光度曲線データを用いた線形判別による変光星の自動分類

広島大学理学部物理科学科
高エネルギー宇宙・可視赤外線天文学研究室
B125916
安部 太晴

主査: 植村 誠　副査: 樋口 克彦

2015 年 2 月

目 次

第 1 章　研究の背景 〜変光星とその分類〜 **1**
1.1　変光星とサーベイプロジェクト 1
1.2　変光星の型の自動判別 2
1.3　本研究の目的 2

第 2 章　判別手法およびその理論 **4**
2.1　線形判別 4
2.1.1　2 クラス線形判別 4
2.1.2　多クラス線形判別 5
2.2　フィッシャーの線形判別 6
2.2.1　フィッシャーの線形判別 – 2 クラス 6
2.2.2　フィッシャーの線形判別 – 多クラス 7
2.3　人工データを用いた判別実験 8

第 3 章　データ解析 **16**
3.1　判別に用いたデータ 16
3.2　判別対象の変光星型 16
3.3　使用パラメータ 22
3.3.1　光度曲線から抽出した特徴量 23
3.3.2　パワースペクトルから抽出した特徴量 23
3.4　判別モデルの性能評価と交差検定 25
3.4.1　10 分割交差検定 25
3.4.2　1 個抜き交差検定 28

第 4 章　結果 **29**
4.1　10 分割交差検定を用いた多サンプルの判別 29
4.2　1 個抜き交差検定を用いた少サンプルの判別 29

第 5 章　考察 **32**
5.1　$\nu_{\max}$ のクラス内分散と判別正答率 32
5.2　セファイドと食連星の判別 32
5.3　L_{mean} の有用性 34
5.4　サンプル数の影響 40

第 6 章　まとめ **41**

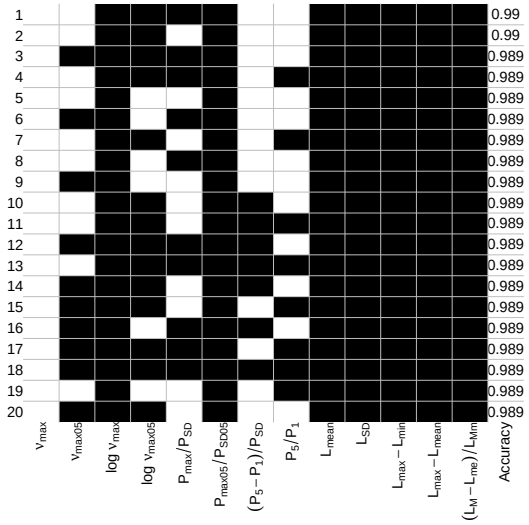
概 要

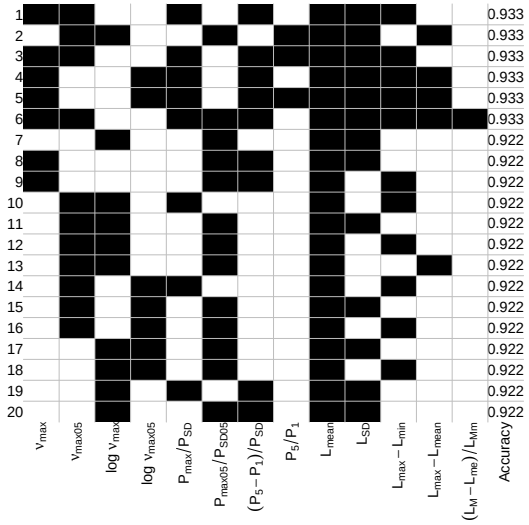
変光星はその変動パターンからいくつかの型に分類することができる。しかしそれは明確な数値的基準があるわけではなく、実際は得られた観測結果を人の目で確認することにより変光星型が推察されることも多い。また、最近は大規模プロジェクトによって大量の新変光星が発見されるようになり、機械的な判別器の必要性が高まっている。

本研究では統計上の根拠がある分類を行うことと、多量の分類を可能にすることを目的に、フィッシャーの線形判別に基づく、機械学習による変光星の分類を試みた。変光星の光度曲線および、光度曲線をフーリエ変換して得たパワースペクトルから抜き出した特徴量を使用して判別器の作成を行う。

しかし特徴量を増やすほど判別精度が上がるとは限らない。そのため抜き出した 13 個の特徴量の組み合わせを変えて全通り試し、最も誤判別の少ない組み合わせを調べるという手法をとった。判別の対象としたものは セファイド型、ミラ型、食変光星、周期の無い人工ノイズデータ、かんむり座 R 型の計 5 種類であり、既存の分類済みデータ群を正しい分類として学習させて判別器を作成し、その性能評価を行った。性能評価には交差検定を用いた。かんむり座 R 型は入手できたデータの数が少なかったので、判別の対象に含む場合と含まない場合でそれぞれ 10 分割交差検定と 1 個抜き交差検定を使用した。

その結果、光度の平均や、標準偏差など、13 個のうち 8 つの重要な特徴量に絞り込むことができた。これらの特徴量を使ったモデルによって、正答率 98.95%の判別器の構築に成功した。





第1章　研究の背景 〜変光星とその分類〜

1.1　変光星とサーベイプロジェクト

恒星の中には明るさが変化するものがあり、それらは変光星と呼ばれる。これらが変光する理由は様々で、膨張や収縮を繰り返すことにより明るさを変える脈動変光星と呼ばれるものや、連星系による周期運動で見かけの明るさが変わる食連星、突発的な爆発で明るさを変える爆発型変光星など、複数の種類が存在する。

変光星の特性を利用して恒星までの距離や、内部構造を求めることも可能である。脈動変光星に属するセファイド変光星は、周期が長い天体ほど光度が高くなるという「周期–光度関係」があるため、周期と絶対等級が求まれば、恒星や近傍銀河までの距離がわかる。脈動変光星は膨張や収縮を繰り返すため、内部構造を光度周期から知ることも可能である。この手法は星振学と呼ばれる。[1]

変光星の明るさの変化を簡単に確認する際には、光度曲線がよく使われる。光度曲線は明るさの時間変動を表したグラフである。図 1.1 は脈動変光星の 1 種であるミラ型変光星の光度曲線である。基本的に横軸が時間スケール、縦軸が等級で表される。この光度曲線から、この天体が約 100 日周期で振幅 1.2 等の変光をしていることがわかる。なお、光度曲線の縦軸の単位が等級である時は、縦軸は明るいほど上へプロットされて見えるよう、反転されていることが多い。

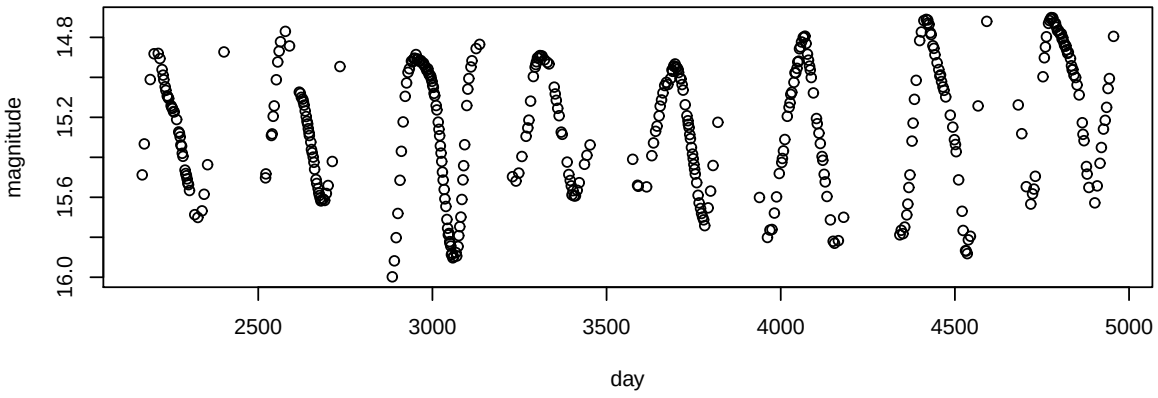


図 1.1: 脈動変光星の一種であるミラ型変光星の光度曲線。

上述のように、変光星は天文学の様々な場面で有用な情報を与えるため、新しい変光星の探索はこれまで熱心に行われてきた。近年、Sloan Digital Sky Survey (SDSS) に代表されるような、大規模サーベイ観測が

盛んに行われている。[2] 変光星の探索に関しても、重力レンズ探査を主な目的とした Optical Gravitational Lensing Experiment (OGLE) など、同じ天域を複数回観測し天体の明るさの時間変動をサーベイ的に観測するプロジェクトが多数進行している。[3] そのようなサーベイ観測によって、多数の新しい変光星が発見されている。例えば OGLE II では変光星でない恒星も含めて、光度曲線が得られている天体の数は 4 千万個に達する。

1.2 変光星の型の自動判別

前節で述べたように、現代では大量の天体データを扱うことができるが、10^7 個にもなる大量データを人の手で 1 つ 1 つ確認することは現実的でなくなっている。例えば、ある特定の型の変光星のデータが必要な場合でも、大量のデータからその型の天体を同定できなければ、データを十分に活用できているとは言えない。

このような問題が起こらないようにするため、海外の大規模プロジェクトでは独自に変光星型の自動分類システムを作っているところも多い。このようなシステムには一般的にベイズモデルや機械学習が使用される。近年、これらは広い分野で使われており、研究が盛んに行われている。

ウェブサービスを手がける Google は 2015 年 11 月にオープンソースの機械学習ライブラリ TensorFlow を公開した。これは元々 Google が自社で開発および使用していたものである。検索システムや自動翻訳、メール分類など、積極的に人工知能の開発を行なってきた企業がライブラリを公開したことで、機械学習の研究がさらに加速すると期待される。

変光星の自動分類には、例えば NASA が 1995 年から公開しているソフトウェア AutoClass が使われることがある。[5] これはベイズモデルに基づいており、光度曲線やそのフーリエ変換によって得られるパワースペクトルから天体の特徴量を入力することで、変光星の型と各サンプルの型の確率が出力として得られる。大規模プロジェクトでは独自の自動判別システムを構築することもあるが、その他の多くのプロジェクトでは AutoClass のような既存の、汎用性の高いソフトウェアを応用することが多いのが現状である。

1.3 本研究の目的

日本国内ではこれまでサーベイ型の大規模な変光星探索プロジェクトが少なかったので、変光星の自動型判別の研究も進んでいない。一方、最近は国内でも大規模な変光星サーベイ観測が行われるようになっている。板らはミラ型星の周期ー光度関係を調べるため、OGLE と南アフリカ IRSF 望遠鏡の SIRIUS プロジェクトによって、大小マゼラン雲をサーベイ的に観測した。その結果、10000 個以上のミラ型星の周期を決定した。[6] 同様にミラ型星については、最近、松永らが東京大学木曽観測所のシュミット望遠鏡を用いて、銀河円盤内のミラ型を探索する KISO-GP プロジェクトが進行している。[7] また、赤外線を使うことでダストの濃い銀河中心方向のセファイドを探索する、岡山観測所広視野赤外線カメラを用いた銀河面のサーベイプロジェクトも現在進行中である。[8] これらのプロジェクトは大量の光度曲線データを生成しているが、主な目的が大振幅かつ周期的な脈動星であるため、光度曲線の変動振幅のみからそれらを容易に抽出することができる。一方で主目的以外の変光星については、それらを自動判別するシステムが存在しないため、データを十分活用できていない。そのため、変光星の自動分類システムの開発は急務であると言え、また、国内にそのようなシステムの開発を行う研究グループを形成することは重要である。

天文分野ではベイズモデルや機械学習による判別問題についての研究は少ないが、他分野では国内でもそのような研究が進んでいる領域は多い。例えば、桑谷らは、過去の津波によって堆積した物質を非津波堆積物と区別するために、堆積物中の元素料を使用した線形判別分析を行い、高い成果を挙げている。[4] この中で、彼らは利用可能な全ての元素量を利用するよりも、特定の組み合わせを用いた方が判別率が上がることを示した。このように、重要な少数のパラメータをデータから選択し、判別モデルを構築する手法は情報科学などの分野で最近注目されており、同様の手法を変光星の型判別の問題に応用することは新しく、価値が高い。

本研究は、進行中の国内サーベイプロジェクトのデータ活用に寄与できる、変光星型の自動判別器を作成することを目標に行った。この卒業研究では判別モデルの中でも基本的な線形判別を基にした、分類済みデータからのモデルの作成とテストを目指す。

本論文の構成は次に従う。まず第 2 章にて本研究で使用する線形判別の理論について述べる。続く第 3 章で使用するデータと実験の内容を説明し、第 4 章で結果について述べ、第 5 章で結果の考察を行う。最後に第 6 章で本研究で得られた知見についてまとめる。

第2章　判別手法およびその理論

複数のサンプルがあって、それらはあらかじめ数種類に分類されているとする。その種族が明らかなデータ群から分類基準を探し出し、種族が不明なデータの分類を行う。本研究ではこのような手法を用いて変光星の型判別を行った。この章では本研究で使用する判別手法およびその理論を説明する [9]。

2.1　線形判別

分類先のパターンや種類のことをクラスと呼ぶ。分類とは、各サンプルに対してサンプルを特徴づける変数ベクトル $(x_1, x_2, \ldots, x_n)$ を離散クラス C_k の 1 つに割り当てることである。今回のような線形識別関数を使った方法では、入力は必ず 1 つのクラスに割り当てられる。分類先のクラスが 2 つや 0 といったことはない。この、入力空間において各クラスが占める領域を決定領域と呼び、その境界を決定境界ないし決定面と呼ぶ。本研究で用いた線形識別モデルは D 次元入力空間を $D-1$ 次元の超平面決定境界で分離するものである。

2.1.1　2 クラス線形判別

一般的な線形判別そのものについて簡単に述べたい。最初に 2 クラスの場合を考えて、後に $K > 2$ クラスへ拡張してゆく。線形判別問題は、入力ベクトルの線形変換として以下のような簡単な式で表現できる。

$$y(\boldsymbol{x}) = w^T \boldsymbol{x} + w_0 \tag{2.1}$$

ここで w を重みベクトル、w_0 をバイアスパラメータと呼ぶ。バイアスの符号を反転した値はしきい値パラメータとも呼ばれる。2 クラスの場合だと、入力ベクトル x は $y(x) \geq 0$ でクラス C_1 へ、それ以外だとクラス C_2 へ割り当てられる。そのため、決定境界は $y(\boldsymbol{x}) = 0$ の超平面に相当する。決定境界の位置は定数項のバイアスパラメータが決定する。一方重みベクトル $\boldsymbol{w}$ は決定面の向きを決める。この w の役割を確認するため、決定面上の任意の 2 点 $\boldsymbol{x}_A$ と $\boldsymbol{x}_B$ を設定する。これらは決定面上の点であるため、$y(\boldsymbol{x}_A) = y(\boldsymbol{x}_B) = 0$ である。これより、$\boldsymbol{w}^{\mathrm{T}}(\boldsymbol{x}_{\mathrm{A}} - \boldsymbol{x}_{\mathrm{B}}) = 0$ が成り立つ。$\boldsymbol{x}_A - \boldsymbol{x}_B$ は決定面上のベクトルであるため、$\boldsymbol{w}$ は決定面上の全ての点に直交する。よって、$\boldsymbol{w}$ は決定面の方向を決定する。

また、$y(\boldsymbol{x})$ の値は点 $\boldsymbol{x}$ と決定面の直交距離にあたる。すなわち、$y(\boldsymbol{x})$ の絶対値が大きいほど、$\boldsymbol{x}$ は決定面から離れたところにある。これを確かめてみる。任意の点 $\boldsymbol{x}$ と $\boldsymbol{x}$ の決定面上への直交射影 $\boldsymbol{x}_\perp$ を考える。$\boldsymbol{x}_\perp$ が決定面上の点であることと、$\boldsymbol{w}$ が決定面に垂直であることを利用すると、$\boldsymbol{x}$ は定数 r を用いて以下のように書ける。

$$\boldsymbol{x} = \boldsymbol{x}_\perp + r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \tag{2.2}$$

