

doi: 10.3969/j.issn.1000-8349.2025.02.XX

基于人工智能的长周期微弱凌星信号探测 进展

闫星辰^{1,2}, 刘擎天^{1,2}, 葛健¹, 阮泽宇^{1,2}, 刘政宏^{1,2}

(1. 中国科学院 上海天文台, 上海 200030; 2. 中国科学院大学, 北京 100049)

摘要: 凌星法是发现系外行星的主要手段, 高精度空间测光巡天积累了大规模光变数据, 但相较短周期目标, 长周期小半径行星的检出与统计仍受限。长周期目标几何凌星概率低、有限基线下可观测的凌星次数少, 且小半径行星凌星深度浅, 易受恒星红噪声与仪器系统误差影响, 导致检测显著性降低并增加候选体判别难度。本文围绕长周期微弱凌星信号的搜寻流程, 光变曲线噪声建模与预处理、初步筛选 (Triage) 与候选体分类 (Vetting) 等环节的关键技术问题与进展, 归纳去共趋势与去趋势处理、经典周期搜寻算法以及人工智能方法在低信噪比条件下的优势与局限, 并讨论完备度与可靠性评估的关键挑战。结合未来空间测光任务, 展望长周期类地行星探测中自动化管线与统一基准评估的发展方向。

关键词: 行星与卫星探测; 恒星行星系统; 测光技术; 统计方法; 巡天; 机器学习; 深度学习
中图分类号: P141.2 **文献标识码:** A

1 引言

1.1 系外行星研究的科学背景

系外行星的搜寻, 尤其是针对位于宜居带内的长周期、微弱凌星信号的探索与分析, 是当代系外行星、天体物理与行星科学的重要研究方向之一。

围绕类太阳恒星的系外行星研究取得突破始于 Mayor 与 Queloz 于 1995 年首次确认 51 Pegasi b 的发现^[1]。在凌星法方面, Charbonneau 等人于 2000 年发表了 HD 209458b 首次凌星信号的独立探测结果^[2], 随后 Henry 等人^[3]的独立观测也确认了这一凌星事件。这两项独立工作的几乎同步发表, 标志着凌星法在系外行星探测领域的重大突破。随着凌星法等探测技术的逐步成熟, 尤其是 Kepler 等空间任务提供的大规模、精确光变数据, 极大促进了系外行星统计研究的进展。基于这些 Kepler 观测数据的分析表明, 在短周期区间内, 小半径行星的发生率整体上高于大半径行星, 且在约 $1.7\text{--}2.1 R_{\oplus}$ 附近存在“半径谷” (Radius Valley)

收稿日期: 2026-03-04; 修回日期: 2026-03-04

资助项目: 太空探源专项 GJ11030405

通讯作者: 葛健, jge@shao.ac.cn

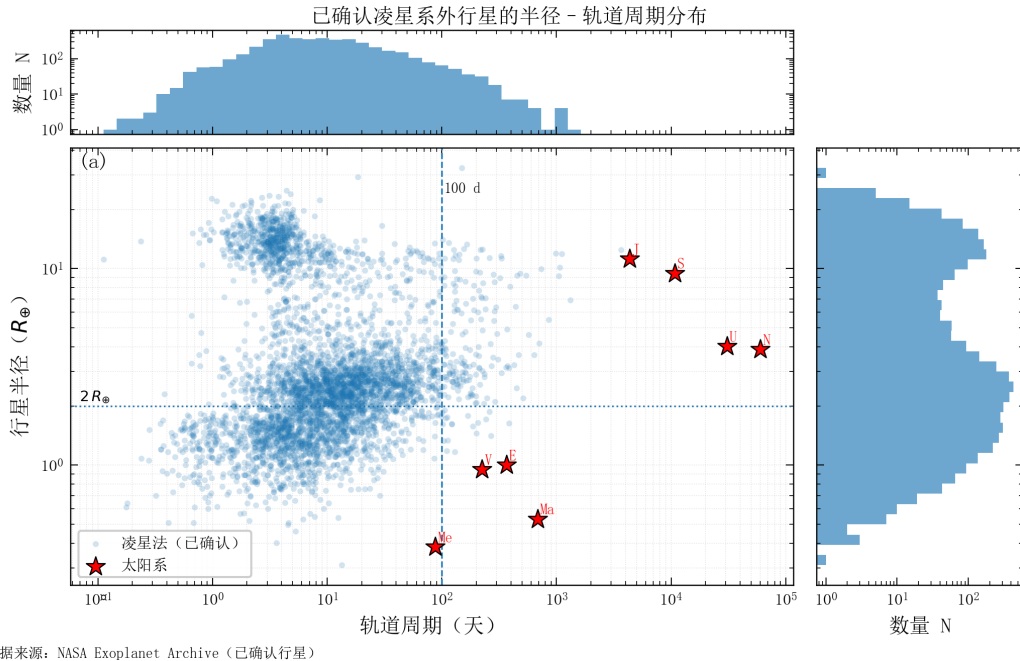


图 1 已确认凌星系外行星周期-半径分布及边缘统计

结构 (图 1)^[4-6]。这些观测特征表明, 小半径系外行星在银河系中可能较为常见, 而半径谷区域的形成机制及其统计性质也仍待进一步研究。在引入探测完备度修正与恒星参数改进后, 研究表明类太阳恒星在宜居带内可能存在岩质行星, 其在特定模型假设下的宜居带类地行星的平均发生率 $\eta_{\oplus}$ 估计约为每颗恒星 0.37–0.60^[7]。这些基于统计推断得到的内禀行星分布与当前已确认的系外行星分布 (图 1) 之间仍存在一定差距, 反映出观测选择效应在样本构成中的决定性作用。根据开普勒候选天体 (Kepler Object of Interest, KOI) 目录的统计, 当轨道周期小于 100 天时, 行星探测完备度高于 85% 且可靠性高于 98%; 而在 200–500 天且信噪比 (signal-to-noise ratio, SNR) 较低的参数空间内, 对于 FGK 型 (F、G、K 型) 主序恒星样本, 探测完备度仅约为 76.7%、可靠性约为 50.5%^[8]。

上述对比表明: 有限观测基线、几何凌星概率随半长轴增大而降低, 以及小半径行星凌星信号更浅等因素, 共同造成长周期参数空间覆盖不足, 并使已观测样本与经完备度校正后的内禀统计之间仍存在差距。针对长周期微弱凌星信号发展高效探测与基于人工智能的候选体筛选, 是提升该区间完备度、完善发生率与轨道统计的关键, 也有助于充分挖掘 *Kepler*^[9, 10] 与凌星系外行星巡天卫星 (Transiting Exoplanet Survey Satellite, TESS)^[11] 等任务的科学潜力。面向宜居带类地行星的长期监测任务, 如欧洲航天局 (European Space Agency, ESA) 行星凌星与恒星振荡探测 (Planetary Transits and Oscillations of stars, PLATO)^[12] 与系外地球 2.0 计划 (Earth 2.0, ET)^[13-15], 有望通过延长基线与优化恒星样本进一步改善数据

条件。

1.2 凌星法在系外行星发现中的地位

据美国国家航空航天局 (National Aeronautics and Space Administration, NASA) 系外行星档案 (NASA Exoplanet Archive) 近年的统计结果, 当前已确认的系外行星总数已超过 6000 颗^[16]。

按发现方法划分, 凌星法贡献的已确认系外行星的规模约占全部已确认样本的三分之二以上; 视向速度法已发现约千余颗行星; 微引力透镜与直接成像方法分别贡献了数百颗和百余颗行星, 其它方法 (如天体测量、测时变分等) 所占比例相对较小^[17]。凌星法在当前已确认系外行星样本中占据主导地位, 其主要原因在于凌星信号在观测幅度上较视向速度与直接成像信号强数个数量级, 使其在探测能力方面占显著优势; 结合空间望远镜的大视场、大样本巡天策略, 显著提升了行星发现速度和效率^[18]。

按主要观测设施划分, *Kepler* 任务已确认的系外行星数量接近 2800 颗, 尚有约 2000 个候选体有待进一步确认; K2 任务已确认行星约 500 余颗, 并提供了近千个候选体; TESS 任务截至目前已确认行星数达数百颗, 其归档的候选体总数已超过 7000 个, 其中仍有数千个目标有待后续观测验证^[17]。在过去十余年中, 空间望远镜因可避免地球大气扰动而在测光精度上普遍优于地基设施, 从而在凌星巡天中占据核心地位^[10, 11]。除 *Kepler* 与 TESS 外, CoRoT^[19]、K2^[20] 以及 WASP^[21]、KELT^[22] 等共同构成多平台、互为补充的凌星观测体系; 胡泉泉等人^[18]对主要凌星设施的光学设计、技术特点与任务规划作了系统综述。

1.3 后文结构安排

本文系统综述长周期微弱凌星信号探测相关方法的发展脉络, 重点分析去趋势处理、经典周期搜索算法及人工智能方法在低信噪比条件下的适用性与局限。其中, 第2节围绕 *Kepler* 等空间测光任务的数据预处理流程, 系统介绍光变噪声来源与分类, 并梳理去共趋势 (Cotrending) 与去趋势 (Detrending) 方法的发展脉络与建模框架^[23], 分析时间相关噪声与系统性误差对长周期、低信噪比凌星信号检出的影响, 并讨论不同技术路径在信号保护与噪声抑制之间的权衡机制; 第3节从经典盒状最小二乘 (Box-fitting Least Squares, BLS) 到凌星最小二乘 (Transit Least Squares, TLS) 算法, 评述现代凌星信号搜寻管线的演进, 并重点讨论机器学习与深度学习方法在自动化搜寻中的应用。第4节在此基础上凝练当前长周期与宜居带类地行星探测面临的核心挑战, 并结合 PLATO、ET 等未来任务展望凌星法结合人工智能方法的发展方向。

2 光变曲线处理与噪声建模

2.1 光变噪声来源与分类

系外行星凌星探测的核心挑战之一在于从复杂的光变噪声中提取微弱的凌星信号。光变噪声根据其时间相关性可分为白噪声 (White Noise) 和红噪声 (Red Noise) 两大类, 二者的物理起源、统计特性和对凌星探测的影响截然不同, 需要采用不同的建模和缓解策略。

白噪声，又称不相关噪声或泊松噪声，是指在时间序列中各时刻的测量误差相互独立、服从相同概率分布（通常为高斯分布）的随机涨落。其主要来源包括光子散粒噪声（Photon Shot Noise）、读出噪声（Readout Noise）和天光背景噪声（Sky Background Noise）。

光子散粒噪声源于光子到达探测器的随机性，根据泊松统计，探测到 N 个光子的不确定性为 $\sqrt{N}$ ，因此相对测光精度为 $1/\sqrt{N}$ ，这意味着积分时间越长或恒星越亮，光子噪声越小。对于 *Kepler* 任务，长曝光采样（Long Cadence, LC）单次积分时间为 29.4 分钟；以 6 小时时间窗口内多个曝光的组合噪声（Combined Differential Photometric Precision, CDPP_{6h}，即六小时组合差分测光精度）衡量，15 等恒星的等效测光噪声约为 50–500 ppm（parts per million，百万分之一），而地球大小行星绕太阳型恒星的凌星深度约为 84 ppm，因此对于这类长周期微弱凌星信号的探测仍具有一定挑战^[24, 25]。

读出噪声来自电荷耦合器件（Charge-Coupled Device, CCD）探测器的电荷读出与模数转换过程，其大小取决于具体仪器设计。对于 *Kepler* 空间望远镜，其 CCD 在 LC 模式下的单次读出噪声平均约为 $86 e^-$ （电子）^[26, 27]。由于 *Kepler* 每次长曝光由多次短积分叠加而成，读出噪声在总噪声预算中通常小于光子散粒噪声，但在暗弱恒星或背景主导的情况下仍可能成为重要贡献项。

天光背景噪声则源于散射光与弥散背景辐射。对于地基观测，主要来源为地球大气分子和气溶胶散射；而对于空间望远镜（如 *Kepler*），则主要来自黄道光和弥散银河背景。其强度依赖于观测波段、天区位置及太阳角距等几何因素。

