DEIM Forum 2023 4a-6-3

周期解析による変動天体の検出

千原　直己† 　高田　唯史†† 　藤原　靖宏††† 　鬼塚　真†

† 大阪大学大学院情報科学研究科　〒 565-0871 大阪府吹田市山田丘 1–5
††† NTT コミュニケーション科学基礎研究所　〒 243-0198 神奈川県厚木市森の里若宮 3-1
†† 自然科学研究機構国立天文台　〒 181-8588 東京都三鷹市大沢 2-21-1
†† 総合研究大学院大学　〒 181-8588 東京都三鷹市大沢 2-21-1
E-mail: †{chihara.naoki,onizuka}@ist.osaka-u.ac.jp, ††tadafumi.takata@nao.ac.jp,
†††yasuhiro.fujiwara.kh@hco.ntt.co.jp

あらまし　変動天体とは, 光度や位置などが時間とともに変化する天体のことであり, その検出は, 高精度な情報や新たな知見の獲得に繋がるため重要である. 既存技術では, 天体ごとの統計量ベースの探索手法が主流であるが, 問題が 2 つ存在する. i) 観測情報を統計的な数値に集約してしまうため, 変動天体の持つとされている周期性関連の情報が損なわれてしまう, ii) 天体の観測可能な時刻が限定されることにより, 基本的に多数の欠損値が含まれたデータしか用意できないことである. これらの問題を解決するため, 本稿では, スパースモデリングを取り入れた周期解析による変動天体の検出を提案する. 具体的には, 周期解析により時系列の重要な情報を抽出した後, それらを活用し, 二値分類にて変動天体かどうかの判定を行う. 実験では, ノイズの混入した人工データに対しても正常な学習が可能であることを明確にした. 加えて, 実データに対しても同様に提案手法の有効性を示した.

キーワード　変動天体, 周期解析, スパースモデリング, 異常検知

1　はじめに

変動天体には, i) 位置が変動する天体, ii) 光度が変動する天体の 2 種類存在する. 変動については, 長期的なものもあれば, 局所的なもの, それらが混在しているものに加え, 一過性であるものや周期性を持つものもなど多岐にわたる [1]. 加えて, それは全天体の約 1% 程度を占める. 変動天体の代表例としては, 閃光星, 新星, 超新星などが挙げられる. 本研究では, 光度が周期的に変動する変動天体のみを対象とする. 非常に稀な天体現象を持つ変動天体を発見することは, 宇宙の構造, 進化に関する研究や, 物理現象の解明などにとって有益となりうる高精度な情報や新たな知見を得ることと密接な関わりがある [2–4].

既存手法には統計的手法を活用したものが多い. ここでは有名な手法を二つ紹介する. 一つ目は B.Sesar らの手法 [2] である. この手法は, 時系列から得られる intrinsic variability σ, χ^2, 等級の測定値の平均, 標準偏差, 歪度といった統計的数値を活用したスコアリングを基にした変動天体検出法の一種である. この手法では, 天体データを統計的な数値に集約するため, 変動天体の持つ時間情報が損なわれる, すなわち, 変動天体の特性として知られる光度曲線の周期性を活用出来ていない, という問題点が考えられる. 二つ目の手法は, lomb-scargle periodogram [5,6] である. こちらは, 最小二乗フィッティングを活用した, 時系列に存在する周期性の分析を行うための手法である. サンプリング間隔が一定でない時系列に対しても適用できるため, 観測環境により一定間隔でのデータ収集が困難である天文分野にてよく活用されている手法である. しかし, この手法にもいくつか問題点がある. 具体的には, i) ノイズや欠損値への耐性が低い, ii) 時系列に複数の周期的な変動が含まれる場合にはそれらを同時に分析することができない場合がある, iii) 周波数分解能が低い, といったものが挙げられる.

本研究では, 機械学習的アプローチを活用した周期解析により既存手法の問題点を解決することで, より高精度な変動天体検出の達成を目的とする. 全体の流れとしては, 初めに回帰分析を活用することで天体データから周期情報を取得する. そして, 後半部分では, 得た情報を元に特徴量設計を行い, 分類タスクの解決により変動天体候補を取得する. しかし, この実現のために幾つかの課題点が存在する. 一つ目が, 利用する天体データに多数の欠損値が含まれていることである (図 1 参照). 原因として考えられることは, 観測の際, 天候や季節などといった外的影響に加え, 望遠鏡の口径による観測範囲や天体の観測時間に関する制限などが挙げられる. そのため, ナイーブな回帰分析を適用すると過学習が生じてしまい, 望ましい結果が得られる見込みが薄い. 二つ目が, 特徴量設計に使用するデータ候補が周期情報のみであるということである. すなわち, 分類タスクによる変動天体の予測が周期解析結果に完全に依存する. 故に, 懸念事項として, 誤った周期解析結果が得られた場合, i) 対象の天体データが含まれたデータセットで訓練を行うことにより, モデルのパフォーマンスが大幅に下がる, ii) 対象の天体データを正しく予測することが大凡不可能である, といったことが挙げられる.

本研究では, 上記の前者の問題点を解決すべく, スパースモデリングの活用を提案する. スパースモデリングとは, データから有為な性質のみを学習し, 全体から重要な変数のみに限定させる, という変数選択手法の一種である. 訓練データから核と

図 1: 使用データセット例

各天体が観測された明るさの時間変化を示したデータである．横軸が時間であり，縦軸が等級を示している．但し，縦軸の等級の値は実際の値を標準化したものである．観測されたデータの割合が全体の約 0.01%と非常にスパースなデータとなっている．

なる箇所を選定するため，ノイズなどの影響による細かなデータに対して過学習しにくいといったメリットが挙げられる．また，後者の問題点を解決するために，ドメイン知識を駆使した特徴量エンジニアリングを提案する．分類タスクの訓練及び予測を行う際に，変動天体の傾向を正確に反映させることで高精度な検出を実現する．

以下に提案手法の貢献点を要約する．

高精度　既存の変動天体検出手法とは異なり，提案手法では重要周期の抽出により，変動天体の特徴である周期性を考慮した検出を行う．さらに，重要周期の抽出は数百件程度の候補から自動探索を行うため，周期長が不明な変動天体に対しても高精度な処理が担保される．

ロバスト性　本研究で取り扱うデータセットは，欠損値が非常に多いことに加え，ノイズが混入している蓋然性が高い．このような特徴を持つデータに対しても，スパースモデリングを活用することにより，ノイズや過学習に対してロバストな処理を実現した．

