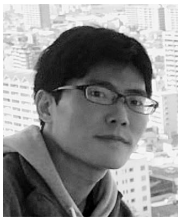
系外惑星探査におけるデータ科学手法



平　野　照　幸
〈東京工業大学理学院　〒152-8551　東京都目黒区大岡山 2-12-1〉
e-mail: hirano@geo.titech.ac.jp

近年系外惑星探査では，観測データからできるだけ正確，精密な情報を取り出す方法論の開発そのものが重要なテーマとなっている．これは，例えば高精度な視線速度観測・トランジット観測では恒星活動や装置に由来する見かけ上の視線速度・フラックス変化が観測精度の何倍も大きいことがしばしばあり，これらの考慮が系外惑星の発見，特徴づけには不可欠であるためである．本稿では，こうした高精度な観測に見られる白色ノイズ以外の変動（相関ノイズ）を扱う方法論の一つとして「ガウス過程」を用いた機械学習的なアプローチを解説する．ガウス過程を用いることで観測データの補間（予測）や相関ノイズの補正をノンパラメトリックに行うことができることを示し，これを実際の視線速度データやトランジットデータの解析に応用した例を紹介する．

1.　はじめに

最初の系外惑星が太陽型星のまわりで発見されてからすでに20年以上が経過し，近年では系外惑星探査における対象，発見方法，その特徴付けの手法が極めて多様化している．代表的な系外惑星探査には，惑星の重力による中心星（恒星）のふらつきをスペクトル線のドップラーシフトとして検出する「視線速度法」，恒星面を惑星が横切る際の恒星の減光を検出する「トランジット法」，惑星を持つ恒星の重力レンズ効果によって背景星の増光パターンから惑星を検出する「マイクロレンズ法」などがあるが，最近ではこれらの各手法でも対象とする天体や観測を実施する波長帯の細分化が進んでいる．

データ科学的な方法論を用いたアプローチは観測・理論を問わず天文学で一般的に用いられるようになってきており，系外惑星研究においても古くは検定やベイズ統計を用いた回帰分析，最近では機械学習等がしばしば利用されている．特に系外惑星の観測的研究では，データの解釈において必ずと言っていいほど何らかのデータ科学的手法が用いられている．

本稿では「データ科学」シリーズの特集記事の一つとして「系外惑星探査」におけるデータ科学的手法を取り上げる．次節以降，系外惑星観測の現状を報告するとともに，系外惑星探査を観測的に制約する点（2章），その問題に対する機械学習的なアプローチを紹介する（3–5章）．なお本稿では筆者が特に専門とする「視線速度法」と「トランジット法」を用いた系外惑星探査を主として取り上げるが，本稿で紹介する方法論はこれらの探査法に限らず応用可能であり，今後広く普及していくものと期待される．観測的宇宙論等，系外惑星探査以外への応用については「データ科学」シリーズのほかの記事を参照されたい[1]．

2.　系外惑星探査はさらなる精密な観測へ

系外惑星探査における対象・観測手法はこの20年で非常に多様化しているが，その一方で惑星探査のフロンティアを制約する点はあまり変化

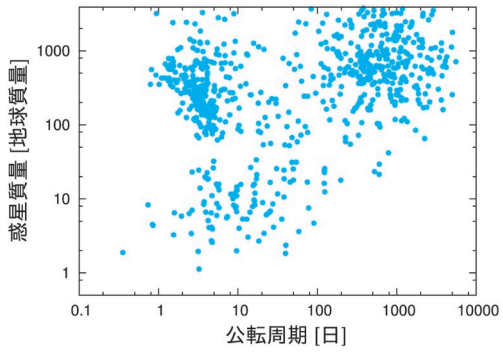


図1　系外惑星の質量と公転周期の分布．トランジットする惑星以外は惑星質量の下限値（$m_p \sin i$）．

していない．すなわち，1）暗い恒星（地球から遠くにある恒星），2）活動度の高い恒星，3）小さい惑星，4）長周期の惑星*1は常に系外惑星探査において困難な対象とされてきた．図1は，これまで見つかっている系外惑星のうち，惑星質量（の下限）が測定されているものを惑星周期の関数としてプロットしている[2]．一般に視線速度法・トランジット法による系外惑星探索の検出限界は，観測精度・観測期間の改善に伴って左上の領域（短周期の巨大惑星）から徐々に右下の領域（長周期の小型惑星）へと遷移する．現在達成可能な視線速度測定精度（～1 m s^{-1}）では地球のように365日程度で公転する地球型惑星の発見は困難であるが（視線速度振幅は10 cm s^{-1}以下），近い将来10 mクラスの望遠鏡に搭載された視線速度観測に特化した分光器（例：VLT/ESPRESSOなど）によりこれまで困難であったパラメータ領域の惑星探査が進むと期待されている．