白噪声的关键特性是其功率谱密度（Power Spectral Density, PSD）在所有频率上均匀分布，即 $P(f) = \text{const}$ 。对光变曲线的离散时间序列 $h(t)$ 进行离散傅里叶变换得频域分量 H_k ，其功率谱密度定义为

$$\text{PSD}(f_k) = (\Delta t)^2 (|H_k|^2 + |H_{N-k}|^2), \quad k = 1, \dots, \left\lfloor \frac{N}{2} \right\rfloor, \quad (1)$$

其中， Δt 为采样间隔， H_k 为第 k 个傅里叶系数， N 为时间序列长度， $f_k = \frac{k}{N\Delta t}$ 为对应频率。

由于白噪声的不相关性，通过 N 次独立测量平均可使噪声水平降低 $\sqrt{N}$ 倍，这一平方根定律是凌星叠加（Phase-folding）提高信噪比的理论基础。白噪声与红噪声在时域形态、自相关特性及振幅分布上的对比如图 2 所示。值得注意的是，二者的单点振幅分布均近似高斯分布（图 2d），其本质区别仅体现在时间相关性上（图 2c）：白噪声的自相关函数在非零滞后处迅速衰减至统计噪声水平（ 2σ 置信带以内），而红噪声则保持显著的长程正相关。例如，对周期三天的热木星，*Kepler* 四年基线可观测约 500 次凌星，叠加后等效噪声降低 $\sqrt{500} \approx 22$ 倍，使单次信噪比仅为 4 的凌星在叠加后达到信噪比 90，远超 *Kepler* 流水线所采用的多事件统计量（Multiple Event Statistic, MES）探测阈值 7.1σ ^[28]。

然而，实际光变数据中普遍存在红噪声（亦称相关噪声），其特征为相邻时刻的测量误差呈正相关，即噪声具有“记忆效应”。红噪声的功率谱密度在低频端显著增强，典型形式为 $P(f) \propto f^{-\alpha}$ （ $\alpha > 0$ ），其中 $\alpha = 1$ 对应闪烁噪声（ $1/f$ 噪声）， $\alpha = 2$ 对应随机游走噪声（布朗噪声）；实际恒星光变中的噪声频谱指数通常介于二者之间，反映了恒星对流与磁活动共同作用下的复合噪声特性。红噪声的主要来源包括：（1）恒星本征活动：黑子、光斑、

耀斑、粒状对流、p 模式振荡等引起的恒星表面亮度时变, 其特征时标自分钟量级 (粒状对流) 至数月 (黑子周期演化) 不等^[29]; (2) 仪器系统误差: 热漂移导致的焦距变化、指向抖动引起的星像位移、探测器像元响应非均匀性 (flat-field 误差) 及望远镜光学系统的长期退化等^[30]; (3) 环境扰动: 地基观测中的大气透明度与视宁度起伏、空间望远镜轨道温度的周期性变化等。上述因素在长时标上累积, 形成准周期或缓慢漂移趋势, 对凌星信号造成严重污染。

为直观展示红噪声对凌星可见性的影响, 图 3 以 *Kepler* 长周期行星 *Kepler-767 b* (KOI-1546.01, KIC 2853093; $P \approx 161.53$ d, $\delta \approx 0.465\%$, $T_{14} \approx 3.69$ h; 参数取自 NASA Exoplanet Archive) 为例, 对比了两种噪声背景下凌星信号的可辨识度。由于红噪声在凌星持续时间尺度上呈现缓慢漂移趋势, 凌星凹陷被噪声的低频起伏所掩盖, 三次凌星事件在视觉上难以与噪声波动区分。该对比表明, 对于轨道周期数十天至数百天的长周期行星, 凌星持续时间 (数小时至一天) 远短于红噪声的相干时标, 凌星信号易被红噪声淹没或误判为噪声起伏的一部分^[25], 凸显了在长周期凌星搜索中对红噪声进行有效建模与抑制的必要性。

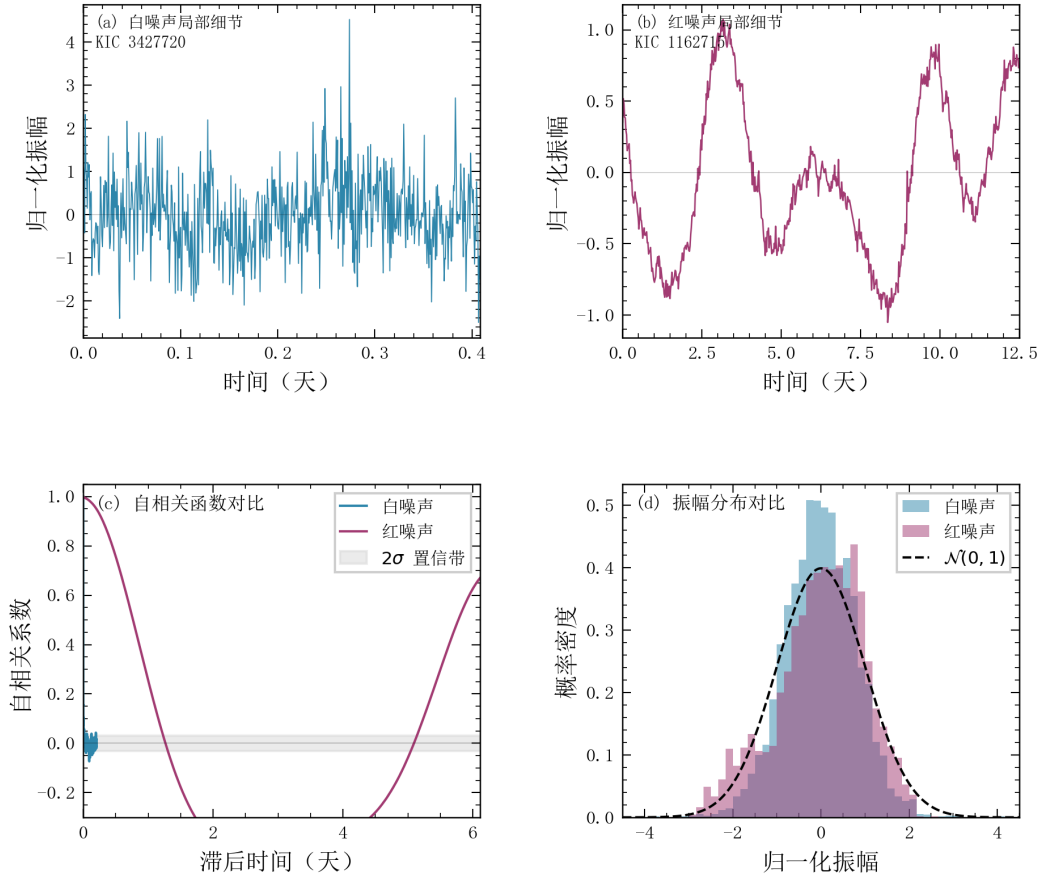
红噪声对凌星探测的影响主要体现在两方面。其一, 红噪声导致凌星信号的统计显著性被系统性高估。传统 χ^2 统计与信噪比计算均假设噪声为白噪声、各数据点相互独立, 典型如 BLS^[34] 与 TLS^[35] 等凌星搜索算法; 当实际噪声为红噪声时, 数据点间的相关性使有效自由度下降、真实不确定性被低估^[25]。Carter 和 Winn 的模拟研究表明, 若忽略红噪声, 长周期行星的虚警概率 (False Alarm Probability, FAP) 可被低估 1–2 个数量级, 从而造成大量假阳性候选体被误报^[36]。其二, 红噪声会增大凌星参数反演的系统误差: 在红噪声存在时, 直接拟合光变曲线会使凌星深度、进出凌时刻等参数的不确定性被严重低估, 进而使后续统计分析 (如行星发生率估算) 产生偏差^[25]。为量化和建模红噪声, 数据预处理与信号搜寻管线中广泛采用高斯过程 (Gaussian Process, GP)、小波、主成分分析及系统误差校正等方法; 其方法分类及在去共趋势、去趋势与候选体验证中的具体应用将在第 2 节及后续章节中述及。

光变噪声的正确识别、分类和建模是凌星探测从“发现时代”迈向“精确科学时代”的关键。白噪声虽可通过延长积分时间和叠加观测降低, 但红噪声的时间相关性要求更复杂的统计方法和物理建模。高斯过程、深度学习和贝叶斯检验的结合为处理红噪声提供了强大工具, 使得长周期、小半径行星尤其是宜居带类地行星的探测成为可能。

2.2 去共趋势预处理方法

去共趋势预处理是当前凌星数据预处理中不可或缺的一步, 其核心思想是利用大规模光变曲线之间的共趋势来建模并去除系统性误差, 从而增强真实天体信号的可检出性^[37, 38]。

在 *Kepler* 科学管线中, 搜索前数据预处理 (Pre-search Data Conditioning, PDC) 模块^[37]负责对原始光变数据的系统性趋势和噪声进行校正, 使之适配后续的凌星搜索算法。更具体地说, 该模块从众多光变曲线中选取统计上高度相关且安静的恒星样本, 进而应用奇异值分解 (Singular Value Decomposition, SVD) 对这些样本的光变矩阵进行分解, 从而构造出一组共趋势基向量^[24] (Cotrending Basis Vectors, CBVs)。这些基向量代表了某个 CCD 通道

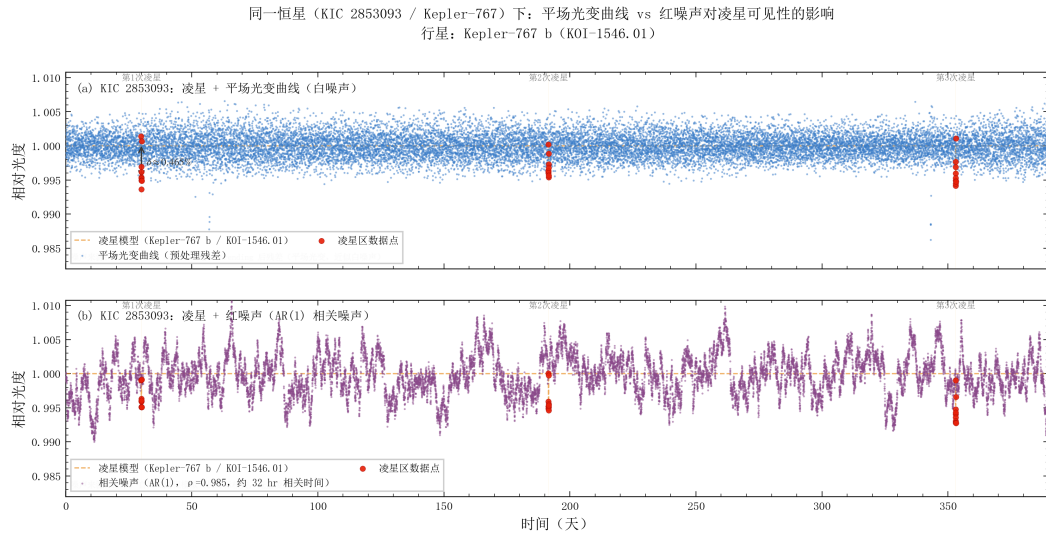


注：数据取自 NASA *Kepler* 任务公开档案^[16]，经 *lightcurve* 工具包^[31]下载与预处理。白噪声样本为安静 G 型矮星 KIC 3427720 的短采样 (Short Cadence, SC) 光变曲线第五季 (Q5) 去趋势残差，光子噪声主导^[32]；红噪声样本为磁活跃恒星 KIC 1162715 的长采样 (LC) 第三季 (Q3) 归一化光变曲线，星斑旋转调制产生显著低频率^[33]。(a) 为白噪声光变在 ~ 0.4 d 时间尺度上的局部细节；(b) 为红噪声光变在 ~ 12.5 d 时间跨度内的代表性片段，可见低频/准周期起伏。(c) 为自相关函数 (ACF) 对比，白噪声在非零滞后处迅速落入 2σ 置信带 (灰色阴影)，红噪声因旋转周期 ($P_{\text{rot}} \approx 5$ d) 而呈现长程周期性相关；(d) 为归一化振幅的概率密度分布，二者均近似标准正态 $\mathcal{N}(0, 1)$ (黑色虚线)，表明两类噪声的主要差异在于时间相关结构而非边际振幅分布。