高スケール性　既存の異常検知手法であると，全体の傾向に基づいて例外的な挙動の発見をする必要があるため，並列化が難しい傾向にある．しかし，提案手法では，天体毎に独立に処理が可能であるため，100 万天体規模に対しても高スケールな処理が期待できる．

2 研究背景

本章では，解決すべき問題の定義並びに本論文で必要な前提知識の説明を行う．表 2 に提案時に使用する記号とその定義を記載する．

2.1 問題定義

本研究の目的は，多数の天体データから変動天体を検出することである．入力は天体の光度の変化を表す時系列データセット $\mathcal{Y} = \{\boldsymbol{y_1}, ..., \boldsymbol{y_N}\}$，周波数候補の数 m，ナイキスト周波数 f_n，特徴量リスト $\mathcal{F}$ に属する各特徴量の内，ドメイン知識に関連したものに限定したもの $\mathcal{F}_d$ であり，出力は変動天体候補セット $\mathcal{L}$ である．本研究にて取り組む問題を数式化すると (1) 式のように与えられる．

$$\mathcal{L} = F(\mathcal{Y}, m, f_n, \mathcal{F}_d) \tag{1}$$

表 2: 記号とその定義

記号	定義
周期解析	
n	時系列の長さ
m	周期解析に用いる周波数の候補数
N	天体の件数
f_n	ナイキスト周波数
X	説明変数 (周波数) $X \in R^{n \times 2m}$
$\boldsymbol{y}$	目的変数 (天体データ) $\boldsymbol{y} \in R^n$
$\mathcal{Y}$	天体データセット $\mathcal{Y} = \{\boldsymbol{y_1}, ..., \boldsymbol{y_N}\}$
$\boldsymbol{w}$	回帰係数 $\boldsymbol{w} \in R^{2m}$
$\boldsymbol{\nu}$	周波数候補 $\boldsymbol{\nu} \in R^m$
Group Lasso	
J	グループ数
g_j	j 番目のグループに属する変数の数　$(1 \leqq j \leqq J)$
$\mathcal{G}$	グループセット $\mathcal{G} = \{\boldsymbol{G_1}, ..., \boldsymbol{G_J}\}$　$(\boldsymbol{G_j} \in R^{g_j})$
二値分類	
$\mathcal{F}$	特徴量リスト $\mathcal{F} \in R^{N \times (m+5)}$
$\mathcal{F}_d$	ドメイン特徴量リスト $\mathcal{F}_d \in R^{N \times 4}$
$\mathcal{L}$	変動天体候補セット

但し，F は本研究における提案モデルを表す関数である．

2.2 技術的課題

本研究の課題は主に 2 種類存在する．1 つ目が，解析すべきデータセットに多数の欠損値が含まれている点である (図 1 参照)．そのため，天体データに直接フーリエ変換を利用するためには欠損値補完を適用する必要がある．しかし，観測データ点の割合は平均 0.01%程度であり，残りの約 99.99%に欠損値補完を適用すると，正確な周期解析が困難であるため，フーリエ変換を直接適用することが困難である．これを解決するため，回帰問題として本研究を扱う．但し，その際に過学習にも気を配る必要がある．こちらの問題を緩和させるため，特徴量選択によりモデルの汎化性を上げる技術としてスパースモデリングを用いる．

2 つ目が，変動天体検出の際に使用する特徴量設計である．現状，特徴量設計に使用するデータ候補が周期情報のみであるため，高精度な分類を実現するためには，正確な周期解析が求められる．しかし，使用する天体データには多数の欠損値が含まれているため，精度には限度があると考えられる．また，例え周期解析により天体の持つ適切な周期情報が得られたとしても，分類の際にそれらが上手く機能する変動天体の傾向に関する情報がなければ高精度な結果は期待できないため，ドメイン知識を考慮した特徴量設計を行う必要がある．

2.3 前提知識

フーリエ変換: 時系列を時間領域から周波数領域へ変換する手法．本研究では，天体データの周期性を調べるために用いる．具体的には，$\mathcal{Y}$ の一要素であり，光度の変化を表現している時系列 $\boldsymbol{y}$ を正弦波，余弦波の和で書き表すために使用する．数式で書き表すと，以下の (2) 式のようになる．但し，t は時刻を表

しており, ν_j は周波数候補の j 番目を表している.

$$\boldsymbol{y} = \sum_{j=1}^{m} sin(2\pi t \nu_j) + cos(2\pi t \nu_j) \tag{2}$$

Lasso [7]: スパースモデリングの一種. L1 正則化項を活用することにより, 更新を繰り返したとしても不必要な特徴量に対応した回帰係数は 0 のまま維持される, すなわち, 特徴量選択が行われる. 詳細については以下の補題 2 にて示し, 本稿における Lasso の損失関数は以下の (3) 式のように書き表す.

補題 1. Lasso を用いて回帰分析を行う際, 説明変数 $X \in R^{n \times m}$, 目的変数 $\boldsymbol{y} \in R^n$, 回帰係数 $\boldsymbol{w_{-k}} \in R^{m-1}$ に対して, w_k が $-n\lambda < w_k < n\lambda$ を満たす場合, 値の更新が行われない. 但し $\boldsymbol{w}_{-k}$ とは, 列ベクトル $\boldsymbol{w}$ から k 番目の要素, すなわち w_k を除いたベクトルを表す.

$$E_{Lasso} = \frac{1}{2n}||\boldsymbol{y} - X\boldsymbol{w}||_2^2 + \lambda||\boldsymbol{w}||_1 \tag{3}$$

証明. 付録 1 参照 □

周期解析の際, 重要度の高い特徴量のみを検出できれば良いため, 特徴量選択が可能な Lasso は本研究にて適切な手段であると考えられる. なぜなら, 特徴量選択は分析の際, 過学習の発生を抑制する効果を持つためである. この仕組みについては, 特徴量選択によりデータの根幹を表現するために不要と判断された特徴量を 0 とすることによるものである. すなわち, データの細部までの表現を行わないことにより, ノイズなどの影響による些細な誤差への過剰なフィッティングが抑制される.

Group Lasso [8]: Lasso の応用手法. Lasso と異なる特徴として, 特徴量選択の際, 一つ一つに対してではなく, グループ単位で行う. すなわち, 事前に関係性のある特徴量群が考えうる場合, それらをグループ化し回帰分析に反映させることで, Lasso 以上の高精度なモデリングが期待できる. 詳細については以下の補題 1 にて示し, 本稿における Group Lasso の損失関数は以下の (4) 式のように書き表す. また, 式中の $\boldsymbol{w_{G_j}}$ は j 番目のグループ $\boldsymbol{G_j}$ に属する特徴量を表すベクトルである.