一方，観測精度以外にも系外惑星の発見，特徴付けを阻む要因はいくつかある．視線速度法・トランジット法においては「恒星表面の活動」がその代表例である．よく知られているように，恒星表面に存在する黒点・白斑等の明るさの「むら」は恒星の自転に伴って見かけ上の視線速度変動を引き起こす．このような恒星活動による見かけ上の視線速度変動は「RVジッター」と呼ばれ，特に若い恒星などで顕著である．図2は系外惑星探査に用いられる代表的な可視高分散分光器ケックI望遠鏡/HIRESとTNG/HARPS-Nで測定された地球型トランジット惑星をもつ恒星Kepler-78の視線速度変動であるが[3]，惑星のサイズ（1.1–1.2地球半径）から予想される視線速度変化（5 m s^{-1}以下）に比べて明らかに大きな変動を示している．これはRVジッターが主な原因と考えられ，Kepler-78のような比較的活動度の高い恒星では，こうしたRVジッターにより地球型惑星の発見確認が困難となる．

トランジット探査においても地球サイズの小型惑星の検出や確認は難しく，小型惑星の観測には光度曲線に影響を及ぼす恒星活動や装置由来の微小な効果を考慮しなくてはならない．例えば，トランジット観測で太陽型星の周りに地球型惑星を検出するにはフラックス比で10^{-4}程度の減光を捉える必要があるが，一般に恒星活動によるフラックス変動は同程度以上であることがほとんどである*2．また特に地上からの測光観測では一般にシグナルノイズ比（SN）に対応する白色ノイズよりも，1）星像の検出器上での位置の変化，2）シーイングや大気エアマスの変化，3）背景光強度の変化，等に起因した「相関ノイズ」が支配的になることが多い．こうした効果はいずれも小型惑星の発見，追観測において極めて大きな障害となり地上からのトランジット探査を特に困難なものにしている．

*1 特に2）と4）は視線速度法とトランジット法において障害となる．マイクロレンズ法，直接撮像などによる惑星探査は比較的長周期の惑星にも感度がある．

*2 ただしトランジットと恒星活動によるフラックス変動は異なるタイムスケールであることが多いため一般に視線速度観測に比べて区別しやすい．

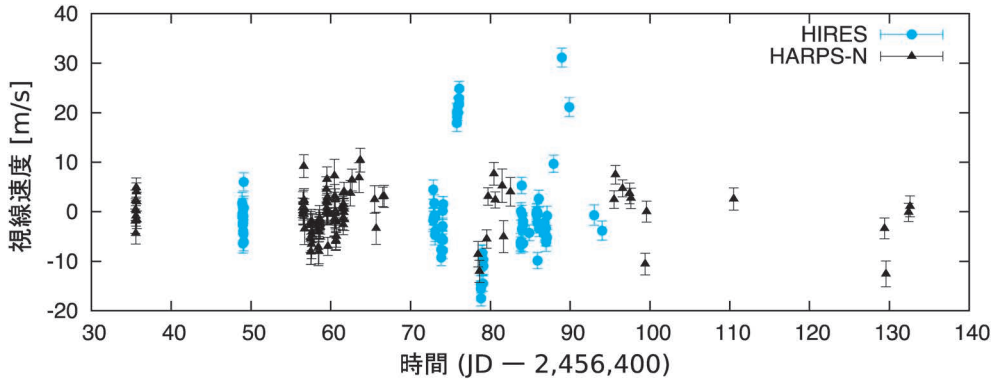


図2　ケックI望遠鏡/HIRES，TNG/HARPS-Nで取得されたKepler-78の視線速度変動．

3.　ガウス過程を用いたデータモデリング

望遠鏡の大型化，観測装置の高度化が進むにつれて当然高精度な測光・分光観測が可能となり系外惑星探査のフロンティアは広がっていくが，上述したような恒星活動等に由来した見かけ上の変動もより鮮明に観測されるようになるため，それらをいかに取り除くかが特に高精度な観測においては重要な鍵となる．幸いなことに，それらは原理的に取り除くのが不可能な白色ノイズとは異なり場合によって原因がわかっている相関ノイズであるため，理論的あるいは経験的な補正を与えることが可能である．

ここでは相関ノイズを取り扱う方法論の一つとして，「ガウス過程」を用いた機械学習法による回帰分析を紹介する．ガウス過程（Gaussian Processes）とは一般にある標本関数が「多次元の正規分布」に従う際の確率過程の総称である．ガウス過程の基本とその応用についてはいくつか有名な文献があるのでそちらを参照されたい[4),5)]．特にガウス過程を用いた機械学習による回帰分析ではデータをモデルに当てはめる際の尤度関数$\mathcal{L}$が多次元正規分布となると仮定をする．すなわち，モデルを当てはめる観測データベクトル（例えば，フラックス値だったり視線速度値の時系列）$\boldsymbol{f}$が

$$\mathcal{L}=\mathcal{N}(\boldsymbol{\mu},\Sigma)\equiv\frac{1}{\sqrt{(2\pi)^{N_{\rm data}}|\Sigma|}}\times\exp\left\{-\frac{(\boldsymbol{f}-\boldsymbol{\mu})^{T}\Sigma^{-1}(\boldsymbol{f}-\boldsymbol{\mu})}{2}\right\}\quad(1)$$