ここで $\|\boldsymbol{w}\|$ はノルムであり、$\boldsymbol{w}$ の大きさを表す。この式に $\boldsymbol{w}^{\mathrm{T}}$ をかけて w_0 を加えると $y(\boldsymbol{x}) = y(\boldsymbol{x}_\perp) + r\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}/\|\boldsymbol{w}\|$ という式になる。$\boldsymbol{x}_\perp$ が決定面上の点であるため $y(\boldsymbol{x}_\perp) = 0$ であることを利用して整理すると、決定面から点 $\boldsymbol{x}$ への直交距離 r は

$$r = \frac{y(\boldsymbol{x})}{\|\boldsymbol{w}\|} \tag{2.3}$$

である。これらの関係を図 2.1 に示す。

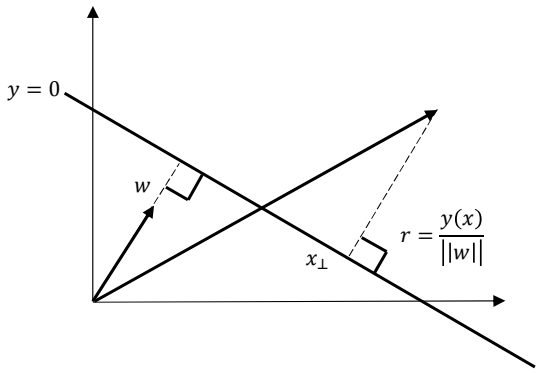


図 2.1: 2 クラス判別での決定面や $y(x)$ の対応。

2.1.2 多クラス線形判別

ここで、先の 2 クラスモデルを拡張した多クラスモデルの話題に入ろうと思う。3 クラス以上の他クラス判別では、識別関数を 1 つしか設定しなかった 2 クラスの場合とは異なり、必然的に複数の関数を設定する必要がある。その際には、K 個の識別関数を設定し、入力ベクトルに対する大小を比較することで他クラスの分類を簡単にしかも明確に行うことができる。

$$y_k(\boldsymbol{x}) = \boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{x} + w_{k0} \tag{2.4}$$

つまりは、関数 $y_1(\boldsymbol{x}), y_2(\boldsymbol{x}), \ldots, y_k(\boldsymbol{x}), \ldots, y_K(\boldsymbol{x})$ を用意して、この値が最も大きいクラス C_k に割り当てるということだ。この方法では 2 クラス $C_k - C_j$ 間の決定境界は $y_k(\boldsymbol{x}) = y_j(\boldsymbol{x})$ で与えられることから、

$$(\boldsymbol{w}_k - \boldsymbol{w}_j)^{\mathrm{T}}\boldsymbol{x} + (w_{k0} - w_{j0}) = 0 \tag{2.5}$$

で決定境界が示される。

この各関数の大小比較を行う方法では、各決定領域は凸領域になる。これは図 (2.2) のように決定領域 R_k 内にある二点 $\boldsymbol{x}_{\mathrm{A}}$ および $\boldsymbol{x}_{\mathrm{B}}$ と、$\boldsymbol{x}_{\mathrm{A}}$ と $\boldsymbol{x}_{\mathrm{B}}$ を結ぶ線分上に存在するものの分類先が不明な任意の点 $\widehat{\boldsymbol{x}}$ を使って説明できる。点 $\widehat{\boldsymbol{x}}$ は $0 \leq \lambda \leq 1$ を使うと次のように書ける

$$\widehat{x} = \lambda\boldsymbol{x}_{\mathrm{A}} + (1 - \lambda)\boldsymbol{x}_{\mathrm{B}} \tag{2.6}$$

これに識別関数を作用させる。いまは線形な識別関数を考えているため、

$$y_k(\widehat{\boldsymbol{x}}) = \lambda y_k(\boldsymbol{x}_{\mathrm{A}}) + (1 - \lambda)y_k(\boldsymbol{x}_{\mathrm{B}}) \tag{2.7}$$

となる。同時に $\boldsymbol{x}_{\mathrm{A}}$ が R_k 内にあると考えているから、全ての $i \neq k$ に対して $y_k(\boldsymbol{x}_{\mathrm{A}}) > y_i(\boldsymbol{x}_{\mathrm{A}})$ が成り立つ。同じことが $\boldsymbol{x}_{\mathrm{B}}$ にも言えるため、$y_k(\widehat{\boldsymbol{x}}) > y_i(\widehat{\boldsymbol{x}})$ であるから、$\widehat{\boldsymbol{x}}$ は R_k 内にある。

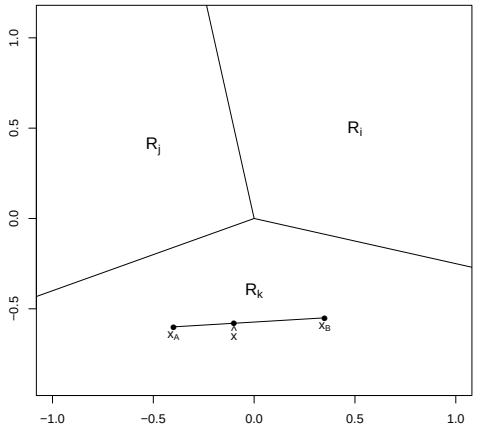


図 2.2: 凸領域になっている決定領域。

2.2 フィッシャーの線形判別

本研究で用いた判別モデルは、線形判別の中でもフィッシャーの線形判別と呼ばれるものであり、多次元入力ベクトル $\boldsymbol{x}$ を低次元に射影して判別を行うモデルである。先と同じように、2 クラスから他クラスへ拡張する形でその射影の方法に関して述べたい。

2.2.1 フィッシャーの線形判別 – 2 クラス

2 クラス判別の場合は、削減後の入力ベクトルは通常 1 次元になる。入力ベクトル $\boldsymbol{x}$ を以下のようにスカラー量 y に変換するとしよう。

$$y = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} \tag{2.8}$$

ここで前節と同様にしきい値 w_0 を設定すると、$y + w_0$ の正負で判別ができる。それはしかしクラス C_1 と C_2 を分離する適切な射影方向 $\boldsymbol{w}$ が選択されていなければ識別モデルとして使えない。ここでクラス C_k の点が N_k 個あるとしてクラスの平均ベクトル $\boldsymbol{m}_k$ は

$$\boldsymbol{m}_k = \frac{1}{N_k}\sum_{n\in C_k}\boldsymbol{x}_n \tag{2.9}$$

である。射影後の平均は以下になる。

$$m_k = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{m}_k \tag{2.10}$$

また、クラス C_k において射影後のサンプルのクラス内分散を以下で定義する。

$$s_k^2 = \sum_{n\in C_k}(y_n - m_k)^2 \tag{2.11}$$

そしてフィッシャーの線形判別モデルでは、射影方向は次の式を最大化するように選択される。

$$J(\boldsymbol{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} \tag{2.12}$$

式 (2.12) の分母は式 (2.11) で定義した各クラス内分散の総和である。一方で、分子は射影先でのクラス平均の差であり、この量をクラス間分散と定義する。フィッシャーの線形判別では、射影先でのクラス間分散を大きくすると共に、クラス内分散が小さくなるような解を選択する。式 (2.12) の右辺を $\boldsymbol{w}$ を含む形にするため、この式を (2.8) と (2.10)、(2.11) 使い、行列の形に書き直す。

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{B}}\boldsymbol{w}}{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{W}}\boldsymbol{w}} \tag{2.13}$$

ここで $\boldsymbol{S}_{\mathrm{B}}$ はクラス間分散行列 (between-class covariance matrix) であり、次の式で与えられる。

$$\boldsymbol{S}_{\mathrm{B}} = (\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^{\mathrm{T}} \tag{2.14}$$

また、$\boldsymbol{S}_{\mathrm{w}}$ はクラス内分散行列 (within-class covatiance matrix) であり、次の式で与えられる。

$$\boldsymbol{S}_{\mathrm{w}} = \sum_{n_1 \in C_1} (\boldsymbol{x}_n - \boldsymbol{m}_1)(\boldsymbol{x}_n - \boldsymbol{m}_1)^{\mathrm{T}} + \sum_{n_2 \in C_2} (\boldsymbol{x}_n - \boldsymbol{m}_2)(\boldsymbol{x}_n - \boldsymbol{m}_2)^{\mathrm{T}} \tag{2.15}$$

我々が求めたいのは $J(\boldsymbol{w})$ の値が最大になる $\boldsymbol{w}$ である。これを得るために (2.13) を $\boldsymbol{w}$ で微分する。行列の微分には次の公式

$$\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}) = (\boldsymbol{A} + \boldsymbol{A}^{\mathrm{T}})\boldsymbol{x}$$

を利用する。その結果、以下の関係が得られる。

$$(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{B}}\boldsymbol{w})\boldsymbol{S}_{\mathrm{W}}\boldsymbol{w} = (\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{W}}\boldsymbol{w})\boldsymbol{S}_{\mathrm{B}}\boldsymbol{w} \tag{2.16}$$

ここで $(\boldsymbol{m}_2 - \boldsymbol{m}_1)^{\mathrm{T}}\boldsymbol{w}$ は

$$y = y^{\mathrm{T}} = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{w}$$

であるため定数となる。これと (2.14) より $\boldsymbol{S}_{\mathrm{B}}\boldsymbol{w}$ は $(\boldsymbol{m}_2 - \boldsymbol{m}_1)$ と同じ向きを持つベクトルである。ここで $\boldsymbol{w}$ の射影方向を得るためにスカラー係数 $(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{B}}\boldsymbol{w})$ と $(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{S}_{\mathrm{W}}\boldsymbol{w})$ を無視して (2.16) の両辺に $\boldsymbol{S}_{\mathrm{W}}^{-1}$ をかけると次の関係が得られる。

$$\boldsymbol{w} \propto \boldsymbol{S}_{\mathrm{W}}^{-1}(\boldsymbol{m}_2 - \boldsymbol{m}_1) \tag{2.17}$$

(2.17) 式がフィッシャーの線形判別の式であり、これに従って射影の方向を決める。これにデータサンプルの確率密度などからしきい値 w_0 を定義することで $y(\boldsymbol{x}) \leq w_0$ の時はクラス C_1、それ以外はクラス C_2 などと判別を行える。

2.2.2 フィッシャーの線形判別 – 多クラス

フィッシャーの線形判別は $K > 2$ クラスに拡張することも可能である。ここでクラス数 K に対し入力空間の次元 D の方が大きいとする。次元 D の入力ベクトル $\boldsymbol{x}$ から次元 D' のベクトル $\boldsymbol{y}$ への射影を考えて

$$\boldsymbol{y} = \boldsymbol{W}^{\mathrm{T}}\boldsymbol{x} \tag{2.18}$$

とおく。ここではまだバイアスパラメータを定めていない。(2.15) の 2 クラスの場合のクラス内共分散を K クラスの場合に書き直して

$$\boldsymbol{S}_{\mathrm{W}} = \sum_{k=1}^{K} \boldsymbol{S}_k \tag{2.19}$$

ここで

$$\boldsymbol{S}_k = \sum_{n_1 \in C_k} (\boldsymbol{x}_n - \boldsymbol{m}_k)(\boldsymbol{x}_n - \boldsymbol{m}_k)^{\mathrm{T}} \tag{2.20}$$

$$\boldsymbol{m}_k = \frac{1}{N_k} \sum_{n \in C_k} \boldsymbol{x}_n \tag{2.21}$$

で、N_k はクラス C_k である入力の個数であり、$\boldsymbol{m}_k$ はクラス C_k の平均ベクトルである。続いてクラス間共分散行列の一般化にあたって以下の式の総共分散行列を考える。

$$\boldsymbol{S}_{\mathrm{T}} = \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{m})(\boldsymbol{x}_n - \boldsymbol{m})^{\mathrm{T}} \tag{2.22}$$

ここで、$\boldsymbol{m}$ は全データ集合の平均

$$\boldsymbol{m} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n = \frac{1}{N} \sum_{k=1}^{K} N_k \boldsymbol{m}_k \tag{2.23}$$

であり、$N = \sum_k N_k$ はデータの総数である。$\boldsymbol{S}_T$ はクラス内共分散行列 $\boldsymbol{S}_{\mathrm{W}}$ クラス間共分散行列に代わる行列 $\boldsymbol{S}_B$ に分解できて

$$\boldsymbol{S}_{\mathrm{T}} = \boldsymbol{S}_{\mathrm{W}} + \boldsymbol{S}_{\mathrm{B}} \tag{2.24}$$

ここで

$$\boldsymbol{S}_{\mathrm{B}} = \sum_{k=1}^{K} N_k (\boldsymbol{m}_k - \boldsymbol{m})(\boldsymbol{m}_k - \boldsymbol{m})^{\mathrm{T}} \tag{2.25}$$

である。以上は入力 $\boldsymbol{x}$ に対して考えたものであるが、射影後の D' 次元ベクトル $\boldsymbol{y}$ にでも同じことが言える。

$$\boldsymbol{S}_{\mathrm{W}} = \sum_{k=1}^{K} \sum_{n \in C_k} (\boldsymbol{y}_n - \boldsymbol{\mu}_k)(\boldsymbol{y}_n - \boldsymbol{\mu}_k)^{\mathrm{T}} \tag{2.26}$$

$$\boldsymbol{S}_{\mathrm{B}} = \sum_{k=1}^{K} N_k (\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^{\mathrm{T}} \tag{2.27}$$

ここで

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n \in C_k} \boldsymbol{y}_n, \quad \boldsymbol{\mu} = \frac{1}{N} \sum_{k=1}^{K} N_k \boldsymbol{\mu}_k \tag{2.28}$$

である。ここで 2 クラスの場合と同様にクラス間共分散が大きいほど大きく、クラス内共分散が大きいほど小さくなるスカラーを設けて、射影後のそれが最大となるような $\boldsymbol{W}$ を決定する。

2.3 人工データを用いた判別実験

フィッシャーの線形判別モデルの実用性を確認するために、人工的に変光星のデータを作成し、判別のテストを行った。0 から 999 までの 1000 個の整数 t_i ($i = 1 - 1000$) を、現実の日数に相当するものとして使用する。また、t_i が 1 刻みの整数であることからナイキスト周波数は 0.5 cycle/d であるため、人工データの周波数 f は $f < 0.5$ cycle/d である必要がある。変数 t_i と周波数 f および振幅 A から成る正弦波を作成し、さらに平均が 0 の正規分布に従う乱数をノイズとして加えたものを人工光度曲線 g' とした。すなわち、

$$g'(t_i) = A \sin(f t_i) + \mathcal{N}(0, \sigma) \ (i = 1 - 1000) \tag{2.29}$$

ただし、$\mathcal{N}(0,\sigma)$ は平均 0、標準偏差 (σ) の正規乱数を表す。また、地上からの天体観測では、天候などの理由から等間隔のデータをとるのは難しい。そのため、各人工光度曲線が持つ 1000 日分の値のうち半分に欠損値として 0 を代入して、サンプル用の信号 g とした。信号 g の例をいくつか図 2.3 に示す。これをフーリエ変換すると図 2.4 のようなパワースペクトルが得られる。ここで、パワースペクトルは周波数 0 からナイキスト周波数 0.5 cycle/day までを等間隔に 500 点パワーを推定した。

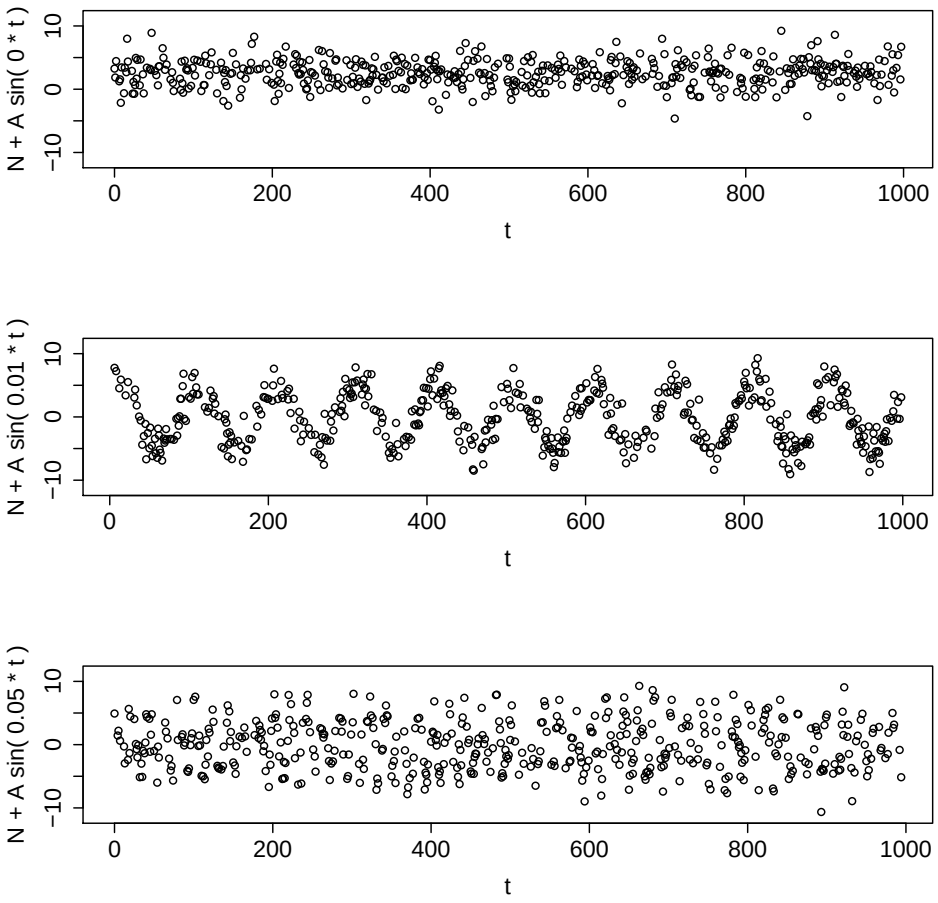


図 2.3: ノイズの乗った異なる周波数を持つ正弦波の図。振幅 5 の波に標準偏差 2 の正規分布に従うノイズを加えた。t_i の単位を日とすると周波数は上から順に 0 cycle/d、0.01 cycle/d、0.05 cycle/d である。