補題 2. Group Lasso を用いて回帰分析を行う際, 説明変数 $X \in R^{n \times m}$, 目的変数 $\boldsymbol{y} \in R^n$, 回帰係数 $\boldsymbol{w} \in R^m$, グループセット $\mathcal{G} \in R^J$ に対して, $\hat{\boldsymbol{w}}_{\boldsymbol{G_j}}(\overset{\text{def}}{\Longleftrightarrow} \boldsymbol{w_{G_j}} - \eta\nabla f(\boldsymbol{w_{G_j}}))$ が $||\hat{\boldsymbol{w}}_{\boldsymbol{G_j}}|| > \eta\lambda$ を満たす場合, 値の更新が行われない.

$$E_{GroupLasso} = \frac{1}{2n}||\boldsymbol{y} - X\boldsymbol{w}||_2^2 + \lambda \sum_{\boldsymbol{G_j} \in \mathcal{G}} ||\boldsymbol{w_{G_j}}|| \tag{4}$$

証明. 付録 2 参照 □

本研究では, 任意の周波数候補に対して, 正弦波及び余弦波の計 2 つの特徴量を与える. 但し, これらは周期解析である以上, 不可分な関係にある. Lasso ではこれらを独立に扱った解析しか行えないが, Group Lasso ではグループ化することにより上記の問題を解消した解析が可能である.

LightGBM [9]: 決定木アルゴリズムに基づいた勾配ブースティングの教師あり機械学習手法 [10] の一種である. すなわち, 比較的浅い決定木を複数用意し, それらを束ねて学習を行う. 目的変数に応じて、説明変数を分類することが可能である. 関連した手法に XGBoost [11] があるが, 最終的には 100 万天体規模の分類を行う必要があるため, より計算コストが小さい LightGBM を採用した.

3 提 案 手 法

本章では, 変動天体を検出するまでの提案モデルについて説明する. 提案モデルは以下の 2 部から構成されている. 概要に関しては図 3 にてまとめている.

周期解析

複数の周波数候補を表現する, 正弦波及び余弦波にて構成された説明変数 X 及び, 天体データである目的変数 $\mathcal{Y}$ を用意する. その後, 各天体データに対して Group Lasso を適用し, 回帰係数 $\boldsymbol{w}$, 決定係数 reg_score を算出する. (図 3 左側の周期解析部分)

変動天体検出

回帰係数 $\boldsymbol{w}$ を始めとした周期情報に加え, ドメイン知識によって構成された特徴量セットを用いて, LightGBM により変動天体かどうかの判定を行う. (図 3 右側の変動天体検出部分)

3.1 周 期 解 析

欠損値が多数混在したデータセットに対して, 周期情報の取得にフーリエ変換を直接適用するためには欠損値補完が必要不可欠であるが, 欠損値が多すぎるために, 有用な補完値を得ることが非常に困難である. 故に, 回帰問題として周期解析を取り扱う. このような周期解析を実行するためには, 初めに説明変数を用意する必要がある. 説明変数に異なる周波数である正弦波並びに余弦波をそれぞれ m ずつ, 合計 $2m$ 個用意する. なお, こちらは加藤らの手法 [12] に基づいている. 具体的には説明変数は以下のように記述することができる.

$$x_{ij} = \begin{cases} \cos(2\pi t_i \nu_j) & (1 \leqq j \leqq m) \\ \sin(2\pi t_i \nu_j) & (m+1 \leqq j \leqq 2m) \end{cases} \tag{5}$$

但し, t, ν は以下のように定義する. すなわち, 表現できる最大周波数が f_n である.

$$t_i = i, \quad \nu_j = j\frac{f_n}{m} \tag{6}$$

また Group Lasso を適用するために, グループセット $\mathcal{G}$ を定義する必要がある. 最終的に抽出したい情報は重要周期長であるため, 任意の t_i, ある周波数 ν_j に対して,

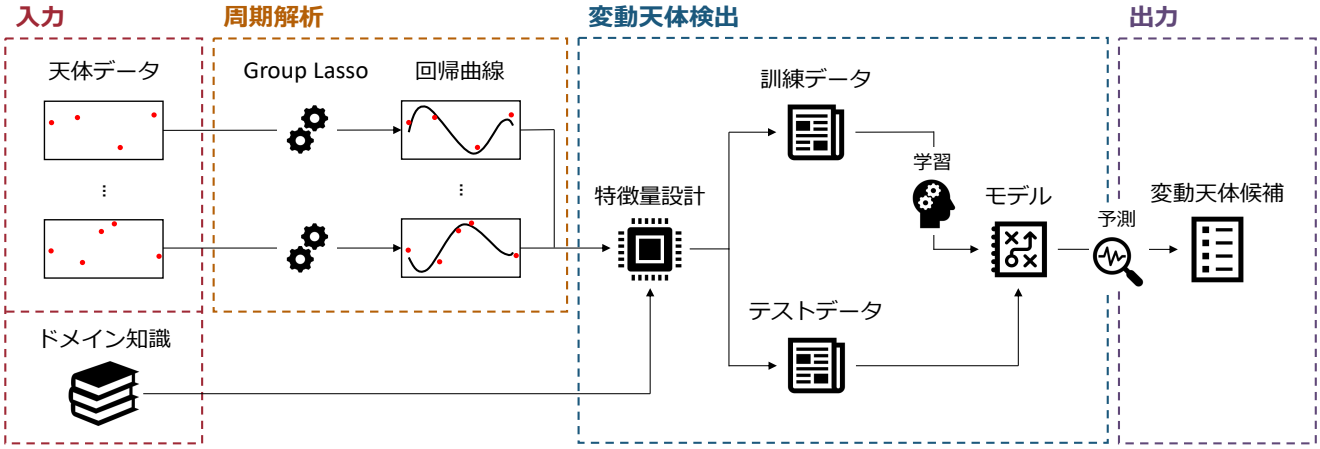


図 3: 提案手法全体像

Algorithm 1 PeriodicAnalysis

Input: i) 説明変数: X
ii) 目的変数: $\boldsymbol{y}$
iii) グループセット: $\mathcal{G}$
Output: i) 回帰係数: $\boldsymbol{w}$
ii) 決定係数: reg_score