图 2 白噪声与红噪声统计特性的对比

中普遍存在的共性系统噪声模式 (例如热漂移、指向抖动等^[38])，并通过列堆叠构成一个设计矩阵。矩阵的每一列对应一个噪声模式随时间的变化响应，而每一行则对应一个时间采样点。在后续使用最小二乘法 (Least Squares, LS) 拟合噪声的过程中 (即 PDC-LS)，该设计矩阵作为自变量矩阵来描述系统性误差对观测光变的贡献。

以第 5 季度 (Quarter 5) 第 42 通道八条光变曲线为例，基于 SVD 的去共趋势流程可概括为三步 (图 4)。(1) 样本选择与归一化：选取该通道内典型简单孔径测光 (Simple Aperture Photometry, SAP) 光变曲线并进行 Z-Scale 标准化 (均值为 0、标准差为 1)，标准化后的光变曲线显示各目标星具有高度一致的长期漂移，该漂移主要源于望远镜的系统性仪器效应等，需从中分离出真实恒星活动信号。(2) 对标准化光变矩阵进行奇异值分解，取前 k 个

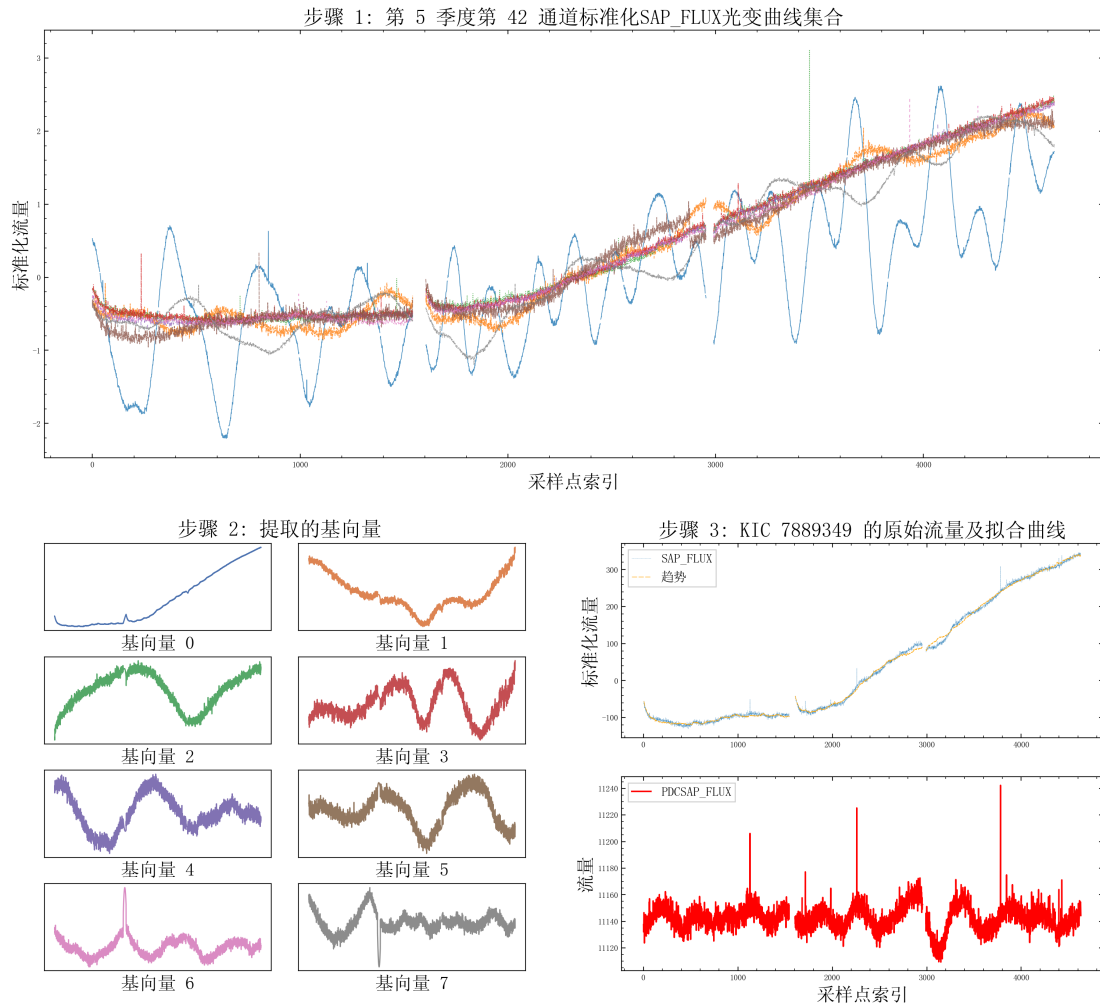


注: 两幅子图均以 KIC 2853093 (*Kepler*-767) 为宿主恒星背景, 行星参数取自 NASA 系外行星档案 (NASA Exoplanet Archive) 中 *Kepler*-767 b (KOI-1546.01) 的观测值: 轨道周期 $P \approx 161.528$ d, 凌星深度 $\delta \approx 0.465\%$ (≈ 4654 ppm), 凌星持续时间 $T_{14} \approx 3.694$ h^[40]。红色圆点标注各次凌星时段内的观测数据点; 橙色虚线为注入的梯形凌星模型; 橙色阴影标注凌星持续窗口。(a) 平场光变曲线 (白噪声): 噪声背景取自 KIC 2853093, 经 *lightkurve*^[31] 下载并进行 CBV 去共趋势及多项式去趋势处理后得到残差。其噪声以散粒噪声与仪器白噪声为主 ($\sigma \approx 1800$ ppm), 在时间上近似独立, 三次凌星信号清晰可辨。(b) 红噪声 (时间相关噪声): 在同量级噪声背景下, 采用一阶自回归过程 AR(1) ($\rho = 0.985$) 模拟时间相关红噪声, 并叠加于注入的凌星信号以构造对比示意。结合 *Kepler* 长时标采样间隔 $\Delta t \approx 29.4$ min, 对应的相时间为 $\tau_{\text{corr}} \approx -\Delta t / \ln \rho \approx 32$ h, 显著长于凌星持续时间 ($T_{14} \approx 3.694$ h)。该相关时间尺度可视为 *Kepler* 光变曲线中恒星活动与仪器系统误差所对应典型低频相关时标的代表性取值, 同时显著长于凌星持续时间, 从而构成与凌星信号相竞争的低频背景, 用于示意时间相关噪声对检出显著性的影响^[25]。

图 3 系外行星凌星信号在白噪声与红噪声背景下的可见性对比 (宿主恒星: KIC 2853093 / *Kepler*-767; 行星: *Kepler*-767 b, KOI-1546.01)

(例子中取八个) 奇异值对应的特征向量作为该通道的 CBVs, 这些正交基向量刻画了通道内共有的仪器效应模式。(3) 系统趋势的拟合与消除: 以目标星 KIC 7889349 为例, 利用上述 CBVs 的线性组合对 SAP 通量中的系统趋势进行最小二乘拟合并扣除, 得到经 PDC 校正的 SAP 通量, 即预搜索数据校正简单孔径测光 (Pre-search Data Conditioning Simple Aperture Photometry, PDCSAP) 通量; 该流程在有效抑制长期仪器漂移的同时保留了恒星固有的测光变化特征。值得注意的是, PDC^[38] 还采用了贝叶斯最大后验估计 (Bayesian Maximum A Posteriori, PDC-MAP), 该方法在构建去共趋势模型时引入了先验分布约束。与传统的 PDC-LS 相比, PDC-MAP 能显著降低噪声的干扰和信号的失真, 同时更好地保留真实的天体变化特征, 从而显著改善了微弱信号的恢复效果。基于 PDC-MAP 方法的发展, 管线进一步引入了 PDC-msMAP^[39] (multiscale MAP, 多尺度 MAP), 该方法使用多尺度分解 (例如小波变换 / 多分辨率滤波) 将光变曲线按不同时间尺度分解, 进而分别对不同尺度的信号应用 MAP 拟合, 从而在去除低频系统性趋势的同时更好保留高频真实天体信号, 是 *Kepler* 最终发布的 DR25 数据中默认的共趋势校正方法^[40]。

除 *Kepler* 官方的 PDC 系列方法之外, 还有多种去共趋势方法, 它们在原理和适用场景



注: 图示基于 SVD 的去共趋势处理流程。(上) 第五季度第 42 通道八条标准化 SAP_FLUX 光变曲线集合; (左下) 从样本集合中提取的前八个基向量; (右下) 以 KIC 7889349 为例, 展示原始流量 (SAP_FLUX) 及经系统误差校正后的光变曲线 (PDCSAP_FLUX)。

图 4 基于奇异值分解的去共趋势预处理流程示意及示例结果

上各有侧重。一类源自地面大视场巡天项目, 这些项目受大视场、天气和大气扰动等影响更大, 因此提出了 SYSREM^[41, 42] 等线性去共趋势方法, 通过迭代拟合或线性组合参考光变曲线来去除共同噪声, 这些方法在地面巡天中用于初步去共趋势但并非针对 *Kepler* 的高精度光变数据设计; 另一个方向是基于统计建模的趋势校正, 例如 K2SC^[43], 利用高斯过程回归同时拟合系统性误差与真实信号, 特别适用于 K2 任务中因反作用轮失效导致的复杂轨道相关系统噪声。还有一些方法采用局部对比或稳健贝叶斯技术来增强趋势校正性能, 例如 DOHA^[44], 它基于改进的差分光度学, 通过两步相关性搜索为每个目标光变曲线选取最匹配的对星并构建主对比光变来去除系统性误差, 与 PDC 方法统一构建全场共趋势基向

量不同, DOHA 更侧重目标与局部对比样本之间的直接相关性, 因此在校正短时和长期系统性噪声方面表现出良好效果; 还有基于贝叶斯框架的稳健扩展, 例如 ARC2^[45] 利用收缩先验对 *Kepler* 官方共趋势基向量系数进行正则化拟合, 从而减少过拟合并更好保护真实信号。整体来看, 这些方法虽然都旨在更有效地校正系统性噪声, 但在共趋势建模的假设、噪声适配策略以及对真实天体信号的保护机制上各有侧重, 体现出从传统线性去趋势技术到现代贝叶斯正则化与非线性统计建模的一系列方法演进。

实际上, 在共趋势去除过程中存在两个本质上紧密相关的噪声问题。其一是恒星自身的时间相关噪声或红噪声, 这种噪声随时间尺度延展会违背传统建立在独立同分布白噪声上的统计假设 (即假定噪声在时间上不相关且方差恒定)。具体而言, 光变曲线中的红噪声主要会由恒星活动引起, 并且具有时间相关性, 这些成分无法被简单的白噪声模型捕捉到。并且 Pont 等人在 2006 年的研究表明, 在存在红噪声的光变曲线中, 传统基于白噪声假设的显著性统计会被高估, 导致检测器对微弱信号的显著性估计偏乐观, 并增加噪声伪信号的误报率^[25]。其二是系统性误差与真实微弱信号的时间尺度重叠问题。去共趋势的基本假设是大多数系统性误差在不同恒星之间具有显著共性, 而真实的凌星信号是个体性且具有确定周期的事件。然而, 在实际观测中, 不同来源的系统性误差在不同恒星的数据中并不完全一致, 这些误差的表现强度和频率结构会随着恒星特性或观测条件的变化而变化, 而其中某些误差的时间尺度可能与长周期凌星事件相近, 这种噪声与信号在时间尺度上的重叠会使得基于全局共趋势模型难以准确区分噪声与真实信号^[38, 46]。

上述噪声问题对长周期微弱信号的检出构成系统性挑战, 尤其是过度去趋势可能导致真实信号被吸收、噪声结构错误建模引发参数估计偏差等 (详见第4节讨论)。因此, 去共趋势方法需要在噪声剥离与信号保护之间达到平衡。