に従うと仮定する．ここで$N_{\rm data}$は回帰を行うデータの数，$\boldsymbol{\mu}$は回帰するモデルを表すベクトル，Σはデータ点の共分散行列（大きさは$N_{\rm data}\times N_{\rm data}$）である．ここで共分散行列が仮に成分同士無相関な対角成分（独立同分布）のみをもつ場合は

$$\mathcal{L}\propto\exp\left(-\frac{\chi^2}{2}\right),\ \chi^2=\sum_i\frac{(f_{{\rm obs},i}-f_{{\rm model},i})^2}{\sigma_i^2}\quad(2)$$

となり，通常のχ^2に帰着する．ただし$f_{{\rm obs},i}$と$f_{{\rm model},i}$は$\boldsymbol{f}$と$\boldsymbol{\mu}$の各成分，σ_iは$f_{{\rm obs},i}$のエラーで白色ノイズに対応する．

ガウス過程が特に有用となるのは，共分散行列Σに非対角成分を導入することであるデータ点と異なるデータ点の相関を定量的に尤度に反映させることが可能となる点である．例えば，異なる時刻の観測点同士の相関を導入すれば，時間に相関した「赤色ノイズ」のモデル化につながる．共分散行列Σの関数形は任意に仮定できるため，この中に離れた観測点との相関や周期性などを導入することが可能である．さらに，観測データ$\boldsymbol{f}$の成

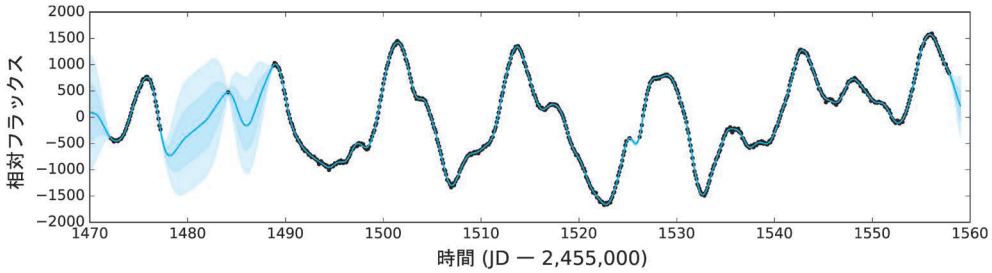


図3　ケプラー宇宙望遠鏡で観測されたKepler-78のフラックス値（黒点）とそのガウス過程による回帰．青の実線とその周りの影付きの領域は，観測データを元に機械学習したテストデータ（各時刻）のフラックス事後確率分布の平均と分散に対応する（濃い領域，薄い領域がそれぞれ1σと2σ）[3]．

分同士の直接的な相関以外にも補助的なパラメータを介して相関をもたせることも可能である．次節以降，これについていくつか系外惑星の観測データの解析例を示しながら具体的に見てみる．

4.　視線速度解析へのガウス過程の応用

ガウス過程の応用例として，相関ノイズ（RVジッター）を含む視線速度データの解析を取り上げ，実際にガウス過程が観測データの解釈においてどのように役立つのかを見てみる．ここでは2節で紹介したKepler-78系に焦点を当て，Grunblattら[3]によって紹介されている手順に従って話を進める．

RVジッターの影響を評価する上で鍵となるのが，RVジッターは主に恒星表面の活動によって引き起こされており恒星活動は視線速度以外にもフラックス変化や活動度の指標となるスペクトル線にも影響を与えるという点である．図3にケプラー望遠鏡で取得されたKepler-78の光度曲線（黒点）の一部をプロットしてある．フラックスの各点は隣り合う（近くの）点と強い相関を持ち，さらにフラックスはおよそ13日程度で周期的に変化していることがわかる．さらに各ピーク周辺の形状は時間とともにゆっくりと変化している．こうした変化は恒星表面の明るさのむらが恒星の自転に伴って動いていることに対応しており，さらに形状の変化は黒点等の形状・配置が時間的に発展している様子を表していると考えられる．これらのフラックス変動は視線速度と強く相関していることが知られており，Kepler-78の場合図2に見られるようなRVジッターとして観測される．このRVジッターの影響を取り除くには直接フラックス値と視線速度値の関係式を求めて（例えば線形回帰など）経験的に補正すれば良いと思われるかもしれないが，図3に見られるように黒点は時間とともに発展するためフラックスのピークとRVジッターのピークは必ずしも対応せず，さらにその位相も時間的に変化する．