1: // Group Lasso
2: $\boldsymbol{w} = \boldsymbol{0}$ // 初期化
3: **while** $\boldsymbol{w}$ is not converged **do**
4: $\hat{\boldsymbol{w}} \leftarrow \boldsymbol{w} - \eta\nabla f(\boldsymbol{w})$ // $\nabla f(\boldsymbol{w}) = X^\top(X\boldsymbol{w} - \boldsymbol{y})/n$
5: **for** $\boldsymbol{G_j}$ to $\mathcal{G}$ **do**
6: $\boldsymbol{w}_{\boldsymbol{G_j}} \leftarrow \max(0, 1 - \eta\lambda/||\hat{\boldsymbol{w}}_{\boldsymbol{G_j}}||)\hat{\boldsymbol{w}}_{\boldsymbol{G_j}}$
7: **end for**
8: **end while**
9: reg_score $= ||X\boldsymbol{w} - \overline{\boldsymbol{y}}||^2/||\boldsymbol{y} - \overline{\boldsymbol{y}}||^2$
10: **return** $\boldsymbol{w}$, reg_score

$\cos(2\pi t_i\nu_j), \sin(2\pi t_i\nu_j)$ の組み合わせで周期を表現する．言い換えると, $\cos(2\pi t_i\nu_j)$ のみが特徴量として採択され, $\sin(2\pi t_i\nu_j)$ は棄却されるといった結果として望ましくないということである．以上の条件を満たす結果を得るために, グループセット $\mathcal{G} = \{\{1, m+1\}, \{2, m+2\}, ..., \{m, 2m\}\}$ のように表現する [1].

3.1.1 アルゴリズム

周期解析の擬似コードを Algorithm1 に記載する．2 行目では, 零ベクトルにて $\boldsymbol{w}$ の初期化を行う． 3-8 行目では, Group Lasso の更新式に従い, $\boldsymbol{w}$ の更新を行う．特に, 4 行目の $\nabla f(\boldsymbol{w})$ は, 損失が小さくなる勾配を示している．また, 6 行目の $\max(0, 1 - \eta\lambda/||\hat{\boldsymbol{w}}_{\boldsymbol{G_j}}||)\hat{\boldsymbol{w}}_{\boldsymbol{G_j}}$ で 0 が採択されない値まで $\boldsymbol{w}_{\boldsymbol{G_j}}$ が大きくなった時, その $\boldsymbol{w}_{\boldsymbol{G_j}}$ に対応したグループをモデリングに重要な特徴量群とみなす．上記の更新を収束まで繰り返し, 収束後の $\boldsymbol{w}$ についての決定係数 reg_score を算出する．ここで得られた $\boldsymbol{w}$ 及び reg_score は以下で説明する分類タスクに使用する特徴量の一部として扱う．

1：$\{j, m+j\} \in \mathcal{G}$ は, それぞれ $\cos(2\pi t_i\nu_j), \sin(2\pi t_i\nu_j)$ と対応している．

3.2 変動天体検出

変動天体検出を実現するためには, 以後教師ラベルとして取り扱う変動天体識別子 objclass が, 0(非変動天体) か 1(変動天体) かを識別する分類タスクとして取り扱う必要がある．

3.2.1 特徴量について

こちらでは二値分類タスクにて実際に使用する特徴量リスト $\mathcal{F}$ に関して説明する．特徴量リスト $\mathcal{F}$ はドメイン特徴量リストである $\mathcal{F}_d$ を構成する 4 種類の特徴量, 回帰係数 $\boldsymbol{w}$, 決定係数 reg_score により構成されている．また, $\mathcal{F}_d$ とは, ドメイン知識を駆使し変動天体検出を実現するために採択に至った特徴量セットであり, 具体的には平均 m_ap40, z_apertureflux_40_mag, diff_ap40 から成る．以下では各要素について説明する [13].

$\boldsymbol{w}$ 周期解析により算出した天体データの各周波数候補に対応した周期関数の係数のベクトル．各係数の大きさは対応する周期長の重要度を表す．各要素は, 特徴量選択により採択された重要周期のみからなるスパースなベクトルである．

reg_score 周期解析により求めた回帰曲線が目的変数をどの程度表現出来ているか表す決定係数．周期解析の精度は特徴量 $\boldsymbol{w}$ の確からしさに相当するため, 変動天体検出に有効であると考えた．

平均 m_ap40 観測点に対応した天体の光度である m_ap40 を平均化した値．なお, m_ap40 とは, 観測結果として得られた画像上の対象天体を中心とした直径 4 秒角の円から得られる光量であり, 図 1 の赤いデータ点に対応している．この値は, 直径 4 秒角の円内に存在する光量を全てを積算して値を算出しているため, 近隣の影響により, 対象天体が持つ本来の光度とは離れた値が観測されてしまっている天体データも存在するという問題点が存在する．

z_apertureflux_40_mag 観測結果として得られた複数の画像を合成することで得られる Coadd 画像上の対象天体を中心とした直径 4 秒角の円から得られる光量．なお, Coadd 画像は, 対象天体が映っている全画像を画像的に足し合わせることで得られる．平均 m_ap40 と比較し近傍天体の明

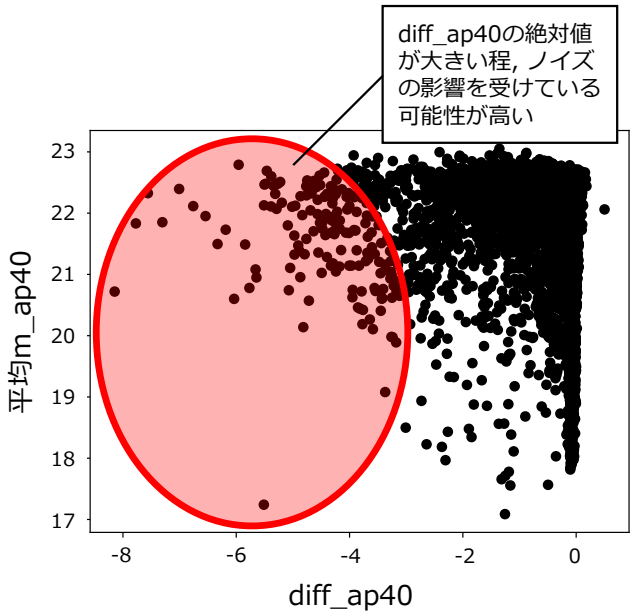


図 4: diff_ap40 と平均 m_ap40 の関係

るさの影響を軽減する等，ある程度のノイズの除去が成されているため，天体本来の明るさに近い値であることが期待される．

diff_ap40　平均 m_ap40 と z_apertureflux_40_mag の差．すなわち，天体が影響を受けているノイズの大きさに相当する特徴量．具体的には，平均 m_ap40, z_apertureflux_40_mag がどの程度異なる値であるかを示す値であるため，z_apertureflux_40_mag が天体本来の光度に近いと仮定すると，diff_ap40 の絶対値が大きい程，ノイズとして考えられる近傍の天体の明るさの影響が大きいことを表す．周期解析では，天体の光度（m_ap40）の変動情報を用いているため，平均 m_ap40 に影響を与えているノイズの大きさは変動天体の識別に影響していると考えられる．そのため，天体が影響を受けているノイズの大きさを表す本特徴量を用いている．

data_size　1 つの天体データに含まれる観測データ点の数．スパースモデリングを利用しているものの，観測回数が少ない一定数の天体データについては過学習を引き起こしてしまう．従って，周期解析結果が過学習を引き起こしている蓋然性に相当する数値だと考え，特徴量として利用することにした．