为了解决去除共趋势过程中的噪声问题, 从而更好地搜寻低信噪比和长周期的凌星信号, 数据驱动的去共趋势方法通过构建更具表达能力的噪声模型, 有效地去除系统性误差并保护真实信号。一类典型方法是因果像素模型^[47] (Causal Pixel Model, CPM), 该模型在像素级别建模, 通过利用来自同一视场的其它像素时间序列来预测目标像素的系统性误差, 同时严格控制可能含有信号的数据泄露到拟合集中, 从而有效分离噪声与信号, 尤其在保护长周期信号方面具有明显优势。另一类代表性方法是 TERRA^[48], 应用至 $12.5 < K_p < 13.5$ 且 $CDPP_{12} < 40$ ppm 的 *Kepler* 目标时, 其在 ~ 12 小时时间尺度上的残差噪声可降至约 33 ppm, 从而使 ~ 100 ppm 量级的地球大小行星凌星信号达到约 3σ 的单次检出水平^[48], 该方法通过数据驱动的噪声建模, 有效缓解时间相关红噪声、系统误差及恒星本征变异 (Star Variability) 的影响。

这些数据驱动的去共趋势方法通过更细粒度的噪声模态提取和建模, 有效缓解了时间相关红噪声与系统性误差重叠问题, 为长周期微弱信号的稳健检测提供了更可靠的技术基础。

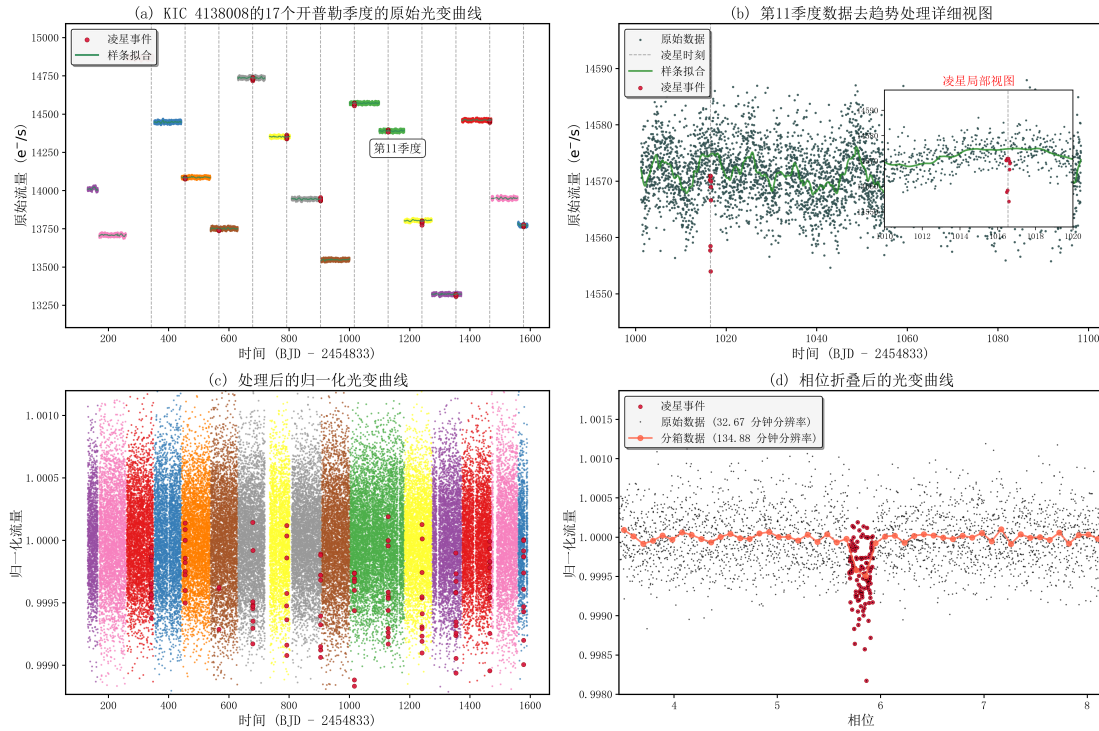
2.3 去趋势预处理方法

在完成去共趋势后仍需要对光变曲线进行专门的去趋势 (Detrending) 处理^[23, 37]。这是因为去共趋势预处理是基于全视场或 CCD 通道多个目标光变曲线之间的共性噪声模式的提取, 其核心假设是系统性误差在不同恒星中具有显著共性, 但这种假设不能完整描述在单条光变曲线中由恒星自身活动、仪器非共性漂移或时间相关噪声等引起的趋势结构^[38]。当这些趋势成分未被有效剥离时, 周期性检出算法的功率谱背景将升高, 长周期、浅信号的显著性容易被淹没或误判为噪声结构, 最终导致检出敏感度与可靠性下降^[34, 35]。

传统的去趋势方法主要基于参数化拟合或滤波技术, 在早期凌星搜索管线中被广泛用于趋势建模与剥离。典型的趋势剥离包括多尺度多项式拟合或样条函数 (Splines)、局部多项式滤波等, 其中 Savitzky-Golay 滤波器作为一种经典的局部多项式平滑滤波方法, 通过拟合子区间内的低阶多项式来实现数据平滑而尽量保留高频信号, 有助于剥离长期趋势^[35]。滑动中值滤波或窗口中值滤波同样是常用的低通趋势去除方法, 通过计算指定窗口内的中值来平滑光变曲线, 但窗口尺寸的选择至关重要; 过大的窗口可能抹平浅的凌星事件, 过小的窗口则无法显著去除低频波动^[49]。这些滤波方法由于对时间相关噪声和红噪声的建模能力有限, 在处理长周期、低信噪比的光变数据时, 往往会残留具有短记忆特征的自相关结构^[50, 51]。

为了有效应对时间相关噪声与红噪声的影响, 统计建模方法如高斯过程回归 (Gaussian Process Regression, GPR) 已成为去趋势处理的核心手段之一。GPR 通过核函数灵活建模时间相关噪声结构, 无需预设噪声形式, 并能在剥离趋势的同时保留周期性信号, 这对于长周期浅信号的保护尤为重要^[43]。常用核包括平方指数核、Matérn 核以及适用于恒星活动引起的准周期光变的准周期核 (Quasi-periodic Kernel); 红噪声相干时标与恒星自转周期和对流翻转时间相关^[29], 采用物理驱动的核函数可更好分离信号与噪声^[52]。例如, 针对 K2 光变曲线, K2SC 方法利用 GPR 同时拟合仪器漂移与恒星活动, 从而提升了长周期信号的恢复能力^[43]。近年来研究还将 GPR 与贝叶斯嵌套采样结合用于长周期候选体的贝叶斯检验, 有效区分微弱真实信号与噪声模型^[53]。除了 GPR 外, 自回归积分移动平均 (Autoregressive Integrated Moving Average, ARIMA) 模型结合凌星梳状滤波器 (Transit Comb Filter, TCF) 在处理具有短记忆自相关性的噪声时已被表明较传统 BLS^[34]更敏感于微弱凌星信号^[54]。

随着机器学习的发展, 数据驱动方法为去趋势处理提供了更强的非线性模式建模能力。例如, Morvan 等人在 2020 年提出的 TLCD-LSTM 模型利用长短期记忆网络 (Long Short-Term Memory, LSTM) 在无需先验假设的前提下, 将可处理时间序列规模从高斯过程约 1000 时间步的上限扩展至更长序列, 并避免因分箱操作导致的信息损失, 从而实现更完整且精确的信号建模^[55]; 此外, 深度学习的应用进一步拓展了去趋势方法的建模能力, 例如基于生成对抗网络的方法 DCGAN (Deep Convolutional Generative Adversarial Network)。相较于主成分分析法在非线性建模能力与可扩展性上实现提升, 消除了对主成分数量这一关键超参数的依赖, 并通过生成逼真的无凌星数据与更优残差分布, 实现更稳健的信号与噪声分离^[56]; 进一步地, 基于 Transformer 的 DTST (Denoising Time Series Transformer) 框架在定量指标上取得直接改进, 相比传统滤波方法将去趋势后光变曲线的 MAE (Mean Absolute Error)



注: 图示 KIC 4138008 光变曲线的去趋势处理流程。(a) 17 个 *Kepler* 季度原始光变曲线及带凌星掩码的样条拟合; (b) 第 11 季度去趋势处理细节 (含局部放大图); (c) 去趋势并归一化后的拼接光变曲线; (d) 按周期折叠后的光变曲线, 其中红点为凌星事件, 橙色曲线为分箱结果。

图 5 *Kepler* KIC 4138008 带凌星掩码的分段样条去趋势处理示意

从约 0.244% 降至 0.235%、IQR (Inter-Quartile Range) 从约 0.393% 降至 0.385%, 同时在推理速度上提升约 5–17 倍, 实现了更低残余噪声与更高计算效率^[57]。

除了单一算法的选择, 为了保护长周期浅信号免受误删, 研究者提出了多种策略^[35, 58]。例如在迭代搜索管线中采用凌星掩码, 避免真实信号被趋势模型吸收; 分段去趋势通过将长时序切分成多个段分别处理, 以捕捉局部系统性误差变化; 自适应窗口选择则利用工具包如 *Wotan* 提供的多项去趋势算法和动态窗口调整, 在不同信噪比阶段自适应优化趋势剥离效果。图 5 展示了在 KIC 4138008 的光变数据中应用带凌星掩码的分段样条拟合去趋势处理的效果。图中, 子图 (a) 展示了原始光变曲线与拟合的趋势模型, 显示出多个数据点在不同时间段的波动; 子图 (b) 则通过放大局部区域, 展示了第 11 季度数据的趋势剥离效果, 表明该方法有效避免了真实信号被误删的风险; 子图 (c) 与 (d) 则分别展示了去趋势处理后的数据与相位折叠后的结果, 进一步验证了分段去趋势的有效性。上述策略与图示表明去趋势效果得到明显改进, 显著提高了长周期候选体的检出完备度与可靠性。

综上所述, 从基础的时域滤波与参数化拟合, 到统计建模、机器学习及集成策略, 去趋势方法经历了从简单滤除趋势到兼顾噪声建模与信号保护的演进。其中, 对于长周期、低信噪比信号, 合理的去趋势方法不仅能剥离趋势噪声, 还能在不破坏周期性信号形状的前



注：箭头表示典型凌星巡天处理主链；各任务具体模块命名可能不同，但“系统噪声剥离—时域趋势去除—周期性搜寻—候选体处置”的逻辑环节具有共性。

图6 空间测光凌星搜寻主流程示意（从原始光变到候选体验证）

前提下提高检出性能，为后续的周期性搜索与信号鉴别打下坚实基础^[23]。

3 凌星信号搜寻

3.1 经典凌星搜寻流程管线

本文第2节所分析的光变噪声来源与数据预处理流程，为后续标准系外行星搜寻管线提供了质量可靠的光变曲线数据。在此基础上，于完成共趋势校正与去趋势化处理的光变曲线上对阈值越限事件（Threshold Crossing Event, TCE）进行初步搜寻与信号分类，构成了整个系外行星搜寻流程中的初筛（Triage）与分类（Vetting）环节。

在大规模空间测光巡天开展之前，凌星信号的初步检出主要依赖盒状最小二乘（BLS）算法^[34]及其变体。该方法在白噪声近似、严格周期性及盒状凌星模板等假设下，对去趋势后的光变曲线进行相位折叠与模板匹配。Jenkins 在早期研究中指出^[59]，凌星光度数据中的噪声往往呈现非白、非高斯特性，传统白噪声假设难以成立，亟需在噪声建模与检测统计层面进行系统性改进。基于此，Kepler 科学处理中心（Science Operations Center, SOC）在官方数据处理管线中实现了凌星行星搜寻模块（Transiting Planet Search, TPS）^[24]。该模块采用基于小波的自适应匹配滤波框架，通过对恒星光变曲线中时间相关噪声进行时变估计与白化处理，将含有色噪声的观测数据转化为近似白噪声过程，并在此基础上对一系列凌星模板执行匹配滤波，计算单事件统计量（Single Event Statistic, SES）并在试探周期下相干叠加得到多事件统计量（MES），以实现周期性凌星信号的统计检测。TPS 采用 7.1σ 的 MES 阈值，并要求识别到至少三次凌星事件。在为期四年的任务以及约 2×10^5 颗目标星的统计框架下，这一阈值的设定旨在将由纯高斯噪声触发的假探测信号 TCE 期望数控制在一个以下^[24, 40, 59]。随着任务推进，TPS 模块引入卡方一致性检验与统计自举（Statistical Bootstrap）等方法，以抑制由非高斯噪声引起的伪高 MES 事件，并在 Q1–17 全基线数据上形成了第 25 次数据发布（Data Release 25, DR25）终版搜寻目录^[40]。在该流程中，通过引入物理意义明确的临边昏暗凌星模板、非线性模型拟合及迭代式多行星搜索算法，实现了对多行星系统的系统性识别。结合数据验证（Data Validation, DV）与自动候选体筛选程序（Robovetter）分类框架^[8]，构建了“TPS 筛选出 TCE - DV - KOI 目录”的完整流水线。除官方管线外，凌星类地系外行星鲁棒降噪算法（Transiting Exoearth Robust Reduction Algorithm, TERRA）、准周期自动化凌星搜索算法（Quasi-periodic Automated Transit Search, QATS）与 TLS 等算法分别从噪声处理、周期扰动容许度及物理模板层面对传统假设进行了拓展^[6, 35, 48, 60]，作为