ガウス過程を用いればこうした相関ノイズのモデル化を具体的な関係式を仮定することなく行うことができる．ただし図2と3について注意が必要なのが，二つの観測量（ケプラー望遠鏡によるフラックス値と地上望遠鏡による視線速度値）が同じ時刻に取得されていないという点である．そのため，ここではガウス過程を用いてまず光度曲線から変動の周期性や相関のタイムスケールを制限して，次にそれらを用いてRVジッターを含む視線速度データをモデル化するという手法をとる．ガウス過程においてデータ点のモデルへの当てはめを行う際の尤度は共分散行列の成分を指定することで求まるが，共分散行列の関数形（カーネルと呼ぶ）は対象とするデータの特徴によってさまざまな関数形を採用する．ここでは，Grunblattら[3]に従ってフラックス（時刻t）に対する共分散行列（i, j成分）として，“quasi-periodic”

カーネル

$$\sum_{ij} = h^2 \exp \times \left[-\frac{\sin^2\{\pi(t_i - t_j)/\theta\}}{2w^2} - \left(\frac{t_i - t_j}{\lambda}\right)^2 \right] \quad (3)$$

を採用する．ただしhは時間的な相関ノイズの振幅，θはフラックス変動の周期（恒星自転周期に対応），λは周期性以外のフラックス相関の特徴的なタイムスケール（黒点の減衰時間に対応），wはコヒーレンススケールを表す．これらのカーネルに登場するパラメータは「ハイパーパラメータ」と呼ばれ，相関の強さやスケールを表すのに用いられることが多い．なおこのカーネルはKepler-78のフラックス変動が，1）恒星の自転に対応した周期性をもつ，2）周期性は厳密ではなくコヒーレント成分があるタイムスケールで減衰する，という二つの物理的考察から導き出された関数形である．

具体的なカーネルが決まったら，あとはデータ点からハイパーパラメータ（ベクトル$\boldsymbol{\theta}$とする）を決定する．このプロセスにはいくつか方法があるが，よく用いられるのがベイズ推定によってデータ点からハイパーパラメータの事後確率分布$p(\boldsymbol{\theta}|\mathrm{data})$を求めるというものである．ただし実際にはしばしば系を記述する物理パラメータ（ベクトル$\boldsymbol{\alpha}$とする）が回帰するモデル$\boldsymbol{\mu}$に含まれているため，

$$p(\boldsymbol{\alpha}, \boldsymbol{\theta}|\mathrm{data}) \propto \mathcal{L}(\mathrm{data}|\boldsymbol{\alpha}, \boldsymbol{\theta}) \cdot p_{\mathrm{prior}}(\boldsymbol{\alpha}, \boldsymbol{\theta}) \quad (4)$$

によって各パラメータを推定する．ここで$p_{\mathrm{prior}}(\boldsymbol{\alpha}, \boldsymbol{\theta})$は$\boldsymbol{\alpha}, \boldsymbol{\theta}$の事前確率分布である[*3]．

この推定にはマルコフ連鎖モンテカルロ法（MCMC）が有効である．ただし式(1)には共分散行列の逆行列が含まれているため，N_{data}の大きさによっては逆行列演算に非常に時間がかかってしまうために注意が必要である．

ハイパーパラメータが最適化（あるいはベイズ推定）によって求まった場合，これを用いてデータ点のない領域における相関ノイズの予測を与える（補間する）ことができる．これは，振る舞いを予測したい任意の点（テストデータ）を加えたデータの新しい集合もまた多次元正規分布となる事が要請されるためである．これを用いれば，観測データを元にしたテストデータ（任意の時刻）のフラックス値の事後確率分布は正規分布となるため，その分布の平均と分散を解析的に求めることが可能である．図3における実線は，テストデータとして時刻を連続的に動かし事後確率分布の平均をプロットしたものである．また濃淡の異なる青色の領域はその分布の1σと2σの領域を表している．

次に，ベイズ推定から求まったフラックス変動を特徴づけるパラメータを視線速度データに適用する．視線速度変化のモデル化には惑星による視線速度への寄与

$$\boldsymbol{\mu} = K \sin\left\{\frac{2\pi(\boldsymbol{t}_i - t_c)}{P_{\mathrm{orb}}}\right\} + \gamma \quad (5)$$

に加えて（Kは惑星による視線速度変化の振幅，P_{orb}は惑星の公転周期，γは各視線速度データセットのオフセット），ここでも共分散行列のカーネル式(3)によって視線速度データに見られる相関を定式化する．すでに述べたように，恒星活動によるフラックス変動と視線速度変動は相関しており，その周期性や減衰のタイムスケールは共有することが可能である．そのため，フラックスの解析で「学習」したハイパーパラメータのうちいくつかはそのまま視線速度変化を記述するカーネルのハイパーパラメータとして採用できる．Grunblattら[3]はこうした周期性や減衰タイムスケールが厳密には光度曲線と視線速度で微妙に異なる事も想定して，光度曲線の解析に基づく

[*3] なおここでは共分散行列に登場するパラメータをハイパーパラメータと呼んでいて，通常の階層ベイズ推定とは異なりデータから式(4)に従って系の物理パラメータとハイパーパラメータを同時に推定する．