図 2.4 では変換元の波の周波数によりパワースペクトルのピーク位置に違いが見られる。例えば中段の、周波数 0.01 cycle/d のスペクトルは横軸が 0.01 の位置にピークが立ち、下段の 0.05 cycle/d のスペクトルは横軸が 0.05 の位置にピークが立っている。また周波数が 0 cycle/d、すなわち周期変動していないものはピークが立たない。これらの違いを利用して判別を行いたい。

判別に使用する変数には、パワースペクトルのピーク位置 $\nu_{\max}$ とパワースペクトルの最大値 $P_{\max}$ が適切であろう。これならば周期変動があるものと無いもの、そして周期変動しているものの中でも、その周期の違いで分けられると期待できる。

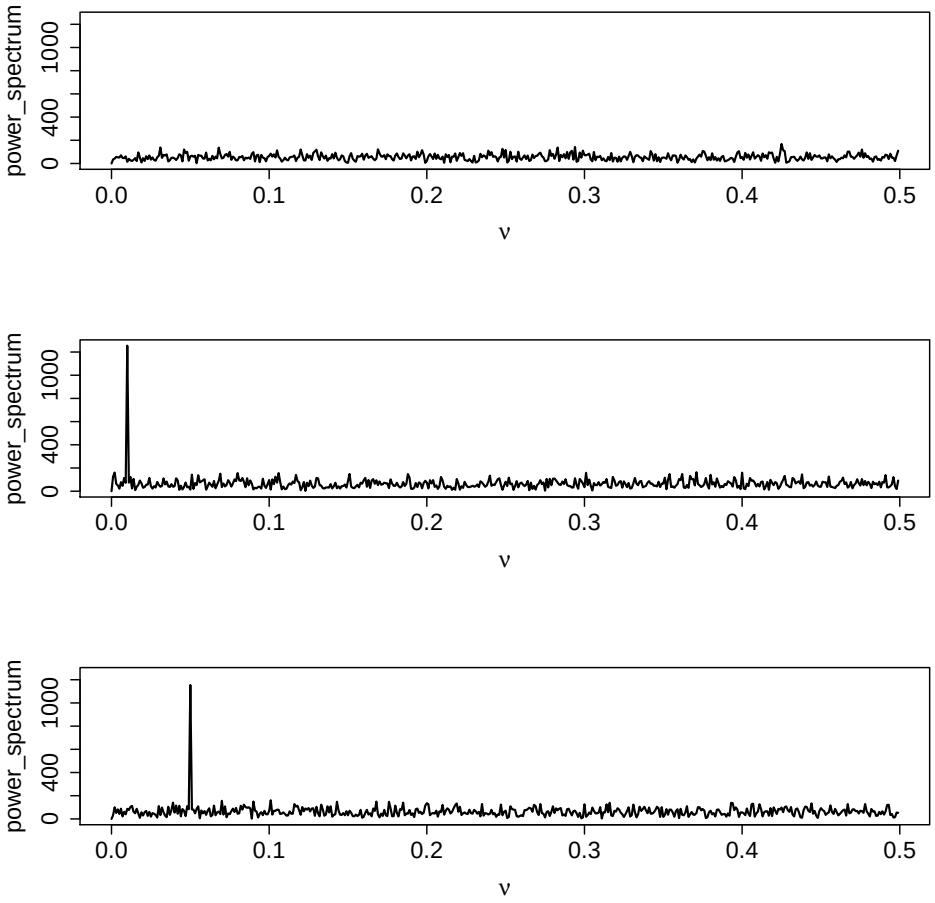


図 2.4: 図 2.3 の波を FFT(高速フーリエ変換) にかけたもの。横軸が元の波の周波数に対応する。下 2 つは強いピークが見られるが、1 番上の周波数 0 のものにはピークが見られない。

この判別をを行うにあたって、周波数 $f = 0, 0.01, 0.05$ cycle/d の三種のデータをそれぞれ 1000 個ずつ用意し、うち半分の各 500 個ずつを学習用、残りの各 500 個ずつをテスト用データとした。すなわち、1500 個のデータから取得した判別式で、別の 1500 個のデータの判別を行う。正弦波の振幅 A が 5 であるのに対して、ノイズの標準偏差は 2 に設定した。むろん、全データに対して異なるノイズを加えて、欠損値 0 を代入する t もランダムにしてあるため、それぞれのデータは異なるものにしている。また、実際の計算には R 言語のライブラリ MASS に含まれる lda 関数を用いた。

このようにして、パワースペクトルのピーク周波数 $\nu_{\max}$ と最大値 $P_{\max}$ の 2 次元入力ベクトルを元データの周波数ごとに分けた 3 クラスへ分離する判別を行った。その結果が図 2.5 と表 2.1 である。

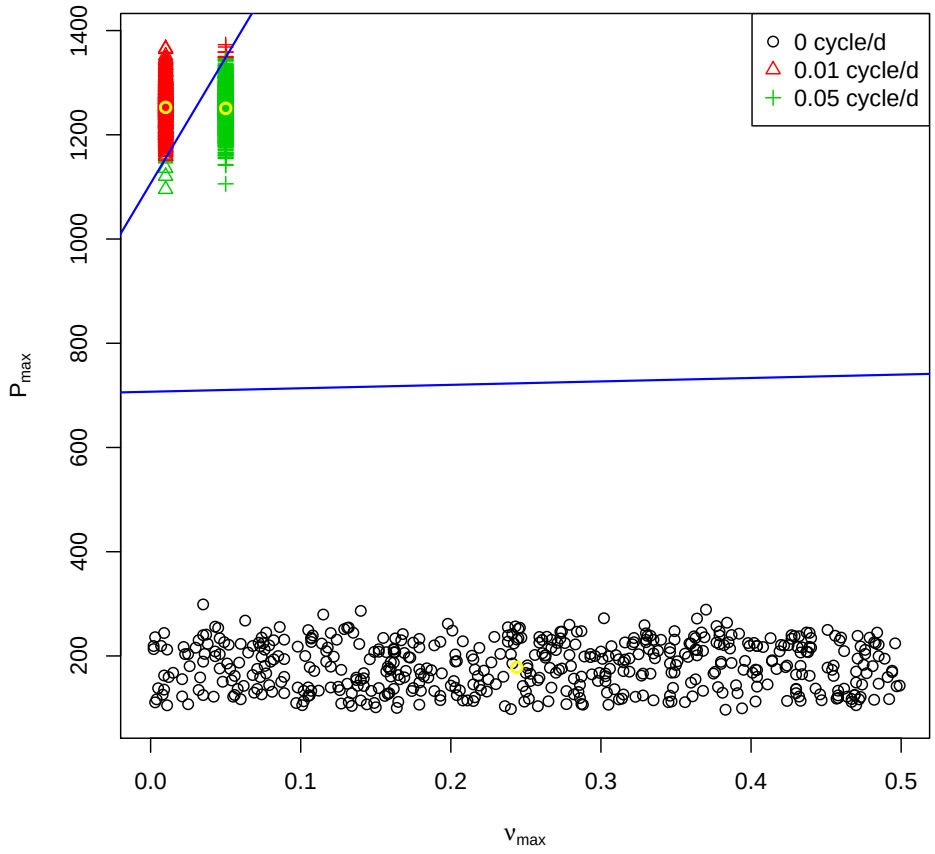


図 2.5: 3 クラスの人工データに対してフィッシャーの線形判別を行った図。青線は決定境界である。 各点の形は元データの種類を、色は判別後のクラスを表している。

黄色の点が学習データの各クラスの平均の値であり、周波数 0.01 cycle/d のクラスと周波数 0.05 cycle/d のクラスとのクラス間分散を大きくしようとして決定境界が傾いてしまっている。この場合、周期を持つデータ間では $P_{\max}$ は判別に必要な変数にはならないはずであり、正しくない解が得られる例である。また、図 2.6 では傾きがさらに大きくなっている場合を示している。このように、学習データが異なると異なる判

		推定されたクラス		
		0 cycle/d	0.01 cycle/d	0.05 cycle/d
正しいクラス	0 cycle/d	500	0	0
	0.01 cycle/d	0	497	3
	0.05 cycle/d	0	4	496

表 2.1: 図 2.5 における判別結果。3 個の 0.01 cycle/d の波が 0.05 cycle/d の波であると誤判別されている。

別結果が得られることがある。

次に振幅が 5 に対してノイズの偏差を 10 にして判別を行った。結果を以下の図 2.7 と表 2.2 に示す。

		推定されたクラス		
		0 cycle/d	0.01 cycle/d	0.05 cycle/d
正しいクラス	0 cycle/d	500	0	0
	0.01 cycle/d	3	454	43
	0.05 cycle/d	2	41	457

表 2.2: 図 2.7 における分類結果

こちらはノイズを大きくしているため $P_{\max}$ が広い範囲に分布している。図 2.5 に比べて決定境界の傾きは小さくなっているものの $P_{\max}$ の範囲が広いため誤判別の数は増えている。

最後に、図 2.7 のデータから $\log \nu_{\max}$ と $\log P_{\max}$ を入力ベクトルにとり直して再び判別を行った。その結果が図 2.8 と表 2.3 である。

		推定されたクラス		
		0 cycle/d	0.01 cycle/d	0.05 cycle/d
正しいクラス	0 cycle/d	498	2	0
	0.01 cycle/d	1	499	0
	0.05 cycle/d	2	0	498

表 2.3: 図 2.8 における分類結果

ここで対数をとったことに物理的意味はない。ただクラス間分散を見かけ上大きくするために変換している。しかしその結果、表 2.2 に比べて判別の精度は向上している。このように、同じ $\nu_{\max}$ や $P_{\max}$ を用いた判別であっても、対数を取るなどの変換によってクラス間分散が変化することで、判別性能が変化する。以上の人工データを用いた実験によって、フィッシャーの線形判別が正しく動作することと、その挙動がいかなるものかを確認することができた。次章では実際の天体データにこの方法を応用する。