独立或后处理手段对 *Kepler* 数据的检出能力提供了重要补充。图 6 概括了从原始测光到候选体筛选的主干流程。

尽管上述方法显著提升了 *Kepler* 数据在中短周期区间的检出完备度与可靠性, 但在长周期、低信噪比凌星探测场景下仍存在若干结构性限制: TPS 的三次凌星触发条件与固定 MES 阈值对长周期候选体形成硬性约束; 同时, 长时间尺度光变去趋势在信号吸收与残余相关噪声之间存在不可避免的权衡; 此外, 基于简化凌星模板的匹配滤波在低信噪比条件下与真实凌星形态存在系统偏离, 进一步降低检测统计量。

3.2 盒状最小二乘 (BLS) 算法

盒状最小二乘算法 (BLS) 由 Kovács 等人^[34]在系外行星凌星探测研究初期被提出, 其核心动机在于解决传统周期搜索方法难以有效识别行星凌星信号这一关键问题。早期广泛使用的傅里叶分析^[61]及相位离散最小化 (Phase Dispersion Minimization, PDM)^[62]方法通常假设信号近似为正弦或可由少数谐波描述, 而行星凌星信号本质上呈现为短时、非正弦的周期性凹陷结构。尤其对于长周期、低信噪比的凌星事件, 其持续时间相对轨道周期更短、凌星深度更弱, 导致信号能量在频域中分散到多个高阶谐波, 从而显著降低传统方法的探测效率。BLS 通过引入针对凌星信号特征的盒状函数模型, 并在不同试周期下对折叠光变曲线进行最小二乘拟合, 一定程度上提高了对浅而短暂周期性凌星事件的检出能力, 因而成为早期大规模光变巡天中系外行星凌星搜索的基础方法之一。

该方法是一种用于在不规则时间采样光变数据中搜索行星凌星信号的周期分析技术, 其理论框架建立在一系列理想化但相互一致的统计与物理假设之上: (1) 目标信号具有严格的周期性, 即存在唯一且稳定的周期 P_0 , 使得凌星事件在每个周期内以相同的形态和深度重复出现, 从而不考虑凌星时间变化、周期漂移或轨道演化等效应。(2) 在信号形态上, BLS 采用盒状模型近似真实凌星过程, 认为一个周期内的信号仅在两种离散的亮度水平之间切换: 凌星期间的低亮度水平 L 与非凌星期间的高亮度水平 H 。该近似忽略了入凌和出凌阶段、恒星临边昏暗效应以及由偏心轨道或冲击参数变化导致的非对称结构, 其目的在于以最简模型最大化对短暂、弱周期性信号的检出效率。(3) 凌星持续时间相对于轨道周期较短, 并用无量纲参数 $q = T_{\text{transit}} P_0^{-1}$ 进行刻画, 通常将其限制在远小于 1 的范围内 ($\approx 0.01 - 0.05$)。在一般情况下, 模型由周期 P_0 、持续时间比例 q 、凌星亮度水平 L 、非凌星亮度水平 H 以及凌星历元 t_0 (即参考凌星中心时刻) 五个参数表征, 但在假定信号在一个周期内的时间平均值为零的条件下, L 与 H 之间满足 $qL + (1 - q)H = 0$, 从而可将 H 表示为 $H = -Lq(1 - q)^{-1}$, 使模型的自由参数数目减少至四个。(4) 在统计假设方面, BLS 假定观测数据由真实信号与服从正态分布的独立噪声线性叠加而成, 噪声均值为零, 不同观测点之间相互独立, 但允许噪声方差随观测而变化。在此框架下, 算法采用观测误差倒数平方作为权重并进行归一化处理, 即 $w_i = \sigma_i^{-2} [\sum_{j=1}^n \sigma_j^{-2}]^{-1}$, 从而在高斯噪声假设下实现极大似然意义上的加权最小二乘估计, 并进一步假设 (5) 加权后的观测数据满足零均值条件 $\sum_{i=1}^n w_i x_i = 0$ 。

在周期搜索过程中, BLS 通过在离散的试探周期网格上逐一对光变时间序列进行相位折叠, 并在每个试探周期下对折叠数据进行盒状凌星模型的最小二乘拟合。该方法假设每个

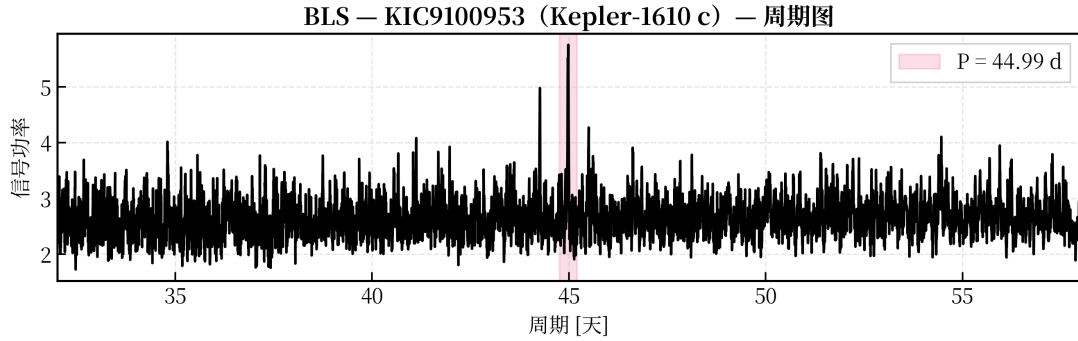
轨道周期内至多存在一次凌星事件，其周期显著性通常通过信号残差（Signal Residue, SR）等统计量来刻画，即由引入盒状模型相对于无凌星的常数模型所带来的残差平方和的改进程度。在显著性评估中，BLS 默认样本数量足够大且噪声近似为白噪声，从而保证统计量分布的近似有效性。从统计意义上看，BLS 因而可被严格解释为一种以盒状函数为信号假设的匹配滤波方法，其统计量在白噪声条件下与匹配滤波信噪比具有一致的最优性性质。由此，BLS 的方法定位得以明确：它并非旨在精确描述凌星光变的物理细节，而是通过引入最小化的信号形态假设，在保证统计可解释性的前提下，实现对不规则采样光变数据中周期性凌星信号的高效、稳健检出。在统计计算中，BLS 的检测量本质上等价于对折叠相位空间中凌星区间加权平均亮度偏离零值的度量。更严格地，在加权最小二乘框架下，若记

$$s = \sum_{i \in \text{tr}} w_i x_i, \quad r = \sum_{i \in \text{tr}} w_i,$$

则其检测统计量（或与之单调等价的形式）可写为

$$\Delta\chi^2 \propto \frac{s^2}{r(1-r)} = \frac{(\sum_{i \in \text{tr}} w_i x_i)^2}{(\sum_{i \in \text{tr}} w_i)(1 - \sum_{i \in \text{tr}} w_i)}. \quad (2)$$

在凌星占空比很小的情形（ $q \ll 1$ ）下，常用近似为将分母中的 $1-r$ 取为 1，得到与文献中常见的简化检测量等价形式；求和均在凌星区间内进行。在白噪声近似下，该统计量对应的有效信噪比满足 $\text{SNR}_{\text{BLS}} \sim \frac{\delta}{\sigma} \sqrt{N_{\text{tr}}}$ ，其中 δ 为单次凌星的平均深度， N_{tr} 为凌星区间内的有效采样点数。图 7 展示了 BLS 在目标 Kepler-1610 c 上 30 至 60 天的试验周期功率。



注：图示展示了 BLS 算法在目标 Kepler-1610 c 上 30 至 60 天的试验周期功率，其中 P 为一颗真实的已确认的系外行星的周期。

图 7 BLS 周期功率图

BLS 的统计效率依赖于信号在相位空间中的高度集中性、足够多的重复次数以及盒状模板与真实信号之间的良好匹配。严格的周期性假设要求通过相位折叠累积信号增益，但长周期系统在有限观测时间内可覆盖的完整凌星次数 $N_{\text{tr,tot}} \sim T_{\text{obs}}/P_0$ 随周期增大而迅速减少；BLS 等盒状近似方法由于将信号限制在理想的阶跃函数空间，难以精确刻画弱信号下真实光变曲线的平滑连续特征，可能导致信号振幅被系统性低估；此外，BLS 对独立高斯白噪声的假设在长时间尺度上并未显式考虑时间相关噪声或低频系统误差的影响。

正是基于对这些内在局限性的认识, 后续研究发展出多种 BLS 变体, 并进一步提出了采用物理驱动凌星模板的凌星最小二乘 (TLS) ^[35] 方法, 以在保持周期搜索框架的同时提高对长周期和微弱凌星信号的统计检出能力。Shahaf 等提出快速折叠版 fBLS, 显著提升速度 ^[63], 同时基于 GPU 的实现 (如 GPU-BLS) 可进一步加速 BLS 搜索。Collier Cameron 等提出混合算法, 将匹配滤波与 BLS 结合以提升地面巡天检出效率 ^[64]。Carter 与 Agol 提出 QATS, 放宽严格周期假设, 允许相邻凌星间隔在界内变化, 适用于多行星系统动力学扰动导致的准周期凌星 ^[60]。

3.3 凌星最小二乘 (TLS) 算法

随着 *Kepler* 等空间测光任务数据质量的提升以及对小半径、类地行星 (浅凌星) 探测需求的增加, BLS 所采用的盒状函数近似逐渐成为限制探测灵敏度的关键瓶颈。从物理角度看, 由于恒星临边昏暗效应以及行星入凌和出凌阶段的渐进过程, 真实的凌星光变呈现为平滑的 U 形结构, 而非理想化的矩形阶跃。基于上述背景, Hippke 与 Heller (2019) 提出了凌星最小二乘法 (TLS) ^[35], 其核心思想在于将搜索模板从几何盒状函数推广为基于物理的凌星模型, 从而在保持周期搜索框架的同时提高对微弱和长周期凌星信号的统计检出能力。与 BLS 的简化盒状模型不同, TLS 显式引入恒星临边昏暗效应, 通常采用二次临边昏暗律

$$I(\mu) = 1 - u_1(1 - \mu) - u_2(1 - \mu)^2 \quad (3)$$

其中 $\mu = \cos \theta$ 表示视线方向与恒星表面法线之间的夹角, u_1 与 u_2 为临边昏暗系数。TLS 假设行星凌星光变可以由理想化的几何与辐射模型准确描述: 恒星与行星均为球对称天体, 行星轨道为开普勒轨道且在凌星阶段可近似为匀速直线运动, 凌星光变遵循 Mandel 与 Agol 提出的解析模型 ^[65]。这一物理化建模使得 TLS 的模板空间从 BLS 的一维盒状函数族扩展到多维物理模型族, 从而对“行星型凌星”这一类物理信号具有更高的敏感性。