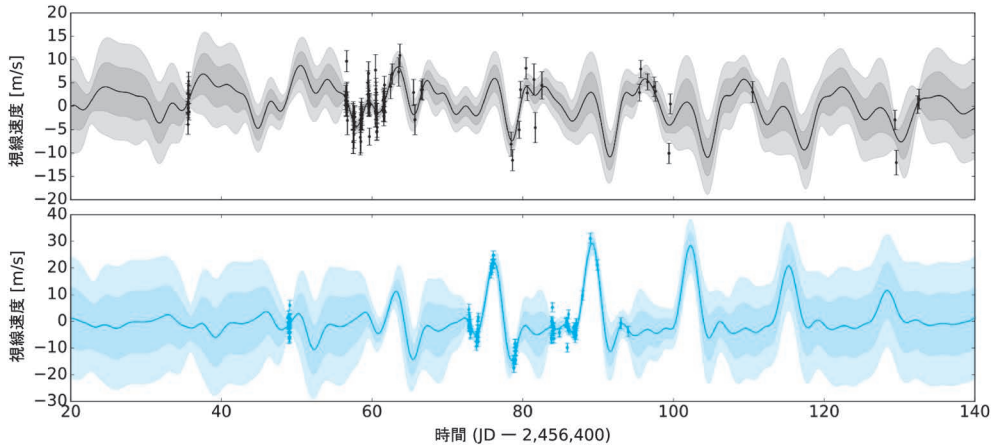


図4　Kepler-78の視線速度値（エラー付きの点）とそのガウス過程による回帰．実線とその周りの影付きの領域は，観測データを元に機械学習したテストデータ（各時刻）の視線速度事後確率分布の平均と分散に対応する（濃い領域，薄い領域がそれぞれ1σと2σ）．HARPS-N（上のパネル）とHIRES（下のパネル）は異なる波長域を用いて視線速度解析を行っていることから恒星活動による視線速度変動も異なる振る舞いをすると予想されるため，ここでは独立な回帰を行っている（ただし惑星軌道に関するパラメータは共通）[3]．

θ等の推定結果を視線速度データの解析の際の「事前確率分布」として採用し，視線速度データの解析で別途ハイパーパラメータを導出している．なお式(3)における相関ノイズの振幅hなどは光度曲線と視線速度でそもそも異なる物理量であるため，前者の結果を後者の解析における事前確率分布としては使用しない．視線速度データのフィットでは，ハイパーパラメータに加えて式(5)による系のモデルパラメータ（Kやγ）を同時に決定する．ここでも光度曲線の解析の際と同様にMCMCを用いる．図4は，視線速度の観測値（誤差付きの点）とハイパーパラメータと系のパラメータの最適値を採用した場合の相関ノイズを含めた視線速度モデル（実線）である．図4とはデータが取得された時刻が異なるため単純な比較は難しいが，視線速度が光度曲線に見られた約13日の変化に対応して大きく変動している様子がよく再現されている．これらの解析によって最終的にKepler-78bの視線速度振幅として$K=1.86\pm0.25$ m s^{-1}という値が得られ，惑星質量は$1.87^{+0.27}_{-0.26}$地球質量と求まる．このようにガウス過程を用いた解析により，一般にKepler-78のように活動度が高くRVジッターの大きな恒星であってもその中に埋もれた惑星による視線速度変化をより正確に抽出することが可能となる．

5.　トランジット解析へのガウス過程の応用

次に地上トランジット観測のデータ解析へのガウス過程の応用例を紹介する．上述したように，地上からのトランジット観測では恒星活動によるフラックス変動に加えて装置や地球大気等の観測条件の変化に起因したフラックス変動が支配的となり，小型惑星のトランジットの検出・追観測を困難にしている．図5は，ケプラー望遠鏡の現在のミッションである「*K2*」で惑星候補をもつと同定された天体K2-28に対しわれわれが地上からトランジットの追加測光観測を実施した結果であるが，トランジットの深さ（約0.6%）と同程度のフラックスのばらつきが見られトランジットが正確にどこにあるのかの判別が難しい．トランジットサーベイで同定された惑星候補に対し地上からトランジット追観測を実施するメリットには，1）惑星の発見確認，2）公転周期・惑星半

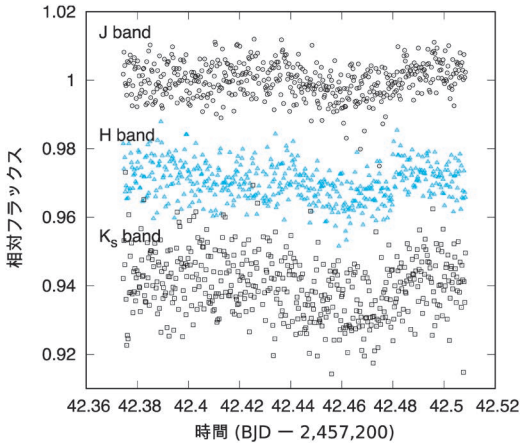


図5　IRSF1.4 m望遠鏡で観測されたK2-28bのトランジット．近赤外三色同時カメラ（SIRIUS）[6]で観測．上からJ，H，K_sバンドの光度曲線がプロットされている．