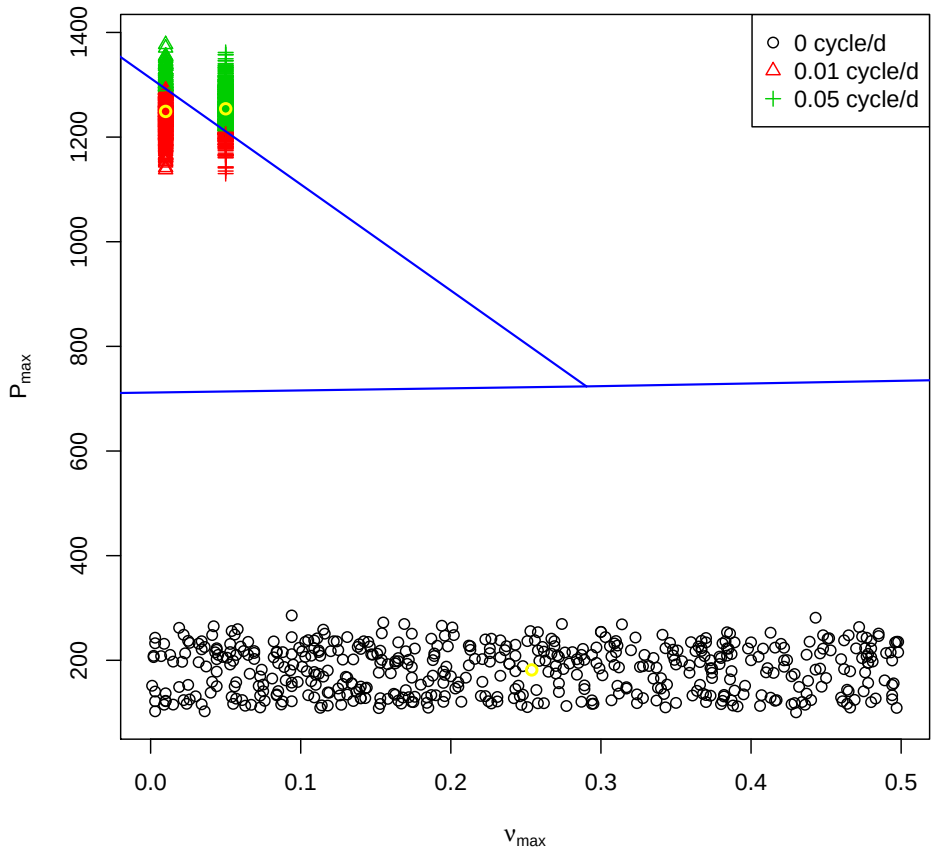


図 2.6: 図 2.5 と設定は同じで乱数を変えて判別したもの。これは誤判別が激しい部類

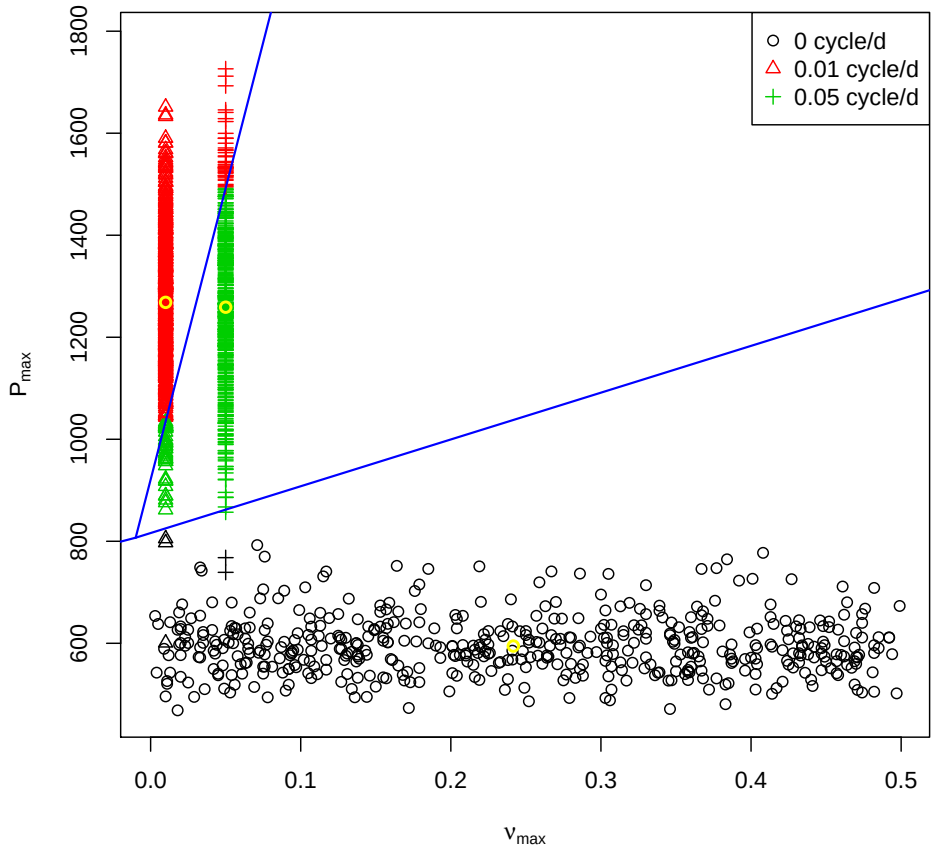


図 2.7: ノイズの偏差を振幅の 2 倍に設定して判別を行った結果

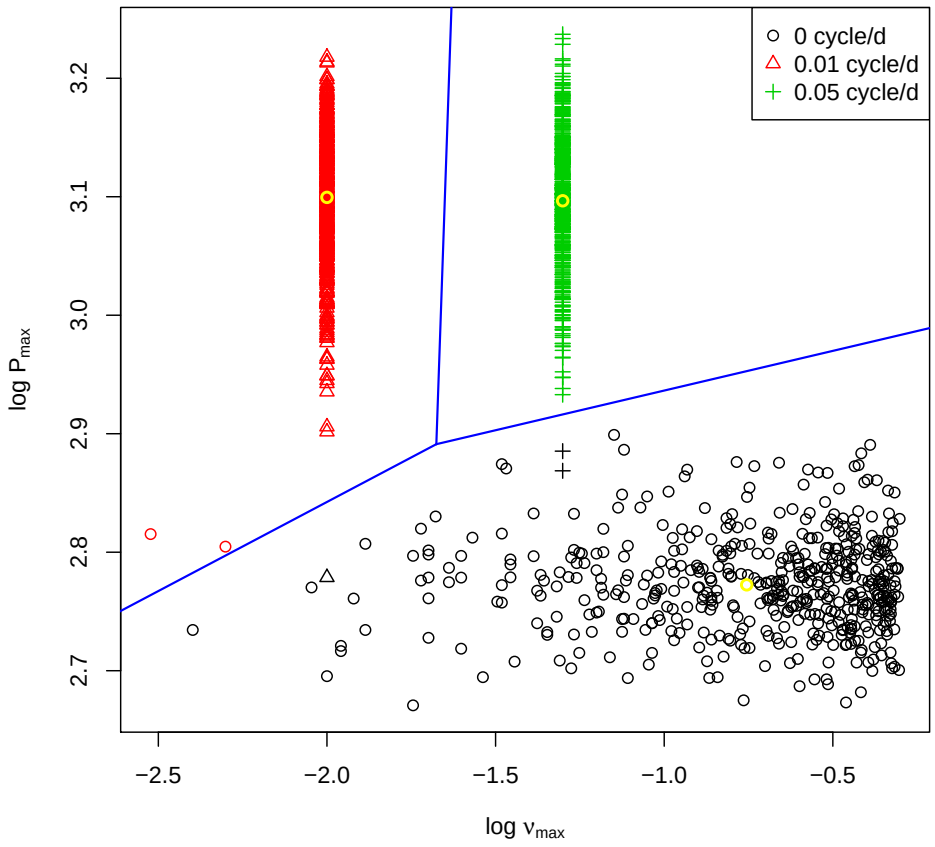


図 2.8: 図 2.7 のデータから各軸の常用対数をとって再判別したもの

第3章　データ解析

本研究で取り扱うデータの説明や、光度曲線から抽出したパラメータ等の説明をする。

3.1　判別に用いたデータ

判別に用いたデータサンプル群は OGLE(Optical Gravitational Lensing Experiment) のものを使用した。OGLE とはワルシャワ大を中心に重力レンズによるによる暗黒物質探査を行っているプロジェクトである。OGLE は重力レンズ現象の探査に伴って観測された光度曲線データを公開しており、独自の基準により観測された変光星の型を分類している。本研究では、この OGLE の分類を正しい分類と仮定してデータ学習を行い、判別器を作成した。

また、OGLE は銀河バルジ方向、大マゼラン雲、小マゼラン雲のそれぞれの領域ごとにデータを公開している。本研究では、観測天体が多いことや距離が等しいため見かけの等級が絶対等級に相当することを考慮して、判別用の天体は大マゼラン雲にあるもののみを用いた。

天体の観測は通常 1 日 1 回であるが OGLE の公開データの中には、一日に複数回の観測を行なっているものもある。したがって、ナイキスト周波数は厳密には 0.5 cycle/day ではなく、より高い周波数の信号を検出できる可能性があるが、一方で、ナイキスト周波数でのパワースペクトルの折り返しも見られる。図 3.1 はパワースペクトルのピークが 0.5 cycle/day よりも高い場合を示している。上から 3 番目と 4 番目の図はそれぞれナイキスト周波数を 0.5 cycle/day と 1.0 cycle/day に設定して得た周期で光度曲線を折りたたんだものである。この 2 つを比較するとナイキスト周波数 1.0 cycle/day で得た周期で折りたたんだものの方がより周期的な光度曲線を描いているため、パワースペクトルの真のピークは 0.59 cycle/day 付近に立っていることが確認できる。このような場合があることを考慮して、特徴量の抜き出しにはナイキスト周波数 0.5 cycle/day と 1.0 cycle/day 両方のパワースペクトルを使用した。

3.2　判別対象の変光星型

本研究では 4 種類の変光星型と人工ノイズデータの計 5 種類を判別の対象とした。人工ノイズデータは複数の観測データの時間情報を複製し、平均 15.2, 標準偏差 0.6 で発生させた正規乱数を等級として代入して作成した。この平均と標準偏差は、今回使用したミラ型変光星の情報を参考にして設定したものである。使用した変光星を光度曲線やパワースペクトルとともに簡単に紹介する。

セファイド型変光星

膨張や収縮を繰り返す脈動変光星の一種。周期は数日から数十日程度。図 3.2 に具体的な光度曲線やパワースペクトルを載せている。

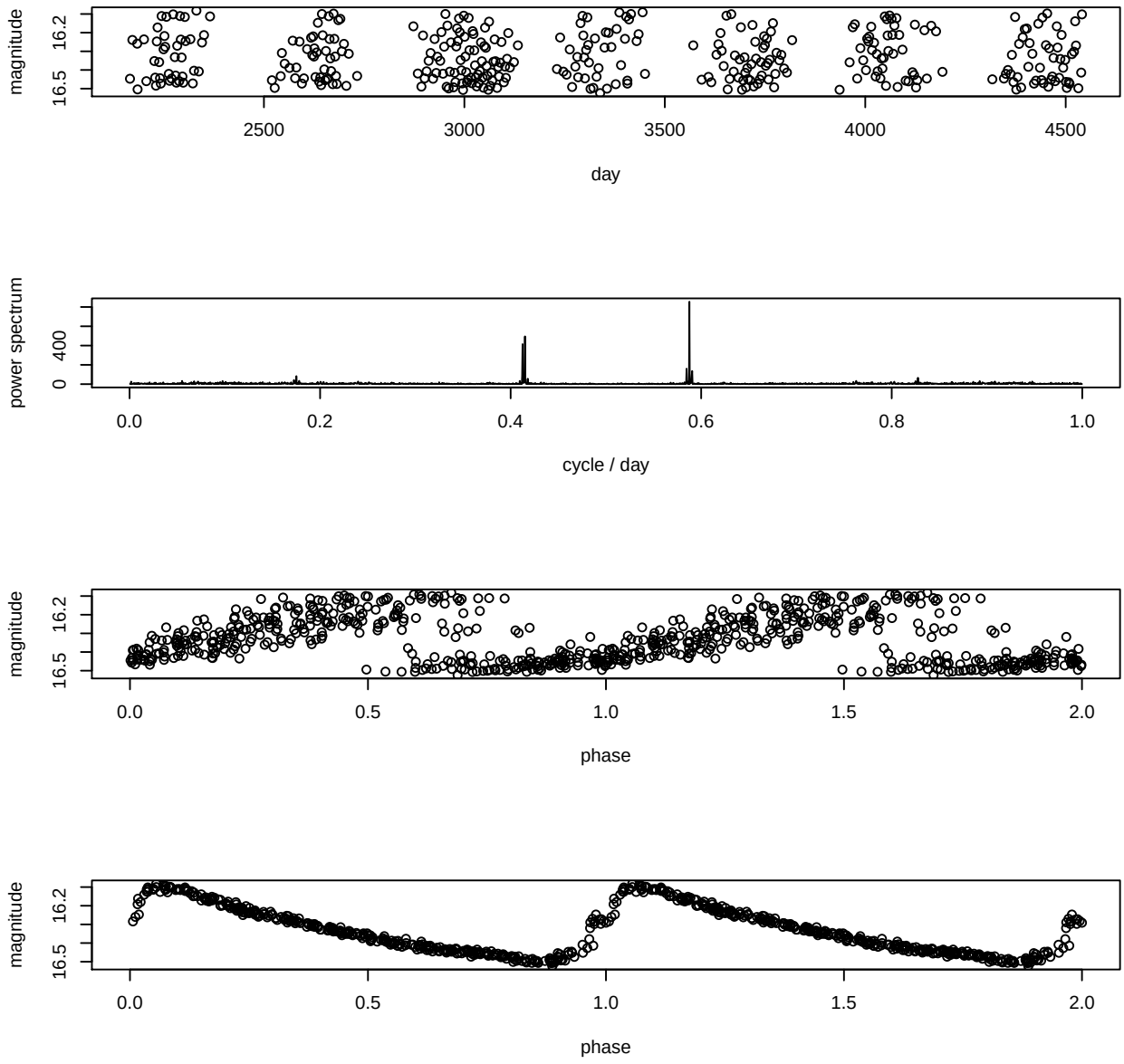


図 3.1: 0.5 cycle/day より高い真の周波数信号が捉えられている例。上から順に、光度曲線 (日-等級)、パワースペクトル、0.41 c/d と 0.59 c/d で畳んだ光度曲線。

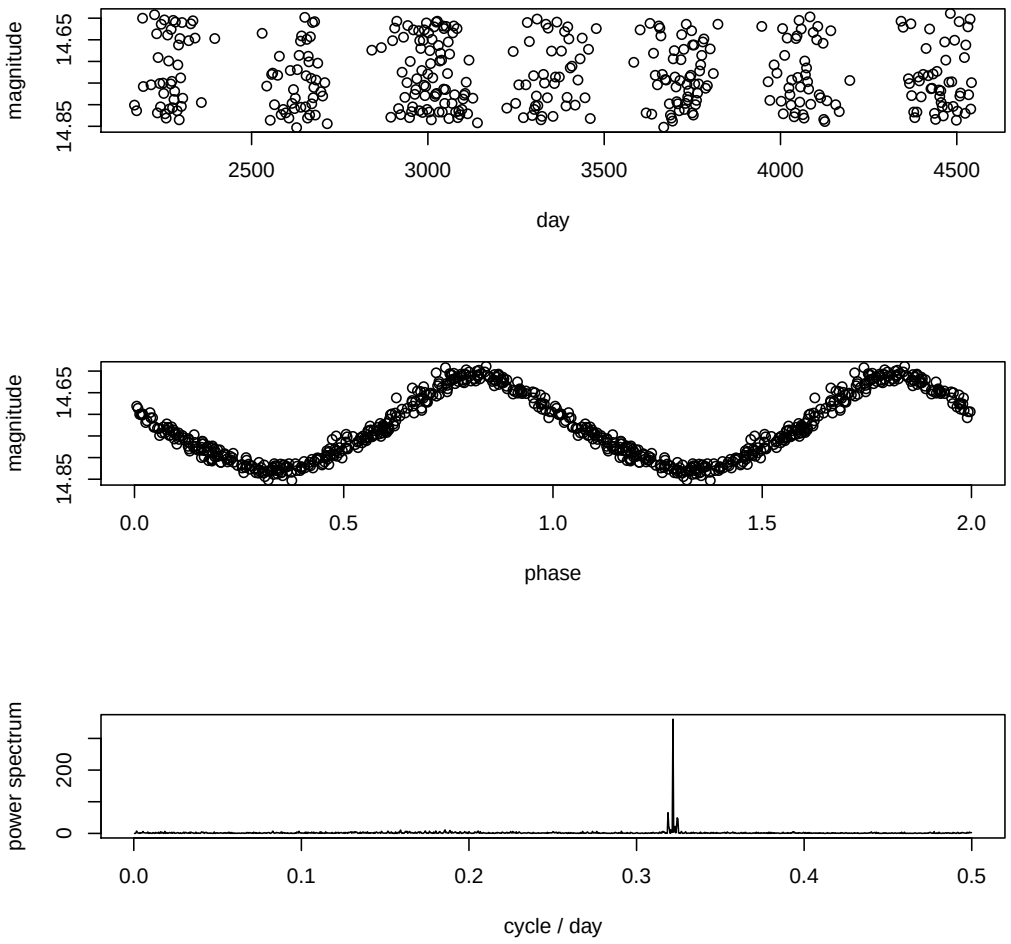


図 3.2: セファイド型の典型的なライトカーブ。中段の図は光度曲線を周期で折りたたんだものである。

ミラ型変光星

周期が数百日、振幅が数等級の脈動変光星。恒星の進化段階でも赤色巨星と呼ばれる年老いたものの一部がこのミラ型の変光を示す。図 3.3 にその一例を示す。

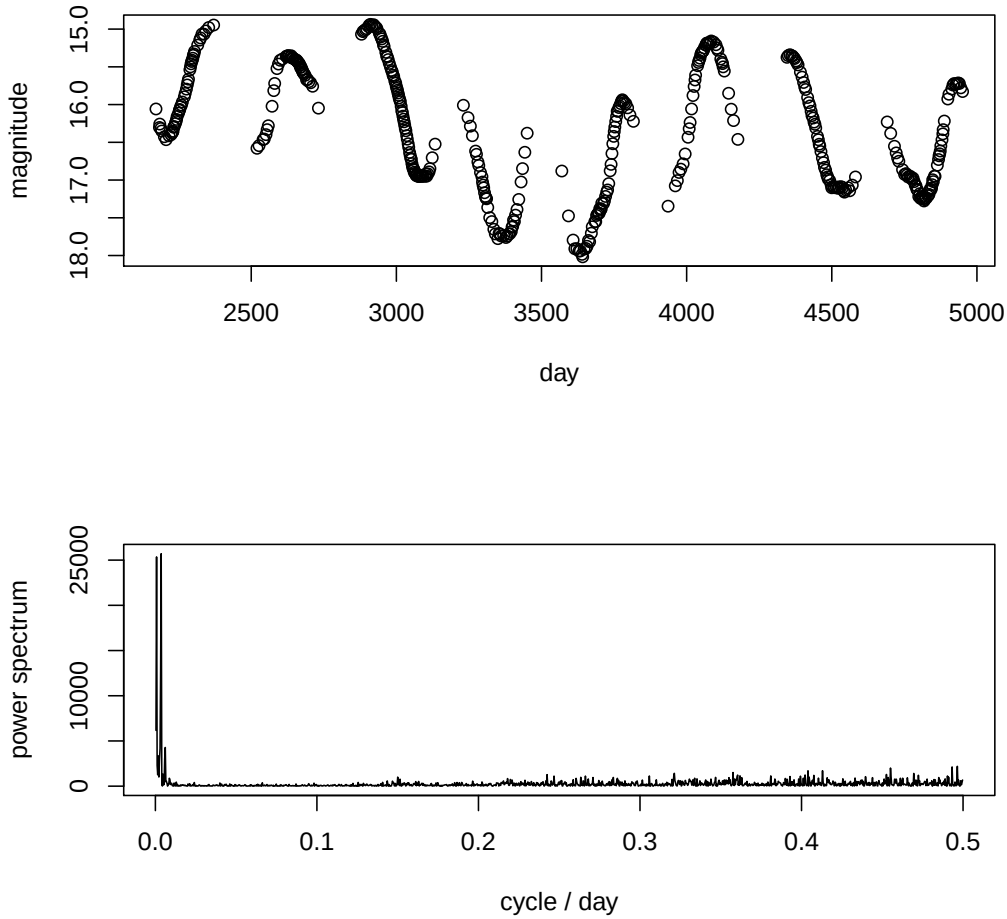


図 3.3: ミラ型の光度曲線およびパワースペクトルの例。

食変光星

2 つの星が一体となって周期運動をすることによって、地球からの見かけの明るさが変動しているように観測される星。相手の星を隠すことで一時的に暗くなり、周期的かつ偏りを持った光度曲線が特徴。図 3.4 および図 3.5 を参照。