在算法实现上, TLS 利用恒星质量 M_* 和半径 R_* 作为先验信息, 通过开普勒第三定律对凌星持续时间施加物理约束, 从而显著缩小参数搜索空间。对于给定的试探周期 P , TLS 对光变数据进行相位折叠, 计算模型光变 $y_{m,i}(P, t_0, d)$ 与观测光变 $y_{o,i}$ 之间的 χ^2 统计量:

$$\chi^2(P, t_0, d) = \sum_{i=1}^N \frac{[y_{m,i}(P, t_0, d) - y_{o,i}]^2}{\sigma_i^2} \quad (4)$$

其中 σ_i 表示第 i 个数据点的测光不确定度。TLS 并非对所有模型参数进行完全数值优化, 而是通过对凌星深度进行解析求解, 将原本高维的非线性最小二乘问题降维为在周期 P 、凌星历元 t_0 及持续时间 d 上的系统性搜索, 这一策略在保证拟合精度的同时使得算法在计算上保持可行。对于每一个试探周期 P , TLS 在所有测试的 t_0 与 d 组合中记录最小值 $\chi_{\min}^2(P)$, 并在整个参数空间内确定全局最小值 $\chi_{\min, \text{glob}}^2$ 。在此基础上, TLS 定义信号残差 (SR) 为

$$SR(P) = \frac{\chi_{\min, \text{glob}}^2}{\chi_{\min}^2(P)} \quad (5)$$

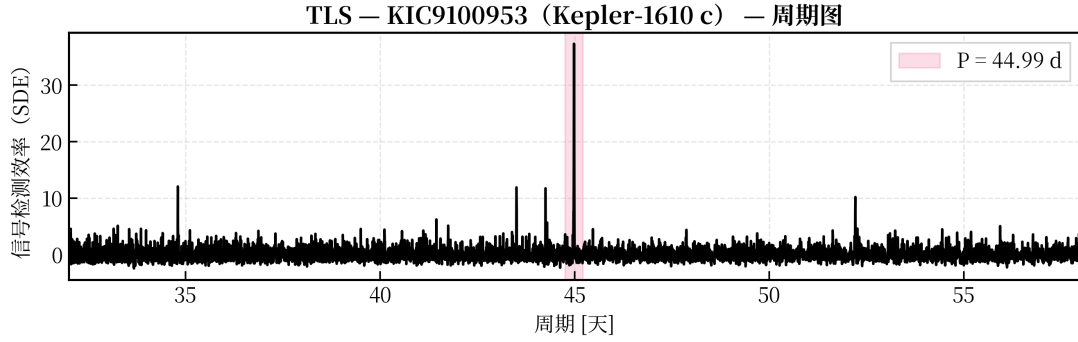
其物理含义是衡量某一试探周期下模型拟合相对于全局最优拟合的相对优劣。进一步地,

通过对所有试探周期的 $SR(P)$ 分布进行标准化, TLS 构造信号探测效率 (Signal Detection Efficiency, SDE):

$$SDE(P) = \frac{1 - \langle SR(P) \rangle}{\sigma[SR(P)]} \quad (6)$$

其中 $\langle SR(P) \rangle$ 与 $\sigma[SR(P)]$ 分别表示 SR 分布的均值和标准差, 显著的 SDE 峰值对应于潜在的凌星周期信号。

在周期采样方面, TLS 采用基于物理约束的非均匀频率网格: 在给定过采样参数 OS (通常取 2–5)、观测时间跨度 T_{total} 、最少凌星次数 N_{transit} 以及恒星质量 M_* 、半径 R_* 先验时, 由最小/最大搜索频率、常数 A 及序列 $\{f_x\}$ 构造试探频率, 进而取 $P_x = 1/f_x$; 推导与显式公式见附录A及文献^[35, 66]。该策略在计算效率与搜索完整性之间取得平衡。图 8 展示了 TLS 在目标 *Kepler*-1610 c 上 30 至 60 天的试验周期功率。



注: 图示展示了 TLS 算法在目标 *Kepler*-1610 c 上 30 至 60 天的试验周期功率, 其中 P 为一颗真实的已确认的系外行星的周期。

图 8 TLS 周期功率图

从方法论角度看, TLS 通过将物理化凌星光变模型嵌入最小二乘搜索框架, 实现了对微弱周期性凌星信号的高灵敏度检测。在高斯白噪声且模型假设成立的理想条件下, 该最小二乘统计量在检测意义上近似等价于匹配滤波^[59, 67, 68], 因此在低信噪比区域相较 BLS 具有更高的探测效率。数值实验表明, 在地球大小行星的典型搜寻场景中, TLS 的信号探测效率可提升约 17%, 并在相同误报率水平下恢复更多真实信号^[35]。TLS 的优势依赖于凌星事件在整个观测基线内保持稳定的周期与形态, 使得多次凌星能够通过相位折叠实现相干叠加。

对于有限观测时间内仅发生少数几次凌星的长周期系统, TLS 与 BLS 一样难以获得充分的统计增益。在极低信噪比条件下, 真实光变数据中普遍存在的低频相关噪声与恒星活动残余, 若未在预处理阶段得到有效抑制, 可能在相位折叠后与物理化凌星模板产生形态相似性, 从而引发非行星起源的伪信号。TLS 搜索所需的最优试验频率数量满足 $N_{\text{freq, opt}} \propto S(f_{\text{max}} - f_{\text{min}})$, 其中 S 为观测时间跨度。在固定频率范围下, 有 $N_{\text{freq, opt}} \propto S^{[35]}$ 。由于每个试验频率均需对全部 N 个数据点进行统计计算, 且均匀采样时 $N \propto S$, 因此总体计算复杂度满足

$C \propto N_{\text{freq,opt}} \times N \propto S^2$ 。这表明, 当时间跨度增加或搜索扩展至长周期范围时, TLS 的计算成本将显著上升。这一特性在实际应用中限制了 TLS 在长周期、极低信噪比情形下的可扩展性, 并促使后续研究转向 GPU 并行化 (如 GPU-TLS^[69]) 与机器学习等替代方案, 例如 GPU 相位折叠卷积神经网络 (GPU Phase Folding algorithm combined with a Convolutional Neural Network, GPFC)^[70] 与剑桥系外行星凌星恢复算法 (Cambridge Exoplanet Transit Recovery Algorithm, CETRA)^[71]。

3.4 自动化凌星搜寻及标准化流程

上述传统最小二乘算法 (BLS 与 TLS) 的核心局限在于其依赖相位折叠与模板匹配的检测范式。对于长周期微弱凌星信号, 这一范式面临统计增益不足 (凌星次数少)、计算代价随周期平方增长、以及低信噪比下模板与噪声形态混淆等多重挑战。

为突破上述局限, 研究者逐渐将目光转向机器学习与深度学习方法。这些数据驱动的方法不依赖于显式的物理模板匹配, 而是通过学习大量标注样本中的复杂模式, 自动提取能够区分真实凌星与假阳性的深层特征, 在集成多域信息、提升计算效率以及自动抑制假阳性等方面展现出相对于传统算法的显著优势。深度学习模型进一步通过多层非线性变换, 能够捕捉光变曲线中的长程依赖关系与复杂时序模式, 为长周期微弱信号的自动化搜寻提供了新的技术路径。

3.4.1 机器学习方法

根据对标注数据依赖程度的不同, 凌星信号自动识别领域中的传统机器学习方法通常可划分为无监督学习与有监督学习两类。前者在缺乏人工标注样本的情形下, 侧重于从观测数据本身挖掘统计结构与形态特征; 后者则依赖已知标签, 通过构建判别模型实现自动分类。

在无监督框架下, 研究工作大体可分为相似性度量驱动的筛选方法与基于聚类的结构划分方法两种思路。Thompson 等^[72]提出了一种结合降维技术与 k 近邻 (k -Nearest Neighbor, KNN) 方法的度量, 用于判断周期性光变信号是否具有凌星形态。在对 *Kepler* 数据的测试中, 该指标在剔除超过 90% 的非凌星信号的同时保留了超过 99% 的已知候选体, 验证了这一度量在大规模数据预处理中的实用性。为进一步区分真实凌星信号与系统误差等干扰信号, Mullally 等^[73]提出了一种基于模型选择的统计判别框架, 该方法通过比较真实凌星模型与系统误差模型的拟合优度, 并利用贝叶斯信息准则 (Bayesian Information Criterion, BIC) 评估每个 TCE 的为真实凌星信号的置信度。在处理更复杂形态和数据空间结构特征方面, 研究者引入了更具拓扑保持能力的聚类方法, Armstrong 等^[74]采用自组织映射 (Self-Organizing Map, SOM) 对 *Kepler* 与 *K2* 任务的光变曲线形状特征进行无监督学习, 并基于形状特征在不依赖预先标记数据的情况下, 真实行星信号与假阳性候选有效区分的成功率达 87%。与此同时, 针对光变曲线中残差数据的异常点剔除, Mislis 等^[75]提出的残差数据异常时间序列分析框架 (Time Series Analysis for Residual Data Irregularities, TSARDI) 利用基于密度的含噪声空间聚类算法 (Density-Based Spatial Clustering of Applications with Noise, DBSCAN) 对去趋势过程时的个别光变数据点进行识别与排除, 从而提高了基于注入模拟信号的整体检

测效率,在宿主星等小于 12 的样本上,该方法对注入信号的检测效率可达约 80%。

有监督学习框架下,模型通过训练带标签的数据样本,学习从观测特征到目标标签的映射关系。在概率性系外行星验证中,Armstrong 等提出了一种结合辅助模型改进的高斯过程分类器 (Gaussian Process Classifier, GPC),通过引入假阳性先验信息实现对候选信号的概率性评估,在区分已确认系外行星 (Confirmed Planets) 与假阳性 (False Positive, FPs) 时,可以达到每个样本 0.54 的平均对数损失 (Mean Log-loss)^[76]。在光变信号分类方面,McCauliff 等率先将随机森林 (Random Forest, RF) 应用于 *Kepler* 系外行星候选体,达到了 5.85% 的总体错误率^[77];随后多项研究在 *Kepler* 数据上进一步验证了决策树模型在候选筛选与特征判别中的有效性。针对 Gaia 数据, Panahi 等采用机器学习对恒星样本进行初步筛选,利用相对 BLS 速度更快的 SparseBLS^[78]方法进行二次筛选,经径向速度观测在 Gaia 首次发现了两颗系外行星 Gaia-1b 与 Gaia-2b^[79]。Malik 等^[80]利用 TSFresh (Time Series Feature Extraction on basis of Scalable Hypothesis tests) 从光变曲线提取 789 个时序特征后,使用 LightGBM 机器学习库构建梯度提升树 (Gradient Boosting Tree, GBT) 分类器,在 *Kepler* 数据集上实现 ROC (Receiver Operating Characteristic) 曲线下面积 AUC (Area Under the Curve) 0.948 和召回率 0.96。Melton 等^[81]开发的 DIAMante 自回归行星搜索 (DIAMante TESS AutoRegressive Planet Search, DTARPS) 首先利用 ARIMA 模型去除光变曲线中的自相关噪声,再通过凌星梳状滤波器 (TCF) 提取显著的周期信号,最终由 RF 对候选信号进行筛选,在训练数据集上实现了约 92.8% 的真阳性率和仅 0.43% 的假阳性率。

综上,传统机器学习在凌星候选筛选与验证中主要形成无监督与有监督两条技术路线。这些方法体现了传统机器学习与信号处理及周期搜索算法的紧密耦合,同时为后续深度学习端到端方法提供了可参考的基线。

3.4.2 深度学习方法

深度学习是机器学习领域中的一个重要分支,其核心在于构建包含多个隐藏层的神经网络结构,通过层次化的非线性变换,从原始输入数据中逐层抽象出更加丰富和复杂的特征表示。相较于传统机器学习方法,深度学习通过端到端的层次化特征学习机制,能够直接从原始数据中自动提取多层非线性特征,减少对人工特征工程的依赖。随着数据规模的增加,其建模能力不断提升,尤其在高维数据和微弱信号识别等复杂任务中,表现出更强的表征能力和泛化性能。卷积神经网络 (Convolutional Neural Network, CNN) 是深度学习的核心方法,它通过局部连接、权值共享以及池化操作,能够高效地学习从局部细节到全局特征的整合,这些优势使得 CNN 在微弱凌星信号的搜寻以及系外行星候选体的自动分类中成为重要的技术手段之一。