径等の系のパラメータの改善，3）惑星大気の調査等が含まれるが，いずれもトランジットの深さや時刻の正確な見積もりが本質的である．そのため，光度曲線に系統的な影響を与える相関ノイズの取り扱いが重要になってくる．

ここでは5章に続き，ガウス過程を用いて光度曲線に見られる相関ノイズを取り扱う．トランジット観測のような数時間のタイムスケールの測光観測では恒星活動由来の相関ノイズ*4よりも観測条件や機器の変化による相関ノイズが支配的であることが多い．特に地上高精度測光観測では1）検出器上の星像位置，2）シーイング（星像の半値全幅），3）背景光強度，等が特にフラックス変化と相関を示すことが知られている[7]．そこで図5のトランジット測光データをガウス過程を用いてモデル化するにあたって，フラックスと同時に取得される補助的なパラメータとの相関を共分散行列に導入する．すなわち，ガウス過程により観測されたデータそのものとそれに付随するパラメータをトレーニングデータとして機械学習し，相関ノイズの影響を補正する．図5のように3バンド同時に測光データが得られている場合，ガウス過程による回帰の尤度関数$\mathcal{L}$は，各バンドの尤度関数の積として

$$\ln\mathcal{L}=-\frac{1}{2}\sum_{i=J,H,K_s}\{(\boldsymbol{f}^{(i)}-\boldsymbol{\mu}^{(i)})^{T}(\Sigma^{(i)})^{-1}\times(\boldsymbol{f}^{(i)}-\boldsymbol{\mu}^{(i)})+\ln|\Sigma^{(i)}|+N^{(i)}_{data}\ln(2\pi)\}\quad(6)$$

と表される．視線速度の場合と同様に，$\boldsymbol{f}^{(i)}$はi（$=J, H, K_s$）バンドの観測データ（各時刻のフラックス）ベクトル，$\boldsymbol{\mu}^{(i)}$はトランジットのフラックスモデルベクトルを表す．

共分散行列$\Sigma^{(i)}$には任意性があり，上で挙げた補助的なパラメータの導入にもさまざまな方法が考えられる．ここでは，近接する補助パラメータ同士の相関を反映する“squared exponential”カーネル

$$\Sigma^{(i)}_{nm}=\sum_j A_j^{(i)2}\exp\left(-\frac{(p^{(i)}_{j,n}-p^{(i)}_{j,m})^2}{2L_j^{(i)2}}\right)+\delta_{nm}\sigma_n^{(i)2},\quad(7)$$

を用いる．jは相関をもたせる各補助パラメータを表すラベルで，$p_{j,n}$（$p_{j,m}$）はその補助パラメータの値である．ここではjとして観測時刻，恒星重心の検出器上での位置（x, y），星像の半値全幅，背景光強度の五つのパラメータを採用する．式(7)は，異なる時刻t_nとt_mに取得された補助パラメータ同士がお互いに値が近ければ相関が強く，遠ければ指数関数的に相関が弱くなるということを反映したカーネルであり，ガウス過程の回帰において最も頻繁に用いられるカーネルの一つである．ハイパーパラメータA_j, L_jはそれぞれ各補助パラメータの相関強度，相関長を表していて，実際の観測データから機械学習によって推定される．なお通常のχ^2計算で出てくる各フラックスの統計誤差（白色ノイズ）$\sigma_n^{(i)}$は，式(7)において第二項の独立同分布として登場する．

*4　黒点の場合は1日程度以上のゆっくりとしたタイムスケールで変化することが多い．

4章で紹介した視線速度データへのガウス過程の応用とは異なり，本章ではモデル化する観測量（フラックス）と「同時に」取得された情報を用いて相関ノイズを取り扱うことができるのが特徴である．そのため，ここでは一度のフィットで通常のトランジット・パラメータ（中心時刻，中心星惑星半径比，軌道長半径，衝突係数等）に加えて観測データを最もよく記述するようにA_j, L_j等のハイパーパラメータを同時に最適化する．これらの同時推定には4章同様MCMC計算を用いるのが便利である．ただし，データ点数が多い場合MCMCの各ステップで式(6)の共分散行列の逆行列を計算するのは非常に計算コストが大きい．そのため，まず式(6)の尤度（の対数）を最大化するようなパラメータの組み合わせを決定し，ハイパーパラメータのみを固定（そうすることで共分散行列$\Sigma^{(i)}$を固定）した上でトランジット・パラメータ推定のためMCMCを実行する「第II種最尤推定」という近似法がしばしば用いられる[*5]．ここでもこの近似を用いる．