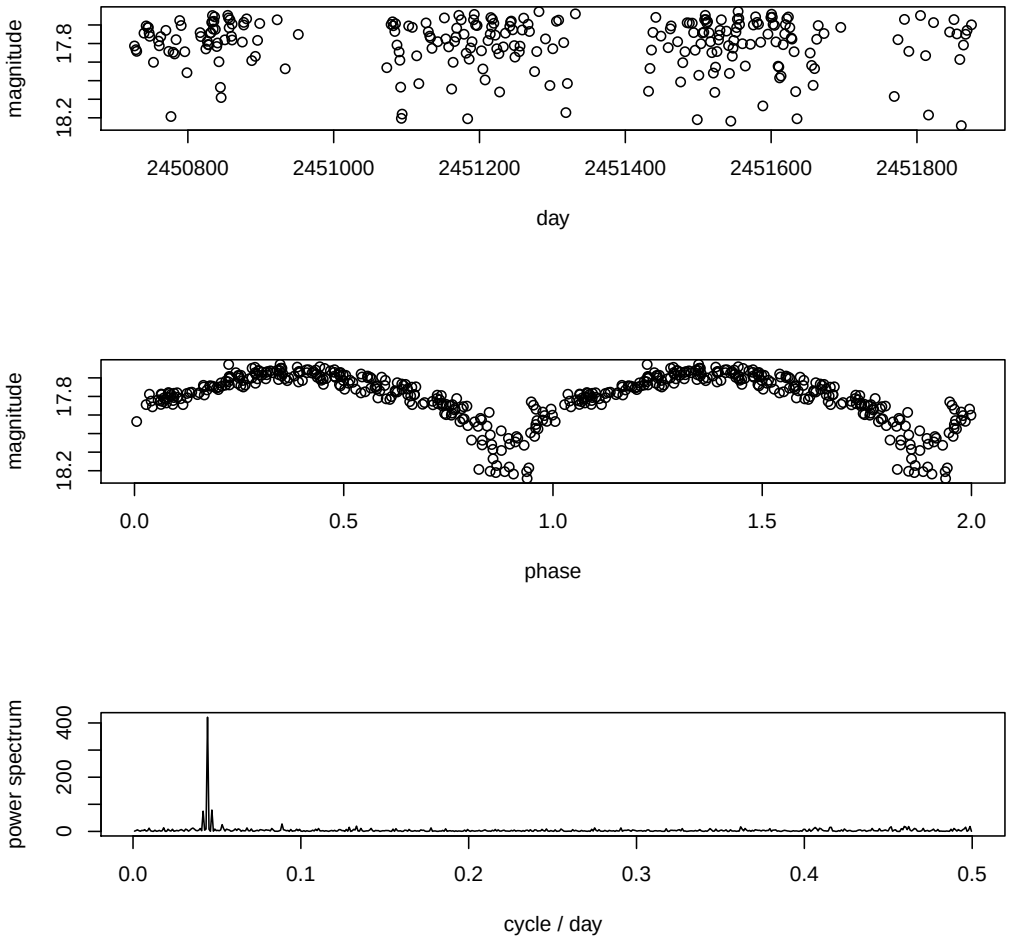


図 3.4: 食変光星の光度曲線およびパワースペクトルの例 1

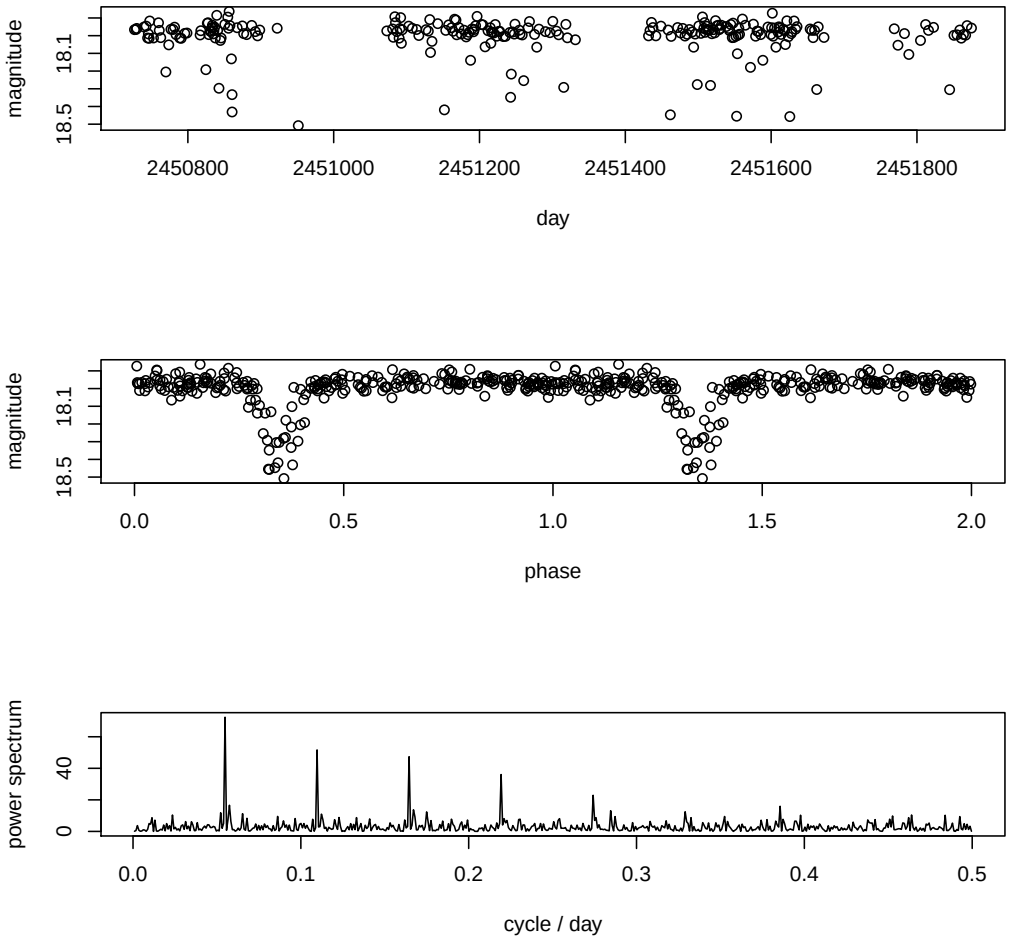


図 3.5: 食変光星の光度曲線およびパワースペクトルの例 2

かんむり座 R 型変光星

明るさは普段一定だが、突発的に減光する星。図 3.6 にその光度曲線やパワースペクトルを示す。

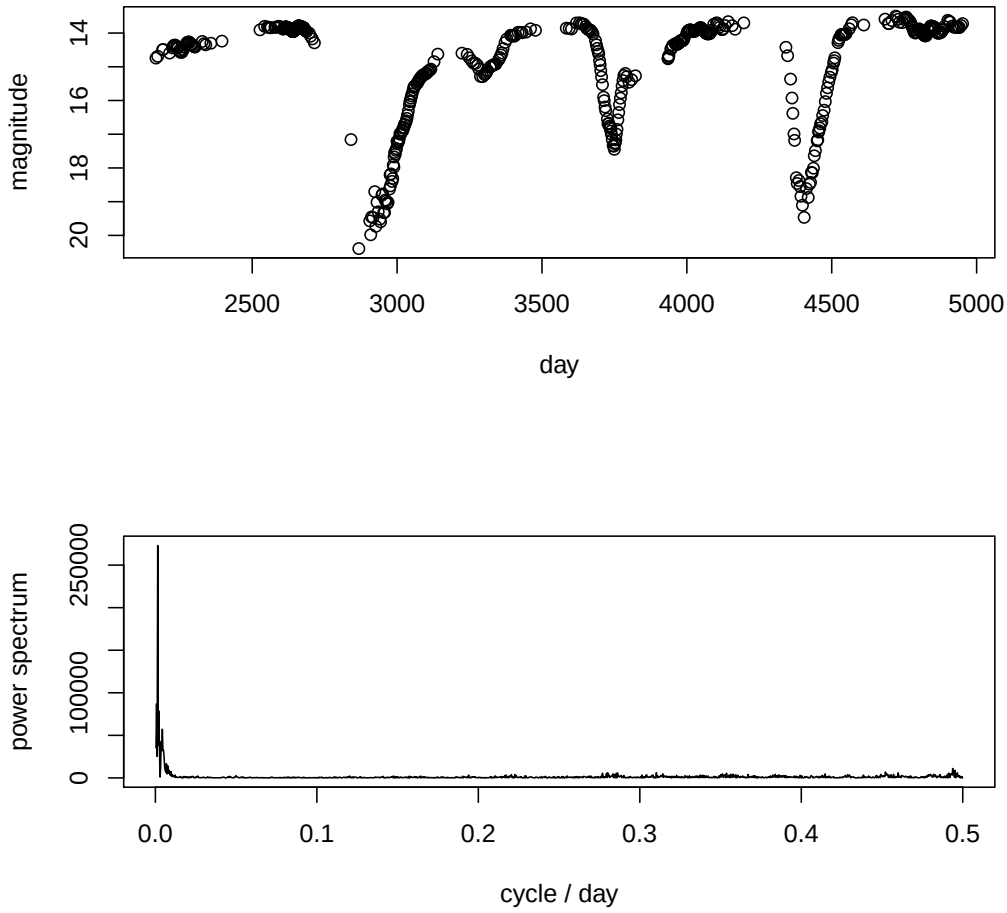


図 3.6: かんむり座 R 型変光星の光度曲線およびパワースペクトルの例

3.3 使用パラメータ

2 章で述べたように、判別に使用する特徴量はスカラー量の集合を用いる。変光星から抜き出す特徴量には、実際の光度曲線やパワースペクトルを参考に、各クラスの判別に適していると予想した量を設定した。

本研究では使用する特徴量の組み合わせを変えて全通り試し、交差検定で性能評価することで重要な特徴量を選択する。すなわち、次に紹介する特徴量は真に重要な特徴量の候補であるが、真に重要かはデータに選ばせる。その意味では、最初から候補を限定するよりも、なるべく多くを試すほうが良い。この理由か

ら、本研究では重要な特徴量をデータに選択させて、その後で選択された理由と意義を考察するというアプローチをとった。

3.3.1 光度曲線から抽出した特徴量

今回使用した光度曲線は 時間–等級 のデータである。光度曲線から抜き出した特徴量をまとめたのが次の表 3.1 である。

表 3.1: 光度曲線から取得した特徴量

特徴量	意味
L_{mean}	等級の平均値。
L_{sd}	等級の標準偏差。光度曲線の振幅に相当する。
$L_{\mathrm{max}} - L_{\mathrm{min}}$	等級の最大値と最小値の差。別途説明。
$L_{\mathrm{max}} - L_{\mathrm{mean}}$	等級の最大値と平均値の差。別途説明。
$(L_{\mathrm{max}} - L_{\mathrm{mean}})/(L_{\mathrm{max}} - L_{\mathrm{min}})$	上記の $L_{\mathrm{max}} - L_{\mathrm{mean}}$ を $L_{\mathrm{max}} - L_{\mathrm{min}}$ で正規化したもの。

表 3.1 中に用いた L_{max}, L_{min}, L_{mean}, L_{sd} の対応を図 3.7 に示す。

また、$L_{\mathrm{max}} - L_{\mathrm{min}}$ と $L_{\mathrm{max}} - L_{\mathrm{mean}}$ に関して、特徴量に設定した理由を補足する。

$L_{\mathrm{max}} - L_{\mathrm{min}}$

L_{sd} と同様光度曲線の振幅に相当する特徴量だが、例えば食連星のように、明るい時間が長く、暗い時間が短い場合は L_{sd} と異なる情報を持つことが期待できる。

$L_{\mathrm{max}} - L_{\mathrm{mean}}$

光度曲線の偏りを示す量と考えて設定した量である。図 3.2 のような正弦波に似た光度曲線ならば L_{mean} は $(L_{\mathrm{max}} + L_{\mathrm{min}})/2$ に近い値を取るが、食連星の場合だと明るい時期が長いため、これより小さい値をとる。これが判別に使えると期待して導入した。

3.3.2 パワースペクトルから抽出した特徴量

一方、パワースペクトルからは表 3.2 に示す特徴量を抽出した。

表 3.2 の P_{sd} はパワースペクトル $P(\nu)$ の標準偏差、P_5 は $0.45 < \nu < 0.5$ における $P(\nu)$ の最大値で、P_1 は $0.1 < \nu < 0.15$ における $P(\nu)$ の最大値である。これらと ν_{max}、P_{sd} の対応を図 3.8 および図 3.9 に示す。

また、P_{sd} と $(P_5 - P_1)/P_{\mathrm{sd}}$、$P_5/P_1$ に関して補足を行う。

P_{sd}

光度曲線に含まれる誤差は、パワースペクトルでは一様なノイズとなって現れる。このノイズを数値化するためパワーの標準偏差 P_{sd} を取得し、異なるノイズレベルの正規化に利用した。

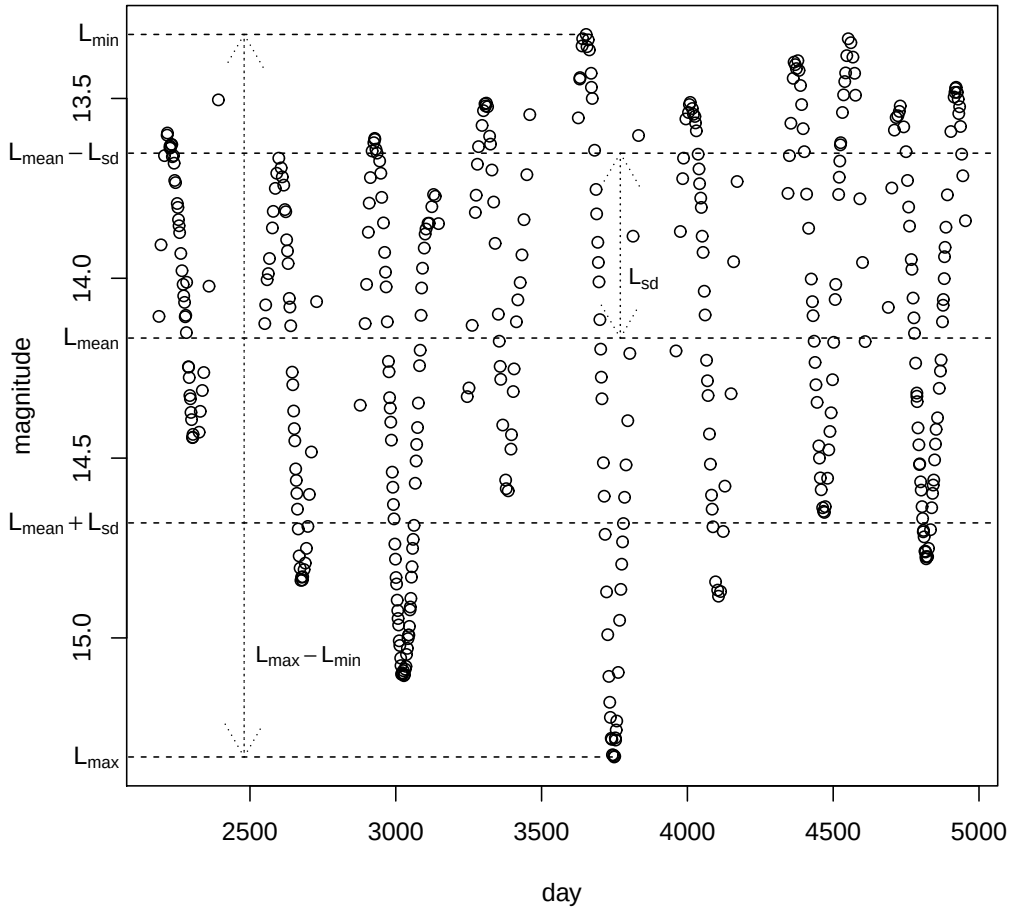


図 3.7: L_{mean}、L_{max}、L_{min}、L_{sd} の定義。典型的なミラ型星の光度曲線を一例として示してある。y 軸は反転してある。

表 3.2: パワースペクトルから取得した特徴量

特徴量	意味
$\nu_{\max}$	パワースペクトル $P(\nu)$ の最大値 $P_{\max} = P(\nu)$ を与える ν。
$\log \nu_{\max}$	$\nu_{\max}$ の常用対数をとったもの。
$P_{\max}/P_{\rm sd}$	パワースペクトルの最大値を標準偏差で正規化したもの。別途説明。
$(P_5 - P_1)/P_{\rm sd}$	$\nu = 0.5$ 付近と $\nu = 0.1$ 付近のパワーの差を標準偏差で正規化したもの。別途説明。
P_5/P_1	$\nu = 0.5$ 付近と $\nu = 0.1$ 付近のパワーの比。別途説明。

P_5/P_1**,** $(P_5 - P_1)/P_{\rm sd}$

図 3.9 のように、R CrB のパワースペクトルはピークを除き、低周波数から高周波数に向けてパワーが増加する傾向がある。これは、周期的でない曲線を高周波でフィッティングしているため見られるもので、これを特徴量として利用するため、低周波数と高周波数のパワーの差と比をとった。

また、P_1 と P_5 以外はナイキスト周波数の設定を 0.5 cycle/day と 1.0 cycle/day の両方を試した。そのため、パワースペクトルから抜き出した特徴量は計 8 種類で、光度曲線と合わせると特徴量は全部で 13 種類になった。