在 *Kepler* 真实观测数据上实现行星候选体与假阳性信号自动分类的首次系统性工作来自 Shallue 与 Vanderburg 提出的 AstroNet 模型,标志着深度学习方法在凌星信号判别中的系统性应用^[82]。后续的研究表明,在噪声显著的光变曲线中检测类地行星时, CNN 在准确率方面优于最小二乘等传统方法,进一步验证了深度学习在系外行星搜寻任务中的潜力^[83]。在此基础上, AstroNet 及其衍生架构被扩展至 K2 与 TESS 观测数据,构建了端到端自动筛选流程,并在不同观测任务与数据分布下验证了良好的泛化能力^[84-86]。

随着研究的深入, 学界逐渐认识到引入领域科学知识对提升模型性能具有重要作用。在 AstroNet 框架基础上, Ansdell 等提出 ExoNet 及其轻量化版本 ExoNet-XS, 通过融合光变曲线、恒星参数及质心信息等多源特征, 在低信噪比样本上的召回率 (Recall) 提高约 15%–20%^[87]。在此思路下, 基于物理先验与数据驱动相结合的分类框架得到进一步发展: Rao 与 Valizadegan 等提出的 ExoMiner 在保持约 99% 准确率 (Precision) 的同时实现了 93.6% 的召回率, 并成功验证 301 颗新行星, 显示出较高的可靠性与实用价值^[88]。进一步地, Valizadegan 等提出多重性提升框架 (Multiplicity Boost), 利用多重先验信息来优化所给定的任一分类器 (Classifier; Vetter, 如 AstroNet、ExoNet), 表明结合天体物理先验能够有效提高模型的鲁棒性与判别可信度^[89]; 其后续工作 ExoMiner++ 在 ExoMiner 框架基础上, 进一步拓展至 TESS 两分钟采样数据。该模型通过引入更多诊断性输入特征 (包括周期图、通量长期趋势、差异图像、相位展开通量以及航天器姿态控制信息等), 并结合迁移学习策略, 显著提升了对复杂假阳性来源的识别能力。相比原始 ExoMiner 模型, ExoMiner++ 在区分真实凌星信号与噪声源方面表现更为稳健, 并据此构建了新的自动化验证目录, 为后续跟踪观测提供了更高可信度的候选体优先级排序。

基于模拟光变曲线数据训练深度学习模型用来实现凌星信号检测的方法, 因其在训练样本规模、标签准确性以及参数空间覆盖范围等方面的显著优势, 被广泛应用至各项研究中^[70, 90–93]。例如, 利用 Lilith 模拟器构建 TESS 人工光变数据集并训练 ExoNet, 并应用至真实 TESS 数据上, 具有较好的泛化能力^[94]。此外, 将不同周期窗口的凌星信号堆叠构造二维特征并输入 CNN 进行分类的做法, 增强了对周期结构信息的表征能力^[95]。

长周期微弱凌星信号的分布缺失, 其中一个原因来源于有限的观测基线, 针对复杂噪声背景下的单次凌星信号检测, 也是填补这一空缺的方法之一。Dvash 等^[96]将凌星检测问题表述为时间序列语义分割任务, 基于 U-Net 构建一维卷积网络, 实现对光变曲线的逐点 (像素级) 概率预测。通过引入生成对抗网络及对抗损失以增强对低信噪比信号的判别能力, 该方法在复杂噪声背景下提升了对单次、浅凌星事件的检测灵敏度。近年来目标检测方法^[97]、语义分割与生成对抗网络 (如 U-Net 与 GAN)、生成模型、Transformer^[98] 以及残差网络 (Residual Network, ResNet) 等也被引入系外行星凌星信号识别任务。例如, 基于 Transformer 架构的模型被用于从 TESS 全帧图像光变数据中直接检测凌星信号, 无需进行相位折叠或假设信号周期, 从而能同时识别多次和单次凌星事件^[99]。面向未来巡天任务的 PLATO 数据, PANOPTICON (Probabilistic Automated Neural Optimization for Promoting Transit Identification and Characterization Of New-planets) 通过改进的 U-Net 结构在未经预先滤波的光变曲线中高效提取多尺度特征, 能够在低误报率条件下恢复超过 90% 的测试样本, 其中包括地球类单次凌星事件^[100]。

除此之外, WATSON-Net (Wavelet-based Automated Transit Search and Optimal Network) 以 vetting 和自动验证的神经网络的思想, 在 Kepler 和 TESS 数据上表现出高精度与良好召回率, 并作为开放源码模块集成到 SHERLOCK (Searching for Hints of Exoplanets Researching Led by Optimizing Counts) 管线中, 用于区分真实凌星信号与假阳性^[101]。DART-Vetter (Deep Learning Tool for Automatic Triage) 以 CNN 为基础构建了自动化凌星分诊工具, 能够从相

位折叠光变曲线中自动提取多维特征,并高效区分行星候选体与假阳性信号,显著降低了人工审查成本^[102]。在超短周期行星 (ultra-short-period planets, USPs) 搜寻方面,基于 GPU 相位折叠与深度学习的探测方法 (GPU Phase Folding and Deep Learning, GFPC) 由 Wang 等提出:通过对试验周期序列上利用 GPU 并行相位折叠,再经模拟光变曲线训练的 CNN 对批量相位折叠的信号进行评分,能够识别微弱凌星信号,并在 *Kepler* KG 型矮星数据中发现了迄今最短周期与最小体积的超短周期行星候选者 (SNR 约为 6-7)^[70, 93]。条件流匹配 (Conditional Flow Matching, CFM) 被用于信号降噪与重构,并与 CNN 形态特征及极端梯度提升 (eXtreme Gradient Boosting, XGBoost) 分类器协同构成生成-判别融合探测框架,显著提升了复杂背景下的探测性能^[103]。

表 1 从任务范式、数据来源、数据表征及样本构建策略四个维度总结了代表性研究。

大分类	细分类别	文献
阶段	搜寻	[70, 83, 90, 93, 95–97, 99, 100]
	分类	[82, 84–89, 98, 101, 103–105]
数据来源	<i>Kepler</i>	[70, 82–84, 88, 93, 101, 103]
	TESS	[85–87, 94, 98, 99, 101, 103–105]
	PLATO	[100]
数据集构建策略	全模拟光变曲线	[70, 90, 91, 93, 94, 100]
	半模拟光变曲线	[92, 96, 97, 99]
	真实光变曲线	[82, 86–88, 101, 103–105]
训练数据形式	相位折叠信号	[70, 82, 83, 87, 88, 90, 94, 95, 98, 101, 103]
	未折叠光变曲线	[91, 92, 96, 99, 100]
	全帧图像	[86]
	多模态 (光变、质心特征、图像等)	[87, 103, 105]

表 1 凌星探测与候选体判别深度学习方法分类 (数据来源中 *Kepler* 含 K2)。

3.5 深度学习搜寻标准化管线

本文在此基础上,参考 GPFC 框架^[70, 93],重点挖掘这类深度学习算法与经典系外行星搜寻管线的共性,标准化了目前在实际应用中表现最为突出的、以深度学习为核心的凌星信号搜寻流程^[69, 71]。该方法在试验周期扫描框架上与传统 BLS、TLS 相似,以深度学习模型作为评分准则实现端到端候选筛选。该标准化流程具有模型无关性,可作为各类深度学习凌星二分类 (仅判断给定信号存在凌星与否) 模型迭代与比较的通用实现框架。

假设在该以深度学习模型为准的凌星信号搜寻流程中,数据来源为去共趋势与去趋势化后的光变曲线,该数据符合较为理想的接近平稳噪声的状态,可能存在未被发现的行星:

- (i) **目标参数空间与数据分布的确定:** 根据望远镜观测数据的采样节律确定时间分辨率,并选定待搜索系外行星的轨道周期、凌星持续时间范围等参数;依据选定的凌星持续时间范围,确定相位折叠后的数据点分箱数量,以在保证时间分辨率足以表征凌星事件的同时提高信噪比。例如,对于 *Kepler* LC 数据,其采样间隔约为 29.4 分钟;在长

周期凌星搜寻中, 试验周期范围可设置为 100–150 天, 或更一般地限定为在观测时间基线内满足至少三次凌星事件出现的周期区间。

- (ii) **合成光变曲线的构建:** 为生成用于模型训练与评估的样本数据, 可采用以下三种策略构建人工或真实光变曲线: (1) 全合成方式: 将模拟的凌星信号嵌入高斯噪声序列或其他噪声模型生成的噪声序列, 生成全模拟的光变曲线; (2) 半合成方式: 将模拟凌星信号注入真实但已掩蔽凌星事件的光变曲线, 以获得兼具真实噪声特性与可控信号参数的半模拟凌星光变曲线; (3) 真实样本方式: 选取已确认的系外行星的凌星光变曲线。
- (iii) **凌星与非凌星样本的生成:** 凌星信号样本由上述凌星光变曲线经相位折叠与数据点分箱后得到, 其相位分布中包含明确的凌星特征; 非凌星信号样本则通过对不存在凌星的光变曲线相位折叠到错误的试验周期上生成, 以确保其不包含周期一致的凌星结构。
- (iv) **数据集规模与参数覆盖的确认:** 在样本构建过程中, 应确保凌星信号在轨道周期、凌星深度、持续时间及信噪比等关键参数空间内具有尽可能全面且均匀的覆盖, 以避免模型训练过程中产生显著的参数偏置。
- (v) **模型结构设计 with 实现:** 根据凌星检测任务的具体需求选定合适的深度学习模型结构, 并明确模型的输入输出形式与特征维度。其中, 模型输入长度需与相位折叠及分箱后的光变数据长度保持一致; 在此基础上完成模型搭建与参数初始化。
- (vi) **模型训练与验证:** 在构建完成的数据集上划分训练集与验证集, 设定合适的损失函数与优化器, 采用监督学习策略对模型进行训练; 通过验证集性能对超参数进行调优, 并结合早停 (Early Stopping) 等策略防止过拟合, 直至模型收敛。
- (vii) **凌星信号的周期搜索与评分:** 在实际搜寻阶段, 针对预设的周期搜索范围生成试验周期序列; 对每一个试验周期, 可采用 GPU 加速的相位折叠与数据分箱方法 (如 GPU 相位折叠或相关实现)^[69, 70], 快速构建对应的模型输入样本 (试验周期–光变对), 并交由训练完成的模型进行统一评分, 从而获得试验周期空间上的评分分布谱, 用以识别潜在的凌星信号候选体。

4 总结、当前挑战与未来展望

本文围绕长周期微弱凌星信号探测任务, 从光变噪声来源与数据预处理、系外行星搜寻、候选体验证等多个维度, 对凌星法与人工智能方法相结合的研究进展进行了系统梳理。

数据预处理是提升信号可探测性的关键环节之一 (第2节)。去共趋势算法由早期线性最小二乘方法 (PDC-LS) 发展至多尺度贝叶斯建模 (PDC-msMAP)^[37, 39], 并通过因果像素模型 (CPM) 与 TERRA 等数据驱动方法, 在像素级或全视场尺度上增强了系统误差的分离能力^[47, 48]。与此同时, 去趋势方法也由传统滤波逐步拓展至高斯过程回归及 LSTM、Transformer 等基于深度学习的时序建模方法^[43, 55, 57], 在噪声抑制与凌星信号保护之间实现更优权衡。

原始光变曲线经由去共趋势以及去趋势后, 凌星信号搜寻构成系外行星探测流程的核心环节 (第3.1节)。*Kepler* 官方 TPS 管线通过小波白化与多事件统计量 (MES) 阈值实现候选信号初筛, 盒状最小二乘 (BLS) 与凌星最小二乘 (TLS) 分别基于盒状模板与物理凌星

模型建立了周期搜索的经典范式^[34, 35]。在此基础上,传统机器学习方法(如随机森林、梯度提升树)被用于候选体筛选与分类。近年来, GPU 加速相位折叠结合深度学习的方案(如 GPFC、CETRA)显著提升了计算效率并拓展了长周期参数空间的搜索能力^[70, 71]; AstroNet、ExoNet、ExoMiner 等卷积神经网络模型在候选体分类中实现了较高精度与召回率^[82, 87, 88];基于 Transformer 的方法则进一步探索了无相位折叠条件下的凌星检测新路径^[98, 99]。