以下では，各特徴量間の関係性について説明を行う．図 4 では，diff_ap40 と平均 m_ap40 の関係を表している．この図から，diff_ap40 の絶対値が極端に大きい値を持つ天体 (図の左側) の多くが暗い天体であることがわかる．故に，暗い天体の方が外部からの影響をより受けやすいという背景が読み取れる．このような背景を変動天体検出に反映させることでより高度な分類を実現する．

3.2.2　アルゴリズム

提案モデル全体を表現した擬似コードを Algorithm2 に記載

Algorithm 2 VariableStarsExtraction

Input: i) データセット: $\mathcal{Y}$
ii) 周波数候補数: m
iii) ナイキスト周波数 (表現可能な最大周波数): f_n
iv) ドメイン特徴量リスト (各天体が持つ固有の天体情報): $\mathcal{F}_d$
Output: 変動天体リスト: $\mathcal{L}$

1: // 変数設定
2: **for** $i = 1$ to n **do**
3:　**for** $j = 1$ to m **do**
4:　　// $t_i = i, \quad \nu_j = j\frac{f_n}{m}$
5:　　$X[i][j] \leftarrow \cos(2\pi t_i \nu_j); \quad X[i][j+m] \leftarrow \sin(2\pi t_i \nu_j)$
6:　**end for**
7: **end for**
8: **for** $j = 1$ to m **do**
9:　$\mathcal{G}[j] \leftarrow [j, j+m]$
10: **end for**
11: // 周期解析
12: **for** $i = 1$ to N **do**
13:　$*w$, reg_score $\leftarrow$ Algorithm1$(X, \mathcal{Y}[i], \mathcal{G})$
14:　$\mathcal{F}[i] \leftarrow [*w$, reg_score, $\mathcal{F}_d[i]]$
15: **end for**
16: // 二値分類
17: $\mathcal{F} \rightarrow X_{train}, X_{test}, y_{train}, y_{test}$
18: lgb = LightGBM.train(X_{train}, y_{train})
19: $\mathcal{L}$ = lgb.predict(X_{test})
20: **return** $\mathcal{L}$

する．2-11 行目では，必要な変数を作成する．具体的には，2-8 行目では説明変数 X を，9-11 行目ではグループセット $\mathcal{G}$ を作成している．その後，13-16 行目では，周期解析を実行し，それらの結果及び $\mathcal{F}_d$ から二値分類の際に必要な特徴量リスト $\mathcal{F}$ を作成する．最後に 18-20 行目では，本章で説明した二値分類を実行している．特に 18 行目で実行している訓練データ及びテストデータへの分割については，テストデータに教師ラベルが振り分けられていないデータセットを用いることで，変動天体の予測に活用することが可能である．

4　実　　験

本章では，実際の天体データを利用した実験を通じて，既存手法との比較を行うことにより，提案手法がより高精度な変動天体検出が可能であることを示す．また，Group Lasso を活用した周期解析により正しい周期情報の算出を行えていること，正しいドメイン知識を特徴量エンジニアリングできていることを明確化する．加えて，提案手法がノイズが混入したデータに対しても正しい周期解析が実行できることを示す．

4.1　実 験 設 定

4.1.1　データセット

本研究では，実天体データセットとして PDR2[2](Public Data Release 2) [13] を利用した．このデータセットのカラムは，天

2：https://hsc-release.mtk.nao.ac.jp/doc/

体識別子 objectid, 観測時刻 mjd, 天体の光度 m_ap40, 変動天体識別子 objclass により構成される．また PDR2 に属する天体データ全てに対して objclass が割り振られているわけではなく, 既に割り振られているデータの総数は計 1,557 件である．これらは, 既存手法の一つである B.Sesar らの手法 [2] で上位を占めるデータであり, 変動天体は約 26.1%を占めている．また, 4.2.2 節では, 周期解析のノイズ及びスパースに対するロバスト性を検証するために人工データを扱った実験を行った．具体的なデータの含有量については, 最小で 0.01%[3]ほどである．なお, 本稿におけるデータの含有量とは, 時系列全体に対して欠損していない観測データ点の割合を示す.

4.1.2 精度評価

ベースラインとして B.Sesar らの手法を利用する．こちらは現状国立天文台の方で変動天体検出のために使用されている手法であるため, ベースラインとして妥当であると考えた．但し, この手法は教師あり学習による手法でないため, 提案モデルと同等の方法である k 分割交差検証を用いた精度評価ができない．そのため, データセット全体を占める変動天体の割合が約 26.1%であることから, 二値分類タスクによって得られる予測結果を占める変動天体の割合も同等であると仮定すると, ベースライン手法によってスコアリングした結果の上位 26.1%を変動天体候補として考えることによって, 比較的公平な条件で精度評価が可能であると考えた．また, 本稿で扱う 1,557 件のデータは光度が相対的に高いレコードの集合であるため, ノイズの影響は無視しても良い程度である．故に, 主要因である標準偏差によりスコアリングを行う.

4.2 実験結果

4.2.1 実天体データへの適用

表 5 は, objclass 割り振られている全データ計 1,557 件に対して, 4.1.2 節で仮定した条件のもと, 既存手法及び提案手法を適用することで得られた評価値 AUC, accuracy, precision, recall, F1 をまとめたものである．提案手法は, ベースライン手法と比較し, 大幅な性能向上を示す結果が見られる．また, スパースモデリング及びドメイン知識の有効性を示すために, それらに関する対照実験も行った．結果は, 線形回帰を使用しているモデルや, ドメイン知識を活用していないモデルと比較し, スパースモデリングの一種である Group Lasso を使用し, かつ, ドメイン知識を活用したケースが最も高精度に変動天体を検出できていることを示すことができた．ここから, 特徴量選択を正しいグループ単位で行い, 得られた周期情報をうまく活用できていることがわかる．また, 周期解析を行う際, Lasso を使用したモデルと Group Lasso を使用したモデルの性能差がほとんど見られない．こちらは, 今回はベースライン手法で得られた結果の上位を占めているデータセットを使用しているため, 特徴量のグループ化を行わずともほとんど望ましい結果が得られたためであると考える．ベースライン手法で取得しきれなかった天体データも含めて同様の実験を行うと, より性能の差が現れると推察する.