3.4 判別モデルの性能評価と交差検定

前節で示したように、本研究では使用する特徴量を変えて複数の判別器を作成し、その性能評価を行うことで最適な特徴量の組み合わせを調べる。この性能を評価するための量として、次の式で定義する判別の正答率 A を導入した。

$$A = \frac{T}{T + F} \tag{3.1}$$

ここで、T は検定データのうち、判別器により正しく分類されたものの総数、F は誤って分類をされたものの総数である。また、$T + F$ は検定対象天体の総数に等しい。この判別の正答率が高いほど、判別器の性能も高いとした。

本研究では手元のデータ群を、判別器を作成するための学習用集団と、判別器の性能評価をするための検定用集団へ、乱数による分割を行って解析した。しかしこれだと、集団の分け方によって結果が大きく変わることがある。この影響を少しでも減らすために使用したのが交差検定 (cross-validation) と呼ばれる方法である。これは、データの分け方を変えて複数回の解析を行うもので、それらの判別正答率の平均値を評価した。本研究では以下 2 種類の方法を使用した。

3.4.1 10 分割交差検定

データ群を 10 個のグループに分割し、うち 9 グループのデータを学習用として解析に使い、残りの 1 グループで解析結果の検定を行う。これを図 3.10 のように 10 回繰り返すことで、偏りのない結果を得ようとする手法を 10 分割交差検定 (10-fold cross-validation) と呼ぶ。本研究では R CrB を対象に含まない判別でこ

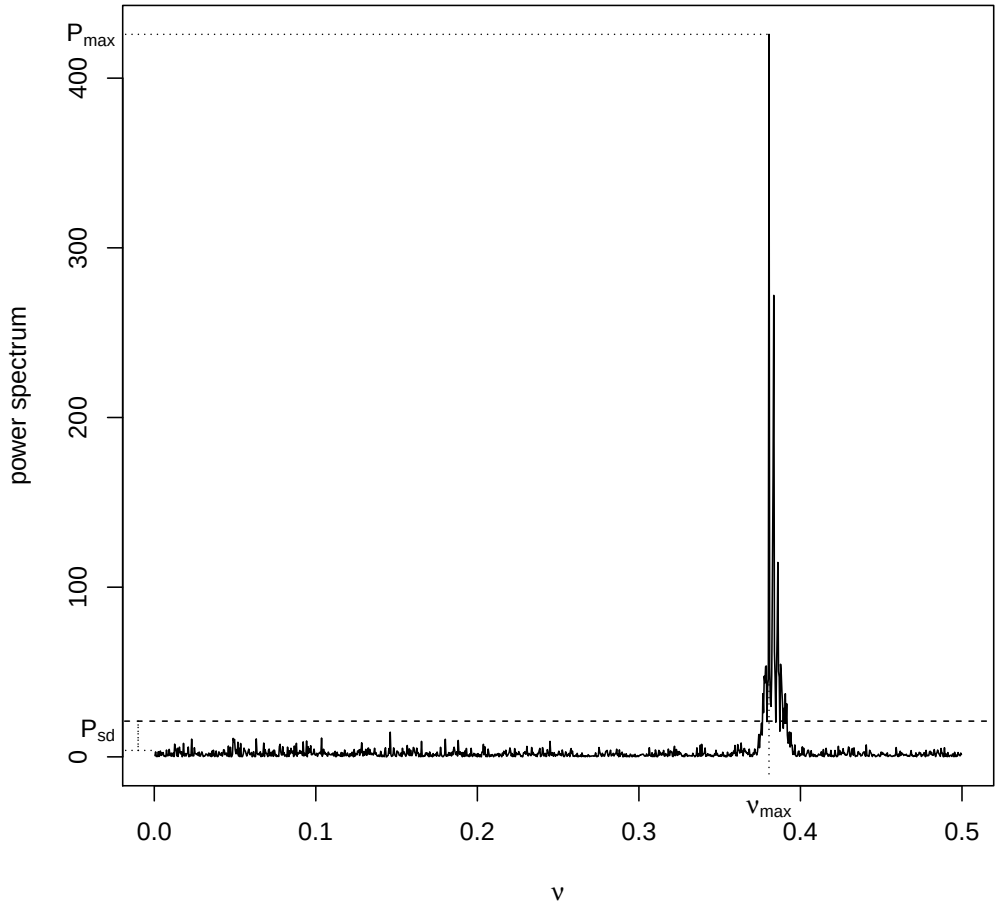


図 3.8: $P_{\max}$ と $P_{\rm sd}$、$\nu_{\max}$ の定義。セファイド型星のパワースペクトルを例に示している。

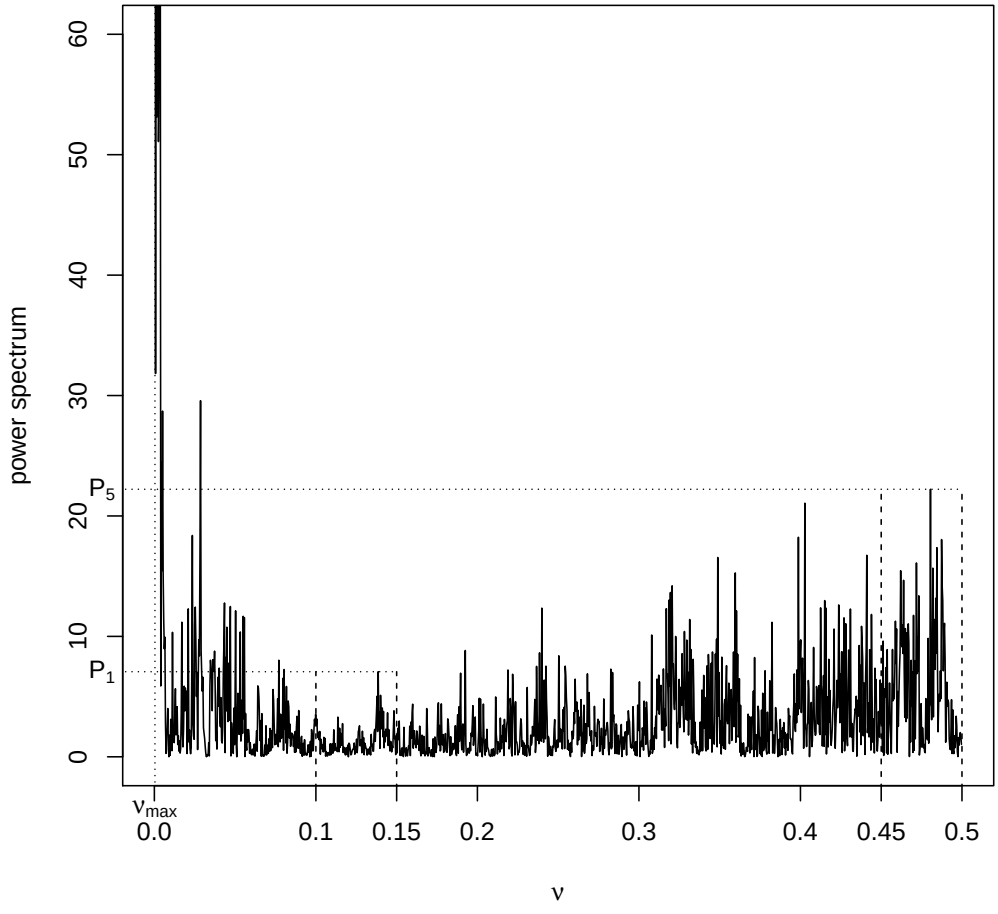


図 3.9: P_1 と P_5 の定義。R CrB のパワースペクトルを拡大したものを一例として示してある。

の手法を使用した。一般に *K* グループへ分割して交差検定を行うことを指す際には *K* 分割交差検定 (*K*-fold cross-validation) とも言われる。

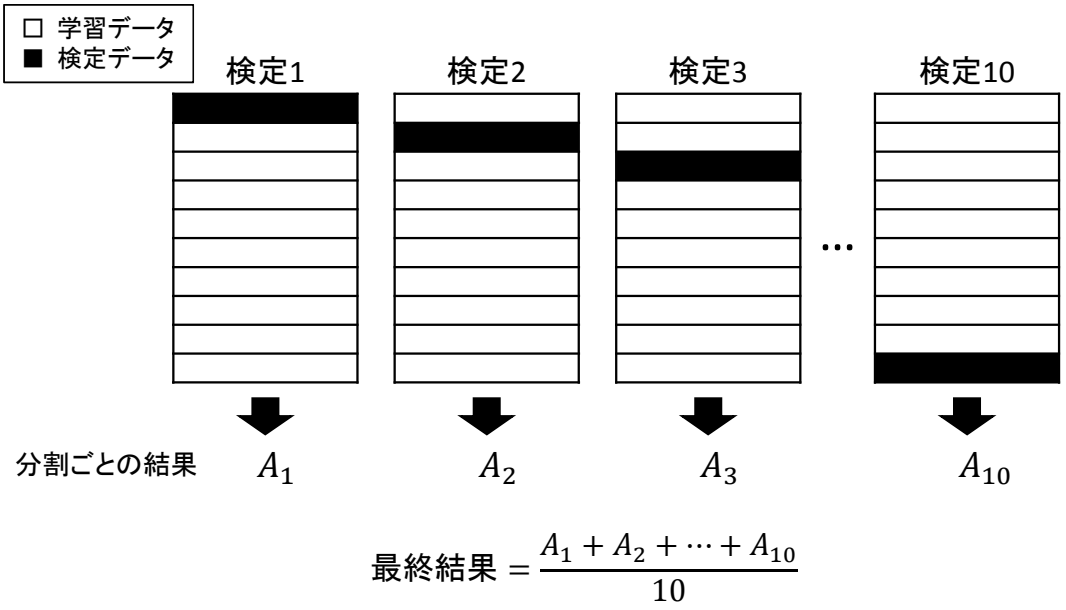


図 3.10: 10 分割交差検定の概念図。

3.4.2　1 個抜き交差検定

n 個のデータ群のうち *n* – 1 個を学習データ、残り 1 つを検定データとした解析を *n* 回繰り返す方法を 1 個抜き交差検定 (leave-one-out cross-validation) と呼ぶ。10 分割交差検定は 10 回の解析を行うのに対して、こちらはデータの数だけ解析を行う。そのため、データの総数が少ない場合に可能な方法であるので、入手できたデータが少ない R CrB を判別対象に含めた解析で使用した。

また、1 個抜き交差検定は *K* 分割交差検定で分割数 *K* をデータの総数 *n* に設定するのと同じことである。

第4章　結果

前にも触れたが、本研究では使用するパラメータの組み合わせを変えて、複数回の判別を行って真に重要な特徴量の選択を目指す。本研究で抽出した特徴量は 13 種類のため、$(2^{13} - 1)$ 回の判別を検定分だけ行った。結果、得られたデータをそれぞれ確認する。

4.1　10 分割交差検定を用いた多サンプルの判別

セファイド型、ミラ型、食変光星、人工ノイズのデータ数は各 1480 個の計 5920 個。それを 10 グループへ分割し、判別を行った。

使用する特徴量の組み合わせを変えて判別を行い、特徴量の組み合わせと正答率を示したものが表 4.1 である。表では、使用パラメータの欄に黒丸を記入、非使用パラメータの欄は空欄にし、正答率が高いものから Case 1, Case 2, と上から順に示してある。また、基本的にパワースペクトルのナイキスト周波数は 0.5 cycle/day だが、1.0 cycle/day に設定して得た特徴量に関しては、プライム記号をつけてある。また、判別の正答率 A に相当する量が Accuracy である。判別に成功した天体が 1 個だけの場合の判別率は、1/5920 個で、およそ 0.00017 である。

表 4.1 では、正答率の差はごくわずかである。そのため、使用パラメータ (特徴量) の細かいバラつきにはあまり意味はない。表全体を見てみると、$\log \nu_{\max}$ と $P'_{\max}$ および全ての光度曲線パラメータが上位 20 位に渡って常に使用されているため、これらは重要な特徴量だといえる。

また、判別正答率が最も高い Case 1 のクラス毎の判別結果を表 4.2 に示す。セファイドと食連星との判別に失敗している数が多いことがわかる。

全ての特徴量を使用した際の正答率は、0.98800 であり、全体のおよそ 60 番目であった。この結果から、パラメータを増やすほど判別精度が上がるとは限らないことが確認できた。

4.2　1 個抜き交差検定を用いた少サンプルの判別

先の 4 クラスに加えて R CrB も対象に加えた 5 クラス判別を行った。こちらのデータ数は、各 18 個の計 90 個であるこれらに、一個抜き交差検定を用いて、全パラメータの組み合わせの判別を行った結果が表 4.3 である。取り扱った天体の総数は 90 天体であるため、天体 1 つの判別成功は判別率 0.01111 である。こちらは、L_{mean} が表中で常に選択されており、重要な特徴量であることがわかる。

全パラメータを使用した場合の判別率は 0.91111 であり、こちらも選択をした場合に比べて劣る結果となった。

また、成績最上位の詳細を、4 クラス判別の場合と同様に表 4.4 に示す。結果を見ると、R CrB がミラ型とされた誤判別が多い。

	$\nu_{\max}$	$\nu'_{\max}$	$\log \nu_{\max}$	$\log \nu'_{\max}$	$\frac{P_{\max}}{P_{\mathrm{sd}}}$	$\frac{P'_{\max}}{P'_{\mathrm{sd}}}$	$\frac{P_5-P_1}{P_{\mathrm{sd}}}$	$\frac{P_5}{P_1}$	L_{mean}	L_{sd}	$L_{\max}-L_{\min}$	$L_{\max}-L_{\mathrm{mean}}$	$\frac{L_{\max}-L_{\mathrm{mean}}}{L_{\max}-L_{\min}}$	Accuracy
Case 1			●	●	●	●			●	●	●	●	●	0.98953
Case 2			●	●		●			●	●	●	●	●	0.98953
Case 3		●	●	●	●	●			●	●	●	●	●	0.98919
Case 4			●	●	●	●		●	●	●	●	●	●	0.98919
Case 5			●			●			●	●	●	●	●	0.98902
Case 6		●	●		●	●			●	●	●	●	●	0.98902
Case 7			●	●		●		●	●	●	●	●	●	0.98902
Case 8			●		●	●			●	●	●	●	●	0.98885
Case 9		●	●			●			●	●	●	●	●	0.98885
Case 10			●	●		●	●		●	●	●	●	●	0.98885
Case 11			●	●		●	●	●	●	●	●	●	●	0.98885
Case 12		●	●	●	●	●	●		●	●	●	●	●	0.98885
Case 13			●	●	●	●	●	●	●	●	●	●	●	0.98885
Case 14		●	●	●		●	●		●	●	●	●	●	0.98868
Case 15		●	●	●		●		●	●	●	●	●	●	0.98868
Case 16		●	●		●	●	●		●	●	●	●	●	0.98868
Case 17		●	●	●	●	●		●	●	●	●	●	●	0.98868
Case 18		●	●	●	●	●	●	●	●	●	●	●	●	0.98868
Case 19			●			●		●	●	●	●	●	●	0.98851
Case 20		●	●	●		●			●	●	●	●	●	0.98851

表 4.1: セファイド型、ミラ型、食変光星、人工ノイズの分類結果。

		推定されたクラス			
		セファイド	食連星	ミラ型	人工ノイズ
正しいクラス	セファイド	1450	25	5	0
	食連星	23	1449	2	6
	ミラ型	0	0	1479	1
	人工ノイズ	0	0	0	1480

表 4.2: 表 4.1 の成績最上位の特徴量の組み合わせで判別を行った結果。

	$\nu_{\max}$	$\nu'_{\max}$	$\log \nu_{\max}$	$\log \nu'_{\max}$	$\frac{P_{\max}}{P_{\mathrm{sd}}}$	$\frac{P'_{\max}}{P'_{\mathrm{sd}}}$	$\frac{P_5-P_1}{P_{\mathrm{sd}}}$	$\frac{P_5}{P_1}$	L_{mean}	L_{sd}	$L_{\max}-L_{\min}$	$L_{\max}-L_{\mathrm{mean}}$	$\frac{L_{\max}-L_{\mathrm{mean}}}{L_{\max}-L_{\min}}$	Accuracy
Case 1	●	●			●		●		●	●	●			0.93333
Case 2		●	●			●		●	●	●		●		0.93333
Case 3	●	●			●		●	●	●	●	●			0.93333
Case 4	●			●	●		●		●	●	●	●		0.93333
Case 5	●			●	●		●	●	●	●	●	●		0.93333
Case 6	●	●			●	●	●		●	●	●	●	●	0.93333
Case 7			●			●			●	●				0.92222
Case 8	●					●	●		●	●				0.92222
Case 9	●					●	●		●		●			0.92222
Case 10		●	●		●				●		●			0.92222
Case 11		●	●			●			●	●				0.92222
Case 12		●	●			●			●		●			0.92222
Case 13		●	●			●			●			●		0.92222
Case 14		●		●	●				●		●			0.92222
Case 15		●		●		●			●	●				0.92222
Case 16		●		●		●			●		●			0.92222
Case 17			●	●		●			●	●				0.92222
Case 18			●	●		●			●		●			0.92222
Case 19			●		●		●		●	●				0.92222
Case 20			●			●	●		●	●				0.92222