図6（上）は，図5と同じK2-28のトランジット前後の観測フラックスに加えて，上記ステップで求めた最適パラメータを採用した場合のガウス過程による回帰曲線（実線）を表しており，細かい変動が相関ノイズによる変動に対応している．各バンドで背景光強度や星像の半値幅等が大きく異なるため相関ノイズの振幅や相関スケールも異なっているが，一般に短波長側でフラックス全体の変動に対する相関ノイズの寄与が大きいことが伺える．図6の下のパネルはこれら相関ノイズを除いた場合の各バンドの光度曲線を表し，分かりやすくするためにフラックスを16点ごとにビニングしてある．こうしてフィットで求まった中心星惑星半径比は*K2*データの解析から求めた半径比と1–2 σ以内で一致したことから，K2-28bは三重連星系などではなく実際に惑星であることが確認された[*6]．

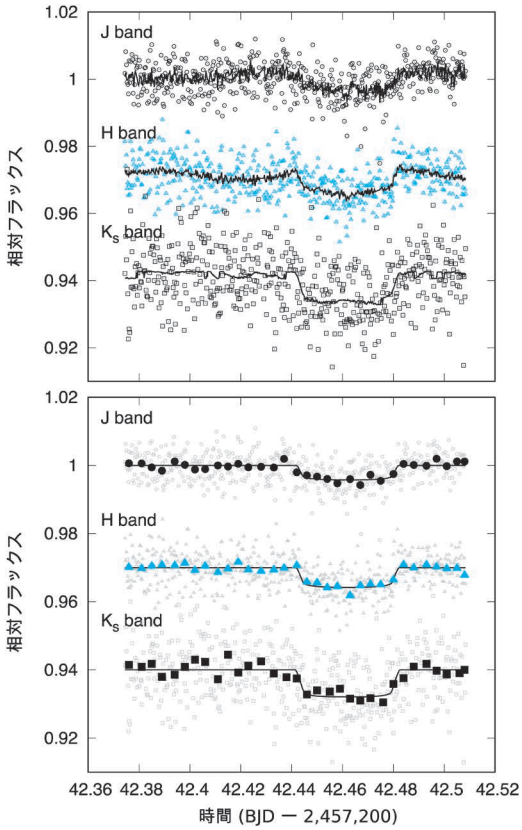


図6　IRSF1.4 m望遠鏡で取得されたK2-28のトランジット前後の光度曲線．（上）図5と同じ観測データがプロットされているが，ガウス過程による回帰で得られたテストデータの事後確率分布の平均を実線で示している．（下）観測された光度曲線から，相関ノイズ成分を除いたデータ（フラックス）．大きいシンボルは各フラックスを16点ごとにビニングしたもの[9]．

[*5] 第II種最尤推定を適用した場合と，ハイパーパラメータを含めた全パラメータをMCMCで動かした場合の結果の違いはよく調べられている[8]．

[*6] 例えば食連星を含む階層三重連星系の場合，各恒星の等級が波長（バンド）によって異なるため，多バンドでトランジット（食）の測光観測を行うことによってそのシナリオを制限することができる．なお，われわれのK2-28bの発見論文[9]ではIRSF1.4 m望遠鏡/SIRIUSのデータに加えて，岡山天体物理観測所188 cm望遠鏡/MuSCAT[10]のトランジット測光データも同時に解析している．

図6では少しわかりにくいが，特にJバンド，Hバンドではガウス過程を用いた回帰により最適モデルの周りでのばらつきが改善している．例えば相関ノイズを考慮しなかった場合，Hバンドの最適トランジットモデルの周りでの観測された相対フラックスのRMSは0.0054であったのに対し，上で示したガウス過程による相関ノイズを含めた解析では最適モデルのまわりのフラックスのRMSは0.0046に改善した．K2-28は比較的活動度の低い中期M型星であったため光度曲線のベースラインにそれほど大きな変動は見られなかったが，星団に属する若い星など活動度の高い恒星の周りで小型惑星を探査する場合はここで紹介したような相関ノイズのモデル化が極めて重要になる．

6. まとめ・今後の展望

本稿では，精密観測により取得されたデータの機械学習による回帰，あるいはデータ中の「相関ノイズ」を取り扱う方法論の一つとしてガウス過程を紹介した．ガウス過程は機械学習によるデータ点の補間や相関ノイズの補正を実際の関数形を指定することなく（＝ノンパラメトリックに[*7]）行うことができるという利点があり，系外惑星関連でも最近急速に利用が拡大している．視線速度観測では，別途取得された光度曲線の情報を元に変動の周期や相関の時間スケールを推定し恒星活動に起因した視線速度変化を定量的に評価・補正する手法を紹介した（4節）．また地上からのトランジット測光観測では，同時に得られた補助的な情報とフラックス変動の相関を調べ，データそのものをトレーニングデータとして機械学習する方法論について述べた（5節）．いずれも系外惑星観測・解析におけるガウス過程の典型的な適用例であるが，本稿で見たようにその応用範囲は非常に広い．