表 5: 各評価指標でのモデル評価

モデル	AUC	accuracy	precision	recall	F1
B.Sesar 手法	0.750	0.743	0.507	0.507	0.507
提案手法					
線形回帰 + 無 [4]	0.633	0.729	0.432	0.109	0.168
線形回帰 + 有	0.846	0.820	0.702	0.538	0.605
Lasso + 無	0.854	0.825	0.769	0.478	0.586
Lasso + 有	0.937	0.897	0.815	0.780	0.795
Group Lasso + 無	0.860	0.828	0.793	0.476	0.590
Group Lasso + 有	**0.939**	**0.899**	**0.820**	**0.785**	**0.801**

4.2.2 周期解析のノイズ及びスパースに対するロバスト性の検証

初めに定性的な評価を行う．図 6 は, 用意した人工データ (a), (b) から約 0.01%[5]ほどのデータ点を抽出し作成したスパースなデータに対して線形回帰及び Group Lasso を利用し, 周期解析を行った結果を描画したものである．(c), (d) が Group Lasso により得られた結果であり, (e), (f) が線形回帰により得られた結果である．使用データ量が極端に少ないため, 線形回帰を活用し算出した回帰曲線は, 技術的課題点に挙げた通り過学習が生じている．また, 特徴量選択が行われていないため, 変動天体の重要周期を表現できていないことも読み取れる．一方, Group Lasso により得られたパワースペクトルからは重要特徴量が一点に集約されていることが汲み取れる．加えて, 正解データと Group Lasso により得られた回帰曲線が類似していることも明白である.

続いて定量的な評価を行う．図 7 は使用データ点の含有率と, 線形回帰及び Group Lasso を使用することで得られる回帰曲線と正解データの整合率の関係を表した折れ線グラフである．評価指標は, 時系列間の周期性の整合度合いを計測可能なコサイン類似度 [14] を用いた．他の評価値として, DTW(Dynamic Time Warping) [13] が存在するが, こちらは時系列間の周期ではなく距離に基準を設けているため, 本実験にはそぐわない評価値であると考えた．実験結果に着目すると, 線形回帰を利用した周期解析では, データの含有率が 1%を下回ったあたりから, 整合度が急激に低下し, 誤差も大きくなっている．但し, 提案手法では, データの含有量が実際の天体データの平均的な欠損値割合である 0.01%程になったとしても, 整合度はほぼ 1 を保っている．また, 誤差も小さく安定している．これらの結果から, 提案手法の周期解析にロバスト性があるということが言える.

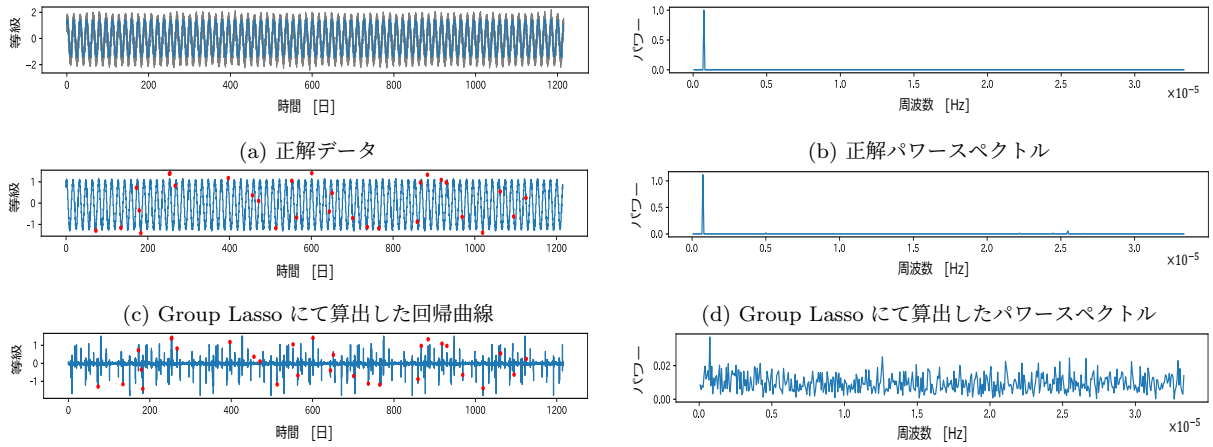
4.2.3 教師ラベルが未定な天体データへの応用

図 8 は教師ラベルが未定なレコードが含まれた天体データ計 5,478 件に対して提案モデルを適用し, 得られた変動天体候補リストの上位 1,000 件を自然科学研究機構国立天文台の方々に調査していただいた結果を, 横軸が上位 k 件, 縦軸がその k 件内の変動天体の割合という形で描画したものである．但し, 学習に用いたデータをテストデータにも含めた形式で予測を行っ

3：天体データの平均的な欠損値数に基づいている

4：有, 無とはそれぞれドメイン知識を活用したか否かを示している.

5：天体データの平均的な欠損値数に基づく



(a) 正解データ　(b) 正解パワースペクトル

(c) Group Lasso にて算出した回帰曲線　(d) Group Lasso にて算出したパワースペクトル

(e) 線形回帰にて算出した回帰曲線　(f) 線形回帰にて算出したパワースペクトル

図 6: 周期解析結果の比較

(a) の人工データを元に周期解析を行い (c)-(f) を算出した．(c), (e) に含まれている赤点が (a) の内, 周期解析に使用したデータ計 30 点である．また, (a) のグラフに描かれている灰色部は元のデータに付与したノイズである．