表 4.3: セファイド型、ミラ型、食変光星、R CrB、人工ノイズの分類結果。

		推定されたクラス				
		セファイド	食連星	ミラ型	人工ノイズ	R CrB
正しいクラス	セファイド	18	0	0	0	0
	食連星	1	17	0	0	0
	ミラ型	0	0	16	1	1
	人工ノイズ	0	0	0	18	0
	R CrB	0	0	3	0	15

表 4.4: 表 4.3 の成績最上位の特徴量の組み合わせで判別を行った結果。

第5章　考察

使用パラメータと判別正答率の表からみられる、重要なパラメータに関して考察を行う。

5.1　$\nu_{\max}$ のクラス内分散と判別正答率

表 4.1 の結果で、$\nu_{\max}$ と $\log \nu_{\max}$ を使用する頻度に大きな差がある。これは $\nu_{\max}$ に比べて $\log \nu_{\max}$ は軸のクラス間分散を大きくし、決定境界の勾配をゆるやかにしているためである。その例が 2 章の人工データ実験の図 2.7 と図 2.8 である。各変光星型での $\nu_{\max}$ と $\log \nu_{\max}$ のヒストグラムを図 5.1 と図 5.2 に示してある。図 5.2 の方では、各々の分散がおおむね等しくなっていることが確認できる。

また、R の lda 関数は各クラスのデータ数が等しい場合、通常は 2 クラス間の平均値をしきい値に設定する。それゆえに偏った分布であると、理想的でない決定境界が定められてしまうことがある。図 5.3 は、$P_{\max}$ に加えて $\nu_{\max}$ もしくは $\log \nu_{\max}$ の 2 つの特徴量のみを使って、セファイド型とミラ型の判別を行った図である。$\nu_{\max}$ そのものを使用した方は決定境界がセファイドの領域に入っているが、対数スケールの方は各々のクラス内分散が近い値となったため、問題が改善されている。

5.2　セファイドと食連星の判別

図 5.4 と図 5.5 はともに食連星がセファイドに誤判別されてしまった例である。この光度曲線だけをみると、両者とも何の型に分類されるべきか、人の目からは見当をつけにくい。増光と減光が対称なセファイドと食が浅い食連星は、類似した光度曲線を描くという問題がある。図 5.6 と図 5.7 がその一例である。これは、セファイドと食連星で誤判別が多い原因の 1 つと考えられる。これらの天体は本質的に判別が困難なものであり、その意味では今回基準とした OGLE の判別結果にも誤判別が含まれている可能性がある。今回は OGLE の判別が完全に正しいものとして正答率を推定したが、OGLE 側に誤判別が存在する場合はこの正答率を過小評価している可能性がある。また、図 5.8 のように、三角関数的な光度曲線を描く食連星では、セファイドとの判別材料が少ない。そのため、天体の色等を特徴量として使うと正しい判別が可能になるかもしれない。

一方、食連星は比較的暗い天体であり、セファイドと食連星の L_{mean} はそれぞれ平均で 15.4 等と 18.0 等であった。対して、食連星のはずがセファイドに誤判別された図 5.6 と図 5.7 の光度曲線を見てみると、L_{mean} が約 14.8 等と、この 2 つが明るい食連星であることがわかる。また、表 4.2 で食連星からセファイドに誤判別された 23 天体のうち 17 天体の L_{mean} が 16.0 等以下であった。対してセファイドから食連星に誤判別されたもので、$L_{\mathrm{mean}} > 17$ 等であるものが 25 天体中の 21 天体であった。また、食連星で $L_{\mathrm{mean}} < 16$ 等であるものは、全体 1480 天体のうち 86 天体、セファイドで $L_{\mathrm{mean}} > 17$ 等であるのは、80 天体であるため、誤判別されたものの明るさは偏っているといえよう。各クラスごとの L_{mean} の平均および、誤判別されたも

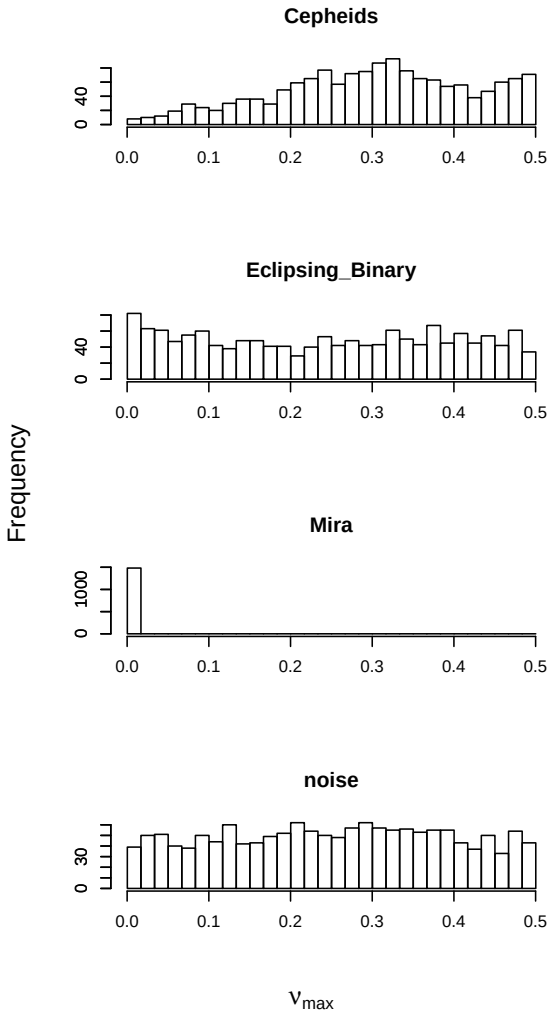


図 5.1: 各変光星型の $\nu_{\max}$ 分布

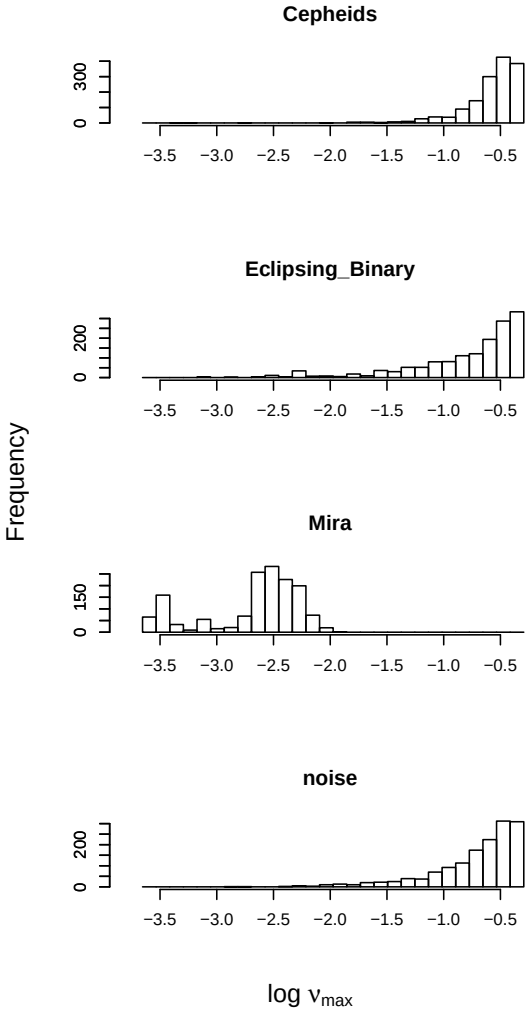


図 5.2: 各変光星型の $\log \nu_{\max}$ 分布

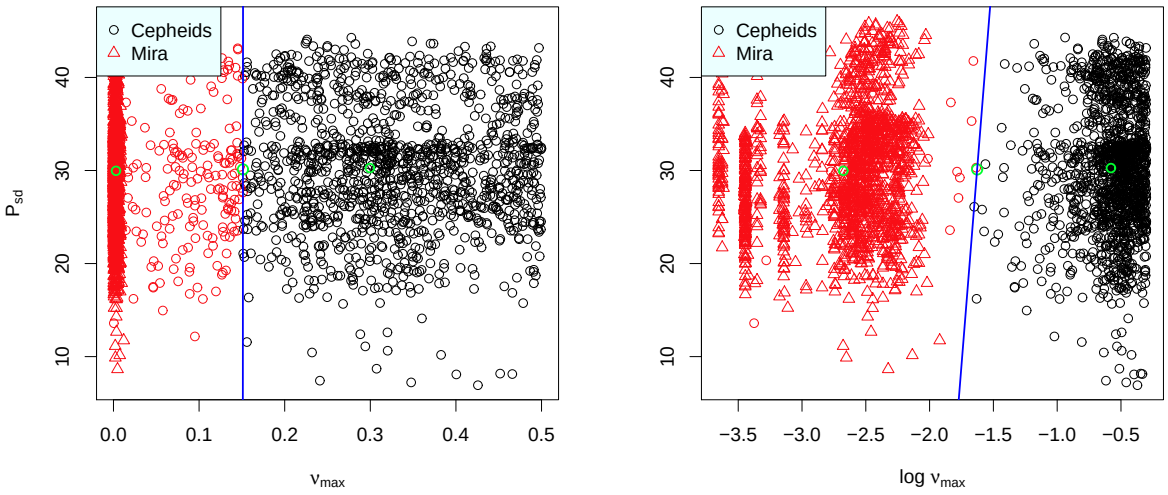


図 5.3: $\nu_{\max}$ と P_{sd} を使って 2 クラス判別をした結果 (左) と $\log \nu_{\max}$ と P_{sd} を使って 2 クラス判別をした結果 (右)。緑点は全体と各クラスの平均。青線は決定境界。プロット記号が正しいクラスを、色が分類後のクラスを表す。

のの平均を比較しているのが図 5.9 である。これで誤判別されたものの偏りが確認できる。つまり、L_{mean} はセファイドと食連星を判別するために重要な特徴量だが、セファイドと食連星の 2 つのクラスの L_{mean} の分布は完全には分離しておらず、それらの境界上にいる天体が誤判別されていると考えられる。これら明るい食連星のや暗いセファイド型の誤判別を減らす案としては、パワースペクトルの倍振動を取ることが考えられる。食外で平坦な光度曲線をもつ食連星とセファイドでは光度曲線の波形が明らかに異なるので、波形の情報を含む 2 倍、3 倍周波数のパワーを利用すれば判別精度が上がる可能性がある。

また、誤判別されたものの光度曲線には外れ値を持つものもいくつかあり、中には図 5.10 のように 90 等級以上の観測記録をもつものも存在した。本研究で使った $L_{\max}$ や $L_{\min}$ という、最大値や最小値を扱う特徴量で、外れ値の影響が強く出て誤判別がでたと考えられる。外れ値を除去する仕様を導入すれば、判別正答率が上昇すると期待される。

5.3 L_{mean} の有用性

表 5.1: 正答率 9 割以上での特徴量の使用頻度 (%)。50%以上のみ表示してある。

L_{mean}	L_{sd}	$\frac{P'_{\max}}{P'_{\mathrm{sd}}}$	$L_{\max} - L_{\min}$	$\log \nu_{\max}$	$\frac{P_{\max}}{P_{\mathrm{sd}}}$	$\frac{P_5 - P_1}{P_{\mathrm{sd}}}$	$\nu'_{\max}$	$\nu_{\max}$	$\frac{P_5}{P_1}$	$L_{\max} - L_{\mathrm{mean}}$
90.09	70.99	65.82	63.70	58.89	58.09	56.78	55.98	53.28	50.22	50.00

4.2 章で述べた 5 クラス判別で、9 割以上の正答率を出した変数組み合わせのうち、使用された各変数の割合を示したのが表 5.1 である。これを見ると、最もよく使用されているのは L_{mean} の 90.1% であり、これ