ガウス過程を用いた回帰では一般に共分散行列のカーネルに適当な関数形を与えることで様々な特徴を持ったデータを扱うことができ，相関ノイズを考慮したより正確な解析が可能となる．ただしこれは必ずしも「精密」な結果につながるとは限らず，視線速度解析などにおいてガウス過程を用いないモデル回帰であっても最終的に得られる物理量（惑星質量等）の誤差はガウス過程を用いた場合とほぼ同程度となることも多い．一方ガウス過程を用いない場合，各視線速度データと最適モデルが統計誤差に比べて大きくずれたフィットとなり，当然χ^2の値も悪くなる．ガウス過程を用いることでよりロバストな解析が可能となると言える．なおガウス過程を用いた回帰ではどのカーネルを仮定するかによって結果が変わることも知られており注意が必要である．一般にはいくつかカーネルを変えてみて，結果の一貫性をチェックしたりモデル選択のための統計量（ベイズ情報量規準，赤池情報量規準等）を比較するなどの作業が必要となる．

系外惑星探査に限ってみても，ガウス過程（機械学習）を始めとするデータ科学的手法を実際の観測データに適用する機会は今後ますます拡大すると予想される．例えば視線速度法による惑星探査では，近年世界のいくつかのグループは近赤外高分散分光観測による低温度星（M型矮星）まわりの惑星探査を実施・計画している．われわれもすばる望遠鏡に最近搭載された近赤外分光器IRD[11]を用いた大規模視線速度サーベイを計画しており，地球型惑星を含めたM型矮星まわり

[*7] ベイズ推定における「ノンパラメトリック」とは，データを記述するパラメータ空間が無限次元であることを表す．ガウス過程の場合データの振る舞いを記述するのは確率分布「関数」であるため，有限次元のパラメータで振る舞いが決定論的に決まる「パラメトリック」な推定とは区別される．例えばデータからある時刻のフラックス（視線速度）値を予測する場合，パラメトリックな推定では時刻とフラックスの1対1の関係式（多項式等）が用いられるのに対し，ガウス過程では他のデータの集合（とハイパーパラメータ）から求める確率分布が与えられる．

のユニークな惑星の発見を目指している．しかしながら，本稿で述べたように小型惑星の発見は常に恒星活動や装置由来の見かけ上の視線速度変動との戦いであり，それらの相関ノイズを補正しない限り信頼度の高い発見は難しい．近赤外領域（IRDは*Y*, *J*, *H*バンドを同時にカバーする）ではパッシェン系列線等が恒星活動度の指標になると考えられ，視線速度との相関をもつことが予想される．また可視光線による視線速度観測とは異なり，近赤外では地球大気吸収による影響を強く受けることから，エアマス等の観測条件に付随した補助パラメータとも視線速度が相関をもつ可能性がある．今後実際に観測される近赤外スペクトルをもとに，ガウス過程等を用いた恒星活動による相関ノイズを補正する手法を確立することが小型惑星探査において必要不可欠である．

謝　辞

ハワイ大学大学院生Sam Grunblattさんには本稿で使用されている図の数値データを提供してもらい，さらに本稿の内容について有益なコメントをいただいた．本稿の執筆を勧めていただいた広島大学の植村誠さん，天文月報編集委員の岡部信広さんにはこの場を借りて御礼申し上げたい．

参考文献

1）西道啓博，2018, 天文月報，111，10月号掲載予定
2）https://exoplanetarchive.ipac.caltech.edu
3）Grunblatt, S. K., et al., 2015, ApJ 808, 127
4）Rasmussen, C. E. & Williams C. 2006, Gaussian processes for machine learning（The MIT Press）
5）Haywood, R. D. et al., 2014, MNRAS 443, 2517
6）Nagayama, T., et al., 2003, Society of Photo-Optical Instrumentation Engineers（SPIE）Conference Series, 4841, 459
7）Fukui, A., et al., 2013, ApJ 770, 95
8）Gibson, N. P., et al., 2012, MNRAS 419, 2683
9）Hirano, T., et al., 2016, ApJ 820, 41
10）Narita, N., et al., 2015, Journal of Astronomical Telescopes, Instruments, and Systems, 1, 045001
11）Kotani, T., et al., 2014, Society of Photo-Optical Instrumentation Engineers（SPIE）Conference Series, 9147, 14

Data Science Methods in Exoplanet Researches

Teruyuki Hirano
Department of Earth and Planetary Sciences, Tokyo Institute of Technology, 2–12–1 Ookayama, Meguro-ku, Tokyo 152–8551, Japan

Abstract: Recently, observations in the exoplanet studies have become so precise that observational signals are often dominated by the intrinsic stellar activities or instrumental artifacts in the data, preventing the discovery and precise characterizations of exoplanetary systems. To handle those “correlated noises,” new data-science approaches have recently been studied and developed. In this article, I focus on a machine-learning technique using “Gaussian processes” as a way to mitigate the impact of correlated noises in the observational data. Introducing some examples for the major observing techniques of exoplanets（the radial velocity method and transit method），I discuss the advantages and limitations of Gaussian process-based analyses.