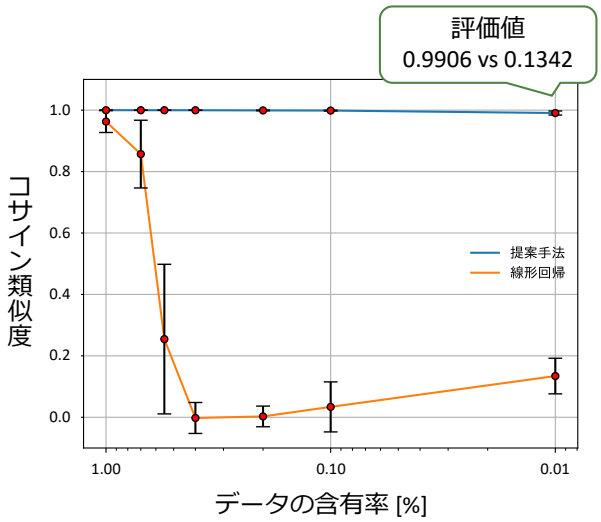


図 7: データ含有量と正解データとの整合度の関係

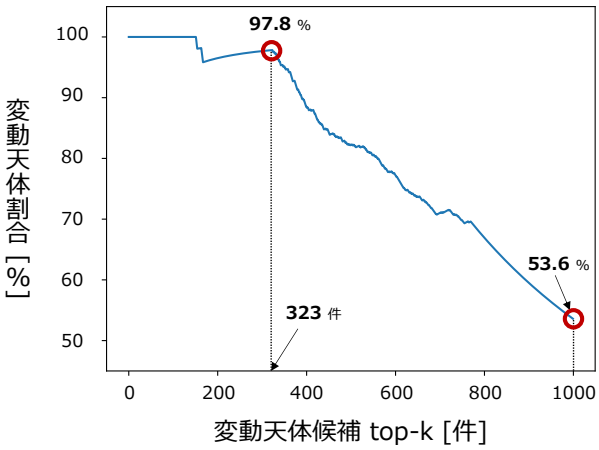


図 8: 候補リスト上位 k 件を占める変動天体の割合

た．このような実験設定にした主な理由は, 訓練データから得られるスコアという基準と予測対象の差, すなわち, 変動天体である確率に関連した情報を得ることができるためである．図より, 上位 323 件内の変動天体の割合は約 97.8%である．この区分の大半は学習に使用した変動天体データによって構成されているので, ここに属する教師ラベルが未定なデータに関しては高確率で変動天体であることが言える．また, 上位 1,000 件内の変動天体の割合は約 53.6%である．但し, 予測対象全件の教師ラベルが明確化されているわけではないため, データセットに属する全変動天体の内, どの程度検出したかどうかが不明である点に注意していただきたい．ところで, 4. 1. 1 節で説明した通り, ベースライン手法によって取得した上位 1,557 件内の変動天体の割合は約 26.1%である．それを踏まえると, 上位 1,000 件を確認するだけで 200 件ほど追加で変動天体の検出できたことがわかる．また, 一般的には訓練データとテストデータの割合が 7:3 程度であるのに対して, 本実験では約 3:7 という訓練データが十分でない環境であったため, 適切な割合で提案モデルを活用することでより高度な検出が期待できる．

5 関 連 研 究

5. 1 変動天体検出

変動天体の検出に関する代表的な手法として, B.Sesar らによる手法 [2] が挙げられる．これは, 提案モデルのような機械学習によるアプローチとは異なり, intrinsic variability σ, χ^2, 等級の測定値の平均, 標準偏差, 歪度などといった統計的数値を活用したスコアリングを行うことで変動天体の検出を行っている．変動天体かどうかの閾値は主に σ である．しかし, 天体の持つ明るさが暗ければ暗いほどノイズの影響を受けやすく, 例え σ が基準を満たしていたとしても, それが変動天体固有の特徴によるものか, ノイズの影響によるものかが判断つかない．このようなケースに対応するため, χ^2 を用いて解決している．

こちらの手法では変動天体の持つと言われている周期情報を活用できていないため, 与えられたデータセットを満足に活用できていない.

5.2 周期解析

周期解析とは, 正弦波や余弦波など周期性を持った関数を基底とし, 入力データを近似・分析を行う手法のことである. また, 今回必要な情報は入力データに潜む重要周期である. CLEAN [15] は雑多なパワースペクトル情報から重要な周期長の抽出を可能にする. しかし, 入力データから重要周期を直接計算できるわけではない. また, lomb-scargle periodogram [5, 6] では, 最小二乗フィッティングを活用することによって特定の周期長に対応したフーリエ振幅を計算するピリオドグラム法の一種である. サンプリング間隔が一定でない時系列に対しても適用可能である反面, ノイズの影響が大きい場合や欠損値がある程度含まれた時系列からは望ましい結果が得られない. また, 複数の周期性が混在したような時系列に対して適用した場合, 複数の周期性の内一つしか取得できない場合や, 存在する周期性が互いに影響を及ぼし合った結果存在しない周期性が誤検出される場合がある, といった問題点が挙げられる.

6 おわりに

本稿では, 周期性に着目した変動天体の検出手法を提案した. 提案モデルは, 変動天体が持つとされている周期性に着目し, 特定の周波数を指定することなく, 加えて, スパースモデリングの恩恵により入力データに潜むノイズの影響も軽減し, 変動天体を検出することが可能である. 実験では, 実際の天体データに対しても既存手法と比べてより高精度な結果が得られることを示した. また, ノイズを加えたスパースな人工データに対しても, 正確な周期解析を行うことが可能なことを明確にした. 今後の課題としては, 並列処理を導入した高速化の実現や他のスパースモデリングを活用したモデルの性能改善などが考えられる.

7 謝辞

この成果は，国立研究開発法人新エネルギー・産業技術総合開発機構（ＮＥＤＯ）の委託業務（JPNP16007) の結果得られたものです.

文献

[1] AC Becker, JJ Bochanski, SL Hawley, Ž Ivezić, AF Kowalski, B Sesar, and AA West. Periodic variability of low-mass stars in sloan digital sky survey stripe 82. *AJ*, 731(1):17, 2011.

[2] Branimir Sesar, Željko Ivezić, Robert H Lupton, Mario Jurić, James E Gunn, Gillian R Knapp, Nathan De Lee, J Allyn Smith, Gajus Miknaitis, Huan Lin, et al. Exploring the variable sky with the sloan digital sky survey. *AJ*, 134(6):2236, 2007.

[3] Eran O Ofek, Maayane Soumagnac, Guy Nir, Avishay Gal-Yam, Peter Nugent, Frank Masci, and Shri R Kulkarni. A catalogue of over 10 million variable source candidates in ztf data release 1. *MNRAS*, 499(4):5782–5790, 2020.

[4] Lorenzo Rimoldini, Berry Holl, Panagiotis Gavras, Marc Audard, Joris De Ridder, Nami Mowlavi, Krzysztof Nienartowicz, Grégory Jevardat de Fombelle, Isabelle Lecoeur-Taïbi, Lea Karbevska, et al. Gaia data release 3: All-sky classification of 12.4 million variable sources into 25 classes. *arXiv preprint arXiv:2211.17238*, 2022.

[5] Nicholas R Lomb. Least-squares frequency analysis of unequally spaced data. *Astrophysics and space science*, 39(2):447–462, 1976.