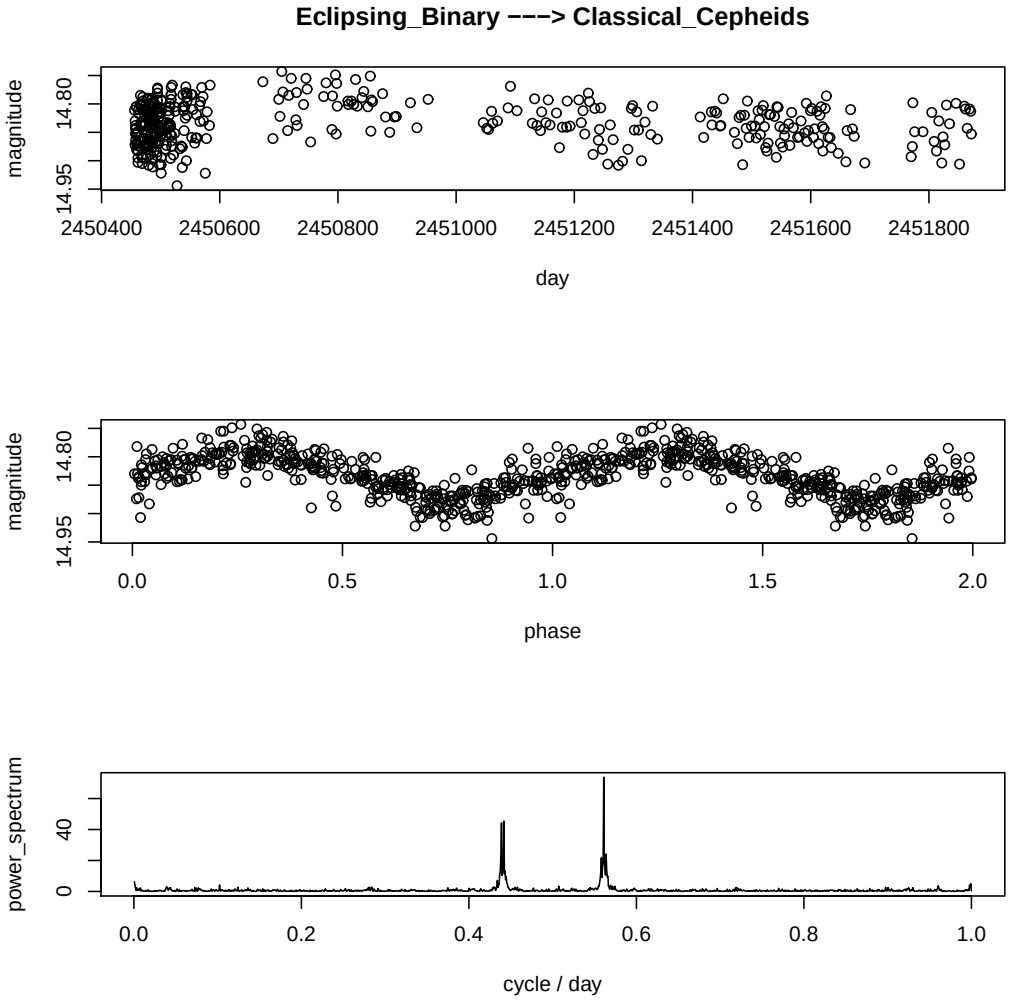


図 5.4: 正しい周期が求まって、セファイドと誤判別された食連星

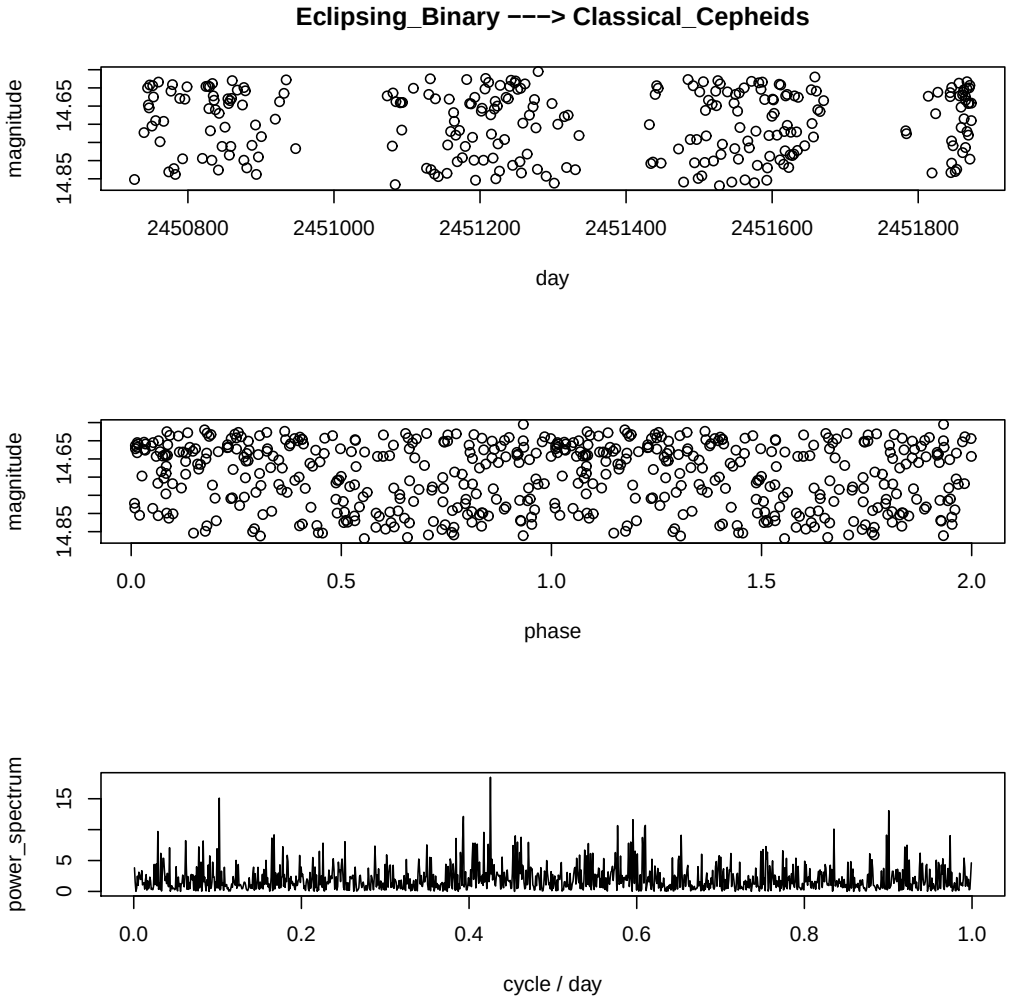


図 5.5: 正しい周期は求まらず、セファイドと誤判別された食連星

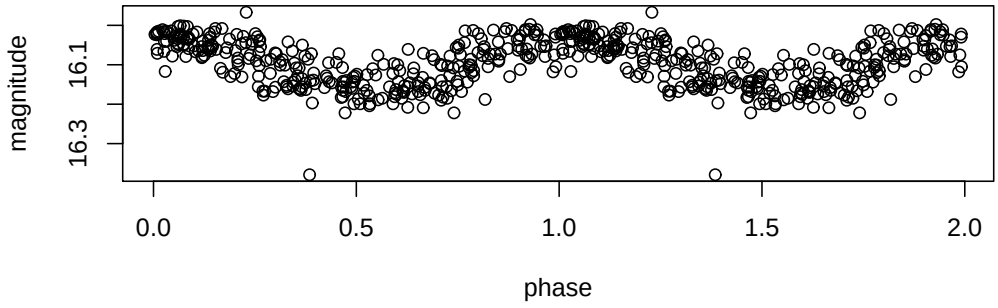


図 5.6: 食連星と誤判別されたセファイドの一例。

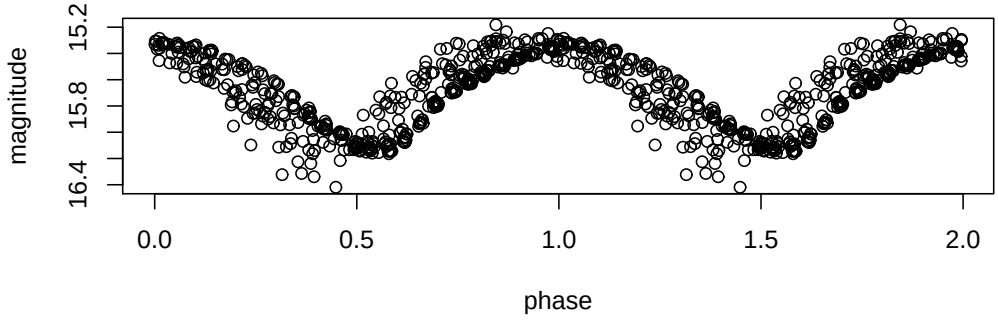


図 5.7: セファイドと誤判別された食連星の一例。

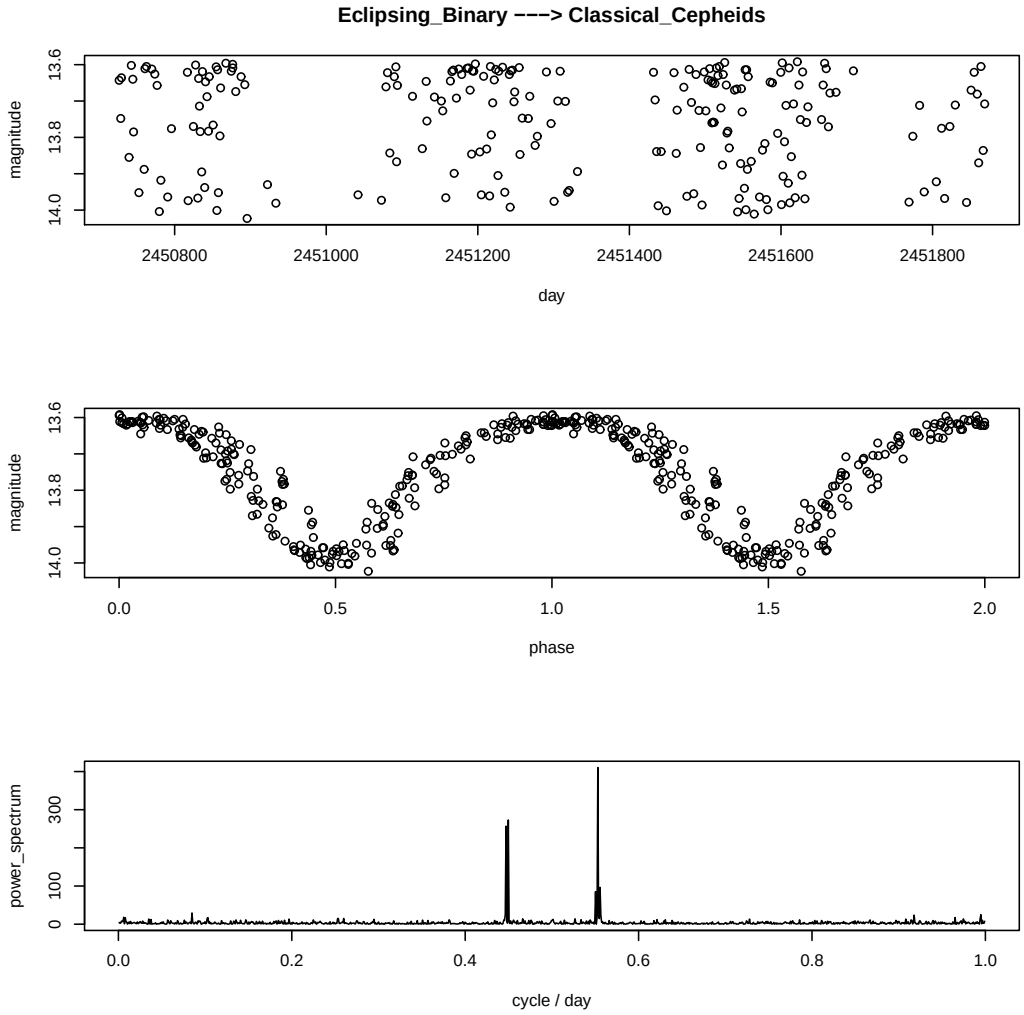


図 5.8: 正弦波を描く食連星の例。

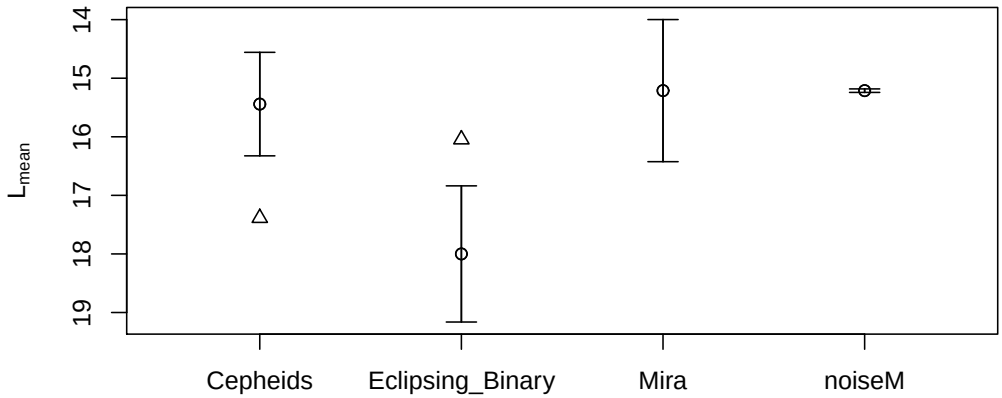


図 5.9: 各クラスの L_{mean} の平均値。丸が全体の平均で、その標準偏差をバーで表している。三角が誤判別されたものの平均。Mira は誤判別が 1 つであるため入れてない。

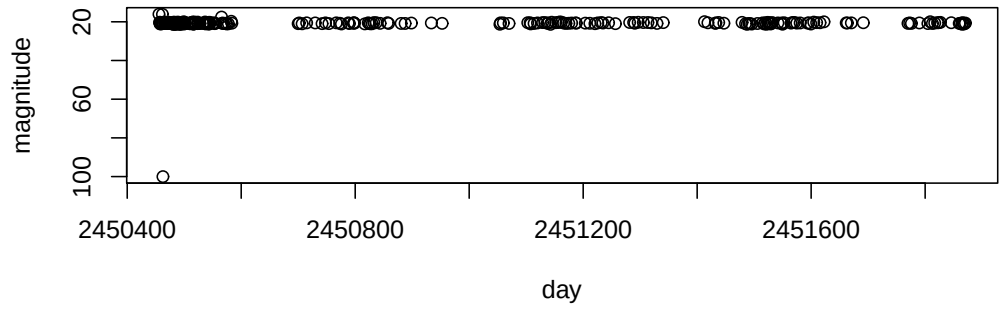


図 5.10: 極端な外れ値を持った光度曲線。

が突出して高い。その下に L_{sd} の 71.0%, P'_{max} の 65.8% と続く。

これは、R CrB が変光星の中で最も明るい星であるために [10]、L_{mean} が使用されていると推察される。今回使用した天体データは大マゼラン雲 (LMC) の領域に限ったものであり、地球からの距離がほぼ一定である。そのため、みかけの等級が絶対等級としての役割も持っている。先にも触れたが、比較的暗い星である食連星にも、L_{mean} が判別の基準になると期待される。

5.4 サンプル数の影響

少サンプル判別では、交差検定の結果に対するデータ 1 つ 1 つの重要性が大きくなっている。

L_{sd}, L_{max}, L_{mean} は光度曲線の振幅に関する特徴量で、図 5.11 のように強い相関関係にある。表 4.3 の Case 7 から Case 20 ではこれらのうち、1 つが選択されて使用されている。これは、少サンプルで使用したデータでは、L_{sd} 等は 1 つで十分であったためと考えられる。

以上の理由もあり、サンプル数が少ないと重要と判断できる特徴量の数も少なくなる。

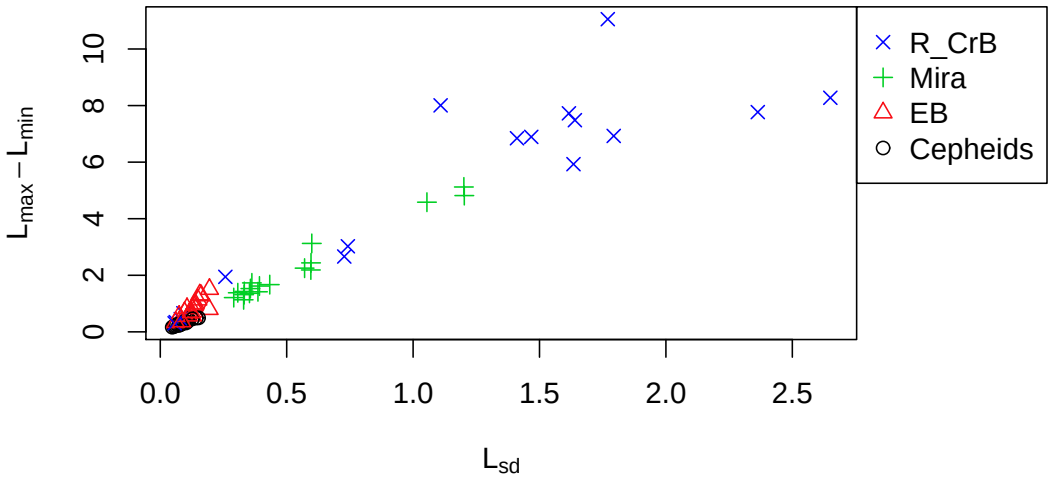


図 5.11: 変数 L_{sd} と $L_{max} - L_{min}$ の相関関係を示してある。

第6章　まとめ

本研究では、分類済みのセファイド、ミラ、食連星、R CrB、人工ノイズの 5 クラスに対する判別器の作成を行った。判別器の作成に必要な特徴量は光度曲線とパワースペクトルから定義した。得られた特徴量と分類先の情報を基に、フィッシャーの線形判別による機械学習で判別器を作成した。この性能評価には、少数データである R CrB を含まない 4 クラスの判別には 10 分割交差検定で、R CrB を含む 5 クラスの判別には 1 個抜き交差検定で行った。

変光星を分類する試みは以前より行われてきたが、今回は重要な特徴量の候補の全組み合わせを試すことで、真に重要な特徴量を選択するという新しい手法を用いた。これにより、天文分野で行う判別の先駆け的役割を目指した。

その結果、4 クラスの判別では元データに対する正答率が 98.95% の、5 クラスの判別では正答率 93.33% の判別器を作成することに成功した。一方、誤判別されたデータは、外れ値を持つものが複数存在し、外れ値を除去する機構を導入すれば、結果の改善が望まれる。また、類似した光度曲線を持つ別の型の天体に関しては、今回は使用しなかった色などの情報を使えば分類可能性が見込まれる。

謝辞

まず卒業研究や論文執筆時に気を遣って頂いた研究室の先輩方に感謝します。同期の方々も数多くの助けをいただきありがとうございました。先生方には物理の知識や考え方を教えて頂きました。ありがとうございます。特に、卒業研究の指導や文章の添削をしていただいた植村先生には感謝してもしきれません。

また研究室や学科の違う友人や遠く離れた旧友たち、私を支え続けててくれた両親にも深く感謝します。本当にありがとうございました。

参考文献

[1] 山本 大空 2015 卒業論文

[2] York, D. G., Adelman, J., Anderson, J. E., Jr., Anderson, S. F., Annis, J., Bahcall, N. A., et al., 2000, AJ, 120, 1579

[3] Udalski, A., Kubiak, M., and Szymanski, M., 1997, AcA, 47, 319

[4] 東北大学大学院環境科学研究科
https://www.tohoku.ac.jp/japanese/newimg/pressimg/tohokuuniv-press_20141127_02web.pdf

[5] NASA HP "AutoClass"
http://ti.arc.nasa.gov/tech/rse/synthesis-projects-applications/autoclass/

[6] Ita, Y., Tanab, T., Matsunaga, N., Nakajima, Y., Nagashima, C., et al., 2004, MNRAS, 347, 720

[7] Matsunaga, N., Sakamoto, T., and Maehara, H., 2012,
http://www.ioa.s.u-tokyo.ac.jp/kisohp/RESEARCH/symp2012/Matsunaga2.pdf

[8] Yanagisawa, K., et al., 2015, http://www.oao.nao.ac.jp/stockroom/extra_content/um15/O17_Yanagi.pdf

[9] C.M. ビショップ, 2012, 『パターン認識と機械学習 上』(元田 浩・栗田多喜夫・樋口知之・松本裕治・村田 昇 監訳), 丸善出版, pp. 178-190

[10] Jeffery,C.S., 2008 http://ads.nao.ac.jp/abs/2008CoAst.157..240J