる細胞周期制御のシステム同定. 日本神経回路学会第 13 回全国大会, pp.138-139, 2003.

2. N. Yukinawa, J. Yoshimoto, S. Oba, S. Ishii. System identification of cell-cycle-regulated genes based on a linear dynamical system model, *Pacific Symposium on Biocomputing (PSB) 2004*, 2004.

3. 行縄 直人, 大羽 成征, 加藤 菊也, 石井 信. 二値分類器組み合わせの確率モデルによる多クラスパターン識別. 電子情報通信学会技術研究報告, NC2004-221, **104**(760), pp.165-170, 2005.

4. 行縄 直人, 大羽 成征, 加藤 菊也, 石井 信. 二値分類器集合による遺伝子発現プロファイルからの癌サブクラス識別法. 情報処理学会研究報告, Vol.2005, No.74, バイオ情報学研究会 (1), 2005.