[6] Jeffrey D Scargle. Studies in astronomical time series analysis. ii-statistical aspects of spectral analysis of unevenly spaced data. *AJ*, 263:835–853, 1982.

[7] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[8] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[9] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017.

[10] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *AS*, pages 1189–1232, 2001.

[11] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on KDD*, pages 785–794, 2016.

[12] Taichi Kato and Makoto Uemura. Period analysis using the least absolute shrinkage and selection operator (lasso). *PASJ*, 64(6), 2012.

[13] Hiroaki Aihara, Yusra AlSayyad, Makoto Ando, Robert Armstrong, James Bosch, Eiichi Egami, Hisanori Furusawa, Junko Furusawa, Andy Goulding, Yuichi Harikane, et al. Second data release of the hyper suprime-cam subaru strategic program. *PASJ*, 71(6):114, 2019.

[14] Anna Huang et al. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, volume 4, pages 9–56, 2008.

[15] David H Roberts, Joseph Lehár, and John W Dreher. Time series analysis with clean-part one-derivation of a spectrum. *AJ*, 93:968, 1987.

[16] Ernest K Ryu and Wotao Yin. Proximal-proximal-gradient method. *arXiv preprint arXiv:1708.06908*, 2017.

付録

A 補題 1 の証明

証明. E_{Lasso} を $\boldsymbol{w}$ について微分を行った結果は (A.1) 式のようになる.

$$\frac{\partial E_{Lasso}}{\partial \boldsymbol{w}} = \frac{1}{n} X^{\top}(X\boldsymbol{w} - \boldsymbol{y}) + \lambda sign(w_k) \qquad \text{(A.1)}$$

よって, E_{Lasso} を w_k について微分を行った結果は (A.1) 式で得られたベクトルの k 番目を考えれば良いため,

$$\frac{\partial E_{Lasso}}{\partial w_k} = \frac{1}{n}\sum_{i=1}^{n} x_{ik}(\sum_{j=1}^{m} w_j x_{ij} - y_j) + \lambda sign(w_k) \quad \text{(A.2)}$$

E_{Lasso} を最小化する w_k を計算するためには, E_{Lasso} が凸関数であるため, (A.2) 式が 0 になるケースを考えれば良い. このようなケースに対して, w_k を計算すると, (A.3) 式のように

なる.

$$w_k = \frac{\sum_{i=1}^{n}(y_i - \sum_{j\neq k}^{m} w_j x_{ij} - w_0)x_{ik} - n\lambda sign(w_k)}{\sum_{i=1}^{n} x_{ik}^2} \tag{A.3}$$

これらの計算により E_{Lasso} を最小化する $\boldsymbol{w}$ を計算できた. 但し, (A.3) 式は $-n\lambda < w_k < n\lambda$ について w_k を定義できない. すなわち, この範囲においては微分不可能であることを意味している. ここで劣微分の考えを適用すると, この範囲において $w_k = 0$ と考えても良いため, 証明が完了した.

□

B 補題 2 の証明

証明. 近接勾配法 [16] を活用する. これは, 微分不可能点が含まれている関数に対して勾配法を適用できるように改良した手法である. 近接勾配法の更新は以下の規則に従い行われる.

$$\begin{aligned} x_{i+1} &= prox_{\eta g}(x_i - \eta \nabla f(x_i)) \\ prox_g(y) &= \underset{x}{\operatorname{argmin}}\{g(x) + \frac{1}{2}||x-y||^2\} \end{aligned} \tag{B.1}$$

Group Lasso の損失関数である $E_{GroupLasso}$ を (B.1) に従い変形すると, 以下のように書き下せる.

$$\begin{aligned} \boldsymbol{w_{i+1}} &= \underset{w}{\operatorname{argmin}}\{h(\boldsymbol{w})\} \\ h(\boldsymbol{w}) &= \nabla f(\boldsymbol{w_i})^\top (\boldsymbol{w} - \boldsymbol{w_i}) + \frac{1}{2\eta}||\boldsymbol{w} - \boldsymbol{w_i}||^2 + \lambda g(\boldsymbol{w_i}) \\ f(\boldsymbol{w}) &= ||\boldsymbol{y} - X\boldsymbol{w}||^2, g(\boldsymbol{w}) = \sum_{j=1}^{J} ||\boldsymbol{w_{G_j}}|| \end{aligned}$$

すなわち, $h(\boldsymbol{w})$ を最小化する $\boldsymbol{w}$ の内, $\boldsymbol{w} = \boldsymbol{0}$ となる $\hat{\boldsymbol{w}}_{\boldsymbol{G_j}}$ の範囲が $||\hat{\boldsymbol{w}}_{\boldsymbol{G_j}}|| < \eta\lambda$ であることを示すことができれば良い. 続いて, あるグループ $\boldsymbol{G_j}$ に対して, $\boldsymbol{w_{G_j}}$ に関する $h(\boldsymbol{w})$ の微分を考えると, 以下の式が成立する.

$$\begin{cases} \boldsymbol{w_{G_j}} = \hat{\boldsymbol{w}}_{\boldsymbol{i}\boldsymbol{G_j}} - \eta\lambda \dfrac{\boldsymbol{w_{G_j}}}{||\boldsymbol{w_{G_j}}||} & (\boldsymbol{w_{G_j}} \neq \boldsymbol{0}) \\ ||\hat{\boldsymbol{w}}_{\boldsymbol{i}\boldsymbol{G_j}}|| \leqq \eta\lambda & (\boldsymbol{w_{G_j}} = \boldsymbol{0}) \end{cases} \tag{B.2}$$

(B.2) 式を解くことで, $h(\boldsymbol{w})$ を最小化する $\boldsymbol{w_{G_j}}$ を求めることができる. 変形することで以下の式が得られる.

$$\boldsymbol{w_{G_j}} = \max(0, 1 - \frac{\eta\lambda}{||\hat{\boldsymbol{w}}_{\boldsymbol{i}\boldsymbol{G_j}}||})\hat{\boldsymbol{w}}_{\boldsymbol{i}\boldsymbol{G_j}} \tag{B.3}$$

(B.3) より, $1 - \frac{\eta\lambda}{||\hat{\boldsymbol{w}}_{\boldsymbol{i}\boldsymbol{G_j}}||} < 0$ の場合, $\boldsymbol{w_{G_j}} = \boldsymbol{0}$ となり, その範囲が $||\hat{\boldsymbol{w}}_{\boldsymbol{G_j}}|| < \eta\lambda$ であるため, 証明が完了した. □