总体而言,尽管上述方法在中短周期参数空间内已实现较高的完备度与可靠性,但在长周期、低信噪比区间,从物理建模到信号检出再到假阳性控制,各处理环节仍面临一系列相互关联的核心挑战。

4.1 凌星法长周期微弱凌星信号的挑战

从几何概率角度看,对于轨道取向随机分布的行星系统,其发生凌星的几何概率近似与恒星半径与轨道半长轴之比成正比^[106, 107]。随着轨道半长轴的增大,满足凌星条件的轨道倾角范围迅速收缩,使得长周期行星发生凌星的概率急剧降低。以类太阳恒星为例,其宜居带行星的凌星概率仅约为 0.5%,从几何上对长周期行星的可探测性施加了严格限制。

在测光信号层面,其凌星深度与行星半径平方成正比。对于类太阳恒星上的地球大小行星,凌星深度仅约为八十四万个分之一,已接近 *Kepler* 任务的测光精度极限^[24, 32]。与此同时,宜居带类地行星的凌星持续时间通常可达十小时以上,其时间尺度与恒星活动(如黑子演化、粒状对流)以及仪器系统效应(如热变形、指向抖动引起的响应漂移)高度重叠,使得单次凌星事件极易与缓慢变化的背景信号相混淆,并在去趋势处理中被部分削弱甚至完全抹除^[29, 30]。基于注入-恢复测试, Burke 等人系统评估了 *Kepler* 的探测完备度,发现对于轨道周期超过三百天、行星半径约为一个地球半径的行星,其探测完备度已降至约 30% 以下,远低于短周期参数空间中 90% 以上的水平^[108],直接反映了长周期微弱凌星信号在观测层面的内在困难。

在数据分析与搜寻算法层面,时间相关噪声(红噪声)的存在进一步构成了长周期凌星探测的核心统计挑战。传统凌星搜索算法如盒状最小二乘(BLS)通常建立在白噪声假设之上,即假定各测光数据点之间相互独立^[34],从而使信噪比可通过多次凌星事件的相位折叠按观测次数的平方根增长。然而,实际光变数据中普遍存在显著的红噪声,其来源包括恒星本征活动在小时至数十天尺度上的准周期变化(如星斑与自转调制)^[29],以及仪器系统误差在长时间尺度上的缓慢漂移^[30]。这些噪声过程具有明显的时间相关性,导致有效自由度显著减少,从而破坏了基于白噪声假设的统计显著性评估,尤其对长周期、低信噪比的微弱凌星信号影响更为突出^[25]。已有研究表明,若忽略红噪声效应,长周期行星候选体的统计显著性将被系统性高估。Carter 和 Winn 通过蒙特卡洛模拟发现,长周期候选体的显著性可能被夸大一至两个数量级,大量由红噪声涨落产生的假信号因此被误判为行星凌星^[36]; Pont 等人亦指出,红噪声会导致凌星深度和时刻等参数不确定性被大幅低估,从而影响行星发生率等统计推断的可靠性^[25]。近年来,高斯过程(GP)建模被广泛用于描述红噪声的时间相关结构,并在贝叶斯框架下与凌星模型联合拟合,以提升长周期信号判别的统计可靠性^[43, 109]。Matesic 等人基于高斯过程与嵌套采样方法重新分析了 *Kepler* 目录中约两百颗小半径长周期候选体,发现其中约 30% 的候选体缺乏足够的贝叶斯因子支持,其凌星性质存

在明显不确定性^[53], 进一步表明在长周期参数空间内, 严格的红噪声建模与统计检验对于区分真实行星信号与噪声涨落具有决定性作用。

除几何概率与时间相关噪声的限制外, 长周期、微弱凌星信号还面临候选体筛选效率与假阳性控制的双重挑战。随着巡天规模扩大, *Kepler* 共产生超过 3.4 万个阈值越限事件 (TCE)^[8], 而 *TESS* 的数据量进一步增长^[110], 传统依赖人工审查的筛选模式会变得困难。为提升自动化与一致性, 机器学习方法已被广泛应用于凌星候选体分类^[82, 87], 在复杂噪声背景下展现出较高的识别能力。然而在长周期区间, 由于样本极度不平衡、凌星次数有限 (通常仅 2–3 次) 且形态多样性更强, 分类性能与统计置信度仍面临挑战^[85, 94]。此外, 恒星活动与仪器系统误差在长时标上产生的准周期结构更易导致误判^[111, 112]。因此, 对于长周期候选体, 仍需结合质心偏移检验及多手段后随验证, 以有效控制假阳性率^[113–115]。尽管近年来在短周期行星探测方面取得显著进展, 但针对长周期、小半径行星的自动化检测与可靠性评估方法仍相对不足。尤其是在低信噪比、少凌星次数条件下的信号识别与统计验证框架方面, 现有研究仍缺乏系统性的总结与方法比较。

4.2 未来展望

4.2.1 下一代空间任务的机遇

面向未来, 多项空间测光任务有望显著改善长周期参数空间的观测覆盖。*ESA* 的 *PLATO* 任务以探测和刻画明亮恒星周围的小半径行星为核心目标, 强调对类地与宜居带行星的敏感性, 并将通过多年高精度测光显著提升对中长周期行星的探测能力^[12]。其长时基、高稳定性和大样本观测设计均为长周期微弱凌星信号探测提供了关键机遇。

中国提出的系外地球巡天 (ET) 则以发现 “Earth 2.0” 为核心科学目标, 拟在日–地 L2 点部署由多台望远镜组成的空间天文台 (含凌星与微引力透镜载荷), 通过超高精度宽视场测光与长期连续监测, 系统提升对微弱凌星信号及长周期行星总体样本的检出能力^[13–15, 116, 117]。相关白皮书与任务综述均明确指出, ET 任务设计强调长期基线与大样本统计, 以拓展 *Kepler*/*TESS* 之后对类地行星参数空间的覆盖^[13, 15]。

与此同时, *TESS* 在扩展任务阶段持续重复观测多个天区, 并通过延长监测时长与多轮回访逐步增强对更长周期候选体的敏感性; 尤其在后续扩展任务规划中, 对部分天区的累积观测基线进一步增加, 为长周期候选体的发现与再确认提供了更有利的数据条件^[11]。

这些任务在带来重要科学机遇的同时, 也将产生更大规模的光变数据集, 对自动化数据处理与候选体识别提出更高要求。随着样本规模持续增长, 传统高度依赖人工审查的筛选模式将难以长期维持, 亟需发展高效、可靠且可扩展的自动化管线, 以支撑长周期、低信噪比候选体的系统搜寻与统计验证。

4.2.2 方法学发展方向

在数据预处理方面, 发展多尺度自适应噪声建模技术仍是关键方向。将去共趋势与去趋势过程在流程上更紧密耦合, 使噪声剥离策略能够随时间尺度与噪声特性自适应调整, 有望在抑制系统性误差的同时更好保留微弱凌星信号。基于 *LSTM* 等深度时序模型的数据驱动去趋势方法已展示出在不显式预设仪器系统误差形式条件下重建基线的能力^[55]。此外, 生

成式模型（如基于 GAN 的非线性重建方法）在天体数据去趋势中的应用也显示出潜力，但目前相关工作更多集中于透射光谱等时序谱数据场景，其向巡天光变预处理任务的迁移能力仍需进一步验证^[56]。

在信号搜寻方面，GPU 加速相位折叠过程与深度学习分类的组合管线代表当前方向下较为先进的方法。以 GPFC 为代表的方法验证了这类框架在大规模搜寻中的高效性^[70, 93]；同时，该系统还成功识别出五颗此前未被发现的超短周期系外行星——*Kepler-158d*、*Kepler-963c*、*Kepler-879c*、*Kepler-1489c* 和 KOI-4978.02。CETRA 则进一步在 GPU 加速的基础上结合线性时域搜索、相位折叠与物理模板建模，在低信噪比（尤其长周期）凌星检出灵敏度与计算效率上同时取得一定的成效^[71]。未来的发展方向包括：（1）端到端检测管线的建立：将初筛（周期估计）与审查（候选体分类）统一为单一深度学习框架，减少对经典算法的依赖；（2）时序模型的深度引入：基于自注意力的 Transformer 架构可直接建模未折叠光变的时间序列结构与长程依赖，并为无需相位折叠的候选体识别提供新路径^[99]；（3）单次或少次凌星专用检测算法：例如 Vivien 等提出的 PANOPTICON 面向 PLATO 光变的单次凌星深度学习检出框架^[100]，以及利用视向速度约束缩小参数空间、开展靶向凌星搜寻的协同策略^[118]。

4.2.3 统一基准与领域协作

当前系外行星深度学习研究面临的一个突出问题是，尚缺乏面向不同任务场景的统一公开基准数据集与标准化评估框架。各研究组通常采用自建训练与测试样本，在样本构成、噪声特性、预处理流程与标注标准等方面存在差异，导致不同方法之间的性能比较缺乏统一、客观的参照体系。未来若能建立覆盖不同轨道周期、行星半径与信噪比区间，并显式纳入多类假阳性场景（如食双星、背景掺杂、系统性噪声伪信号等）的标准化基准数据集，并配套统一评估指标（例如分周期段的精确率-召回率曲线、长周期区间的完备度-可靠性评估，以及注入-恢复一致性检验），将有助于推动方法学的系统性比较与可重复改进。

在领域协作层面，NASA Exoplanet Archive 及 ExoFOP 等公共平台^[17]已为数据汇交、查询接口、候选体信息整合与后随观测协同提供了重要基础设施支撑，可为未来构建统一 benchmark、共享评测流程与沉淀可复现工具链提供良好平台条件。

A TLS 试探频率网格 (补充)

下述(7)–(10)给出 Hippke 与 Heller^[35] 所采用的非均匀频率构造, 供需复现网格的读者查阅; 一般读者可仅阅读第3节中的文字概括。

设观测时间跨度为 T_{total} , 最少要求的凌星次数为 N_{transit} , 则最小与最大搜索频率分别为

$$f_{\min} = \frac{N_{\text{transit}}}{T_{\text{total}}}, \quad f_{\max} = \frac{1}{2\pi} \sqrt{\frac{GM_*}{(3R_*)^3}}. \quad (7)$$

引入过采样参数 OS 后, 定义常数

$$A = \frac{(2\pi)^{2/3}}{\pi} \frac{R_*}{(GM_*)^{1/3}} \frac{1}{T_{\text{total}} OS}. \quad (8)$$

最优测试频率数目满足^[66]

$$N_{\text{opt}} = \left(f_{\max}^{1/3} - f_{\min}^{1/3} \right) \frac{3}{A} + 1, \quad (9)$$

频率序列取为

$$f_x = \left(\frac{Ax}{3} + f_{\min}^{1/3} - \frac{A}{3} \right)^3, \quad x = 1, \dots, N_{\text{opt}}, \quad (10)$$

周期网格由 $P_x = 1/f_x$ 给出。

参考文献:

- [1] Mayor M, Queloz D. Nature, 1995, 378: 355
- [2] Charbonneau D, Brown T M, Latham D W, et al. ApJ, 2000, 529: L45
- [3] Henry G W, Marcy G W, Butler R P, et al. ApJ, 2000, 529: L41
- [4] Howard A W, Marcy G W, Bryson S T, et al. ApJS, 2012, 201: 15
- [5] Fressin F, Torres G, Charbonneau D, et al. ApJ, 2013, 766: 81
- [6] Petigura E A, Marcy G W, Howard A W. ApJ, 2013, 770: 69
- [7] Bryson S, Kunimoto M, Kopparapu R K, et al. AJ, 2020, 161: 36
- [8] Thompson S E, Coughlin J L, Hoffman K, et al. ApJS, 2018, 235: 38
- [9] Borucki W J, Koch D, Basri G, et al. Science, 2010, 327: 977
- [10] Koch D G, Borucki W J, Basri G, et al. ApJ, 2010, 713: L79
- [11] Ricker G R, Winn J N, Vanderspek R, et al. JATIS, 2014, 1: 014003
- [12] Rauer H, Aerts C, Cabrera J, et al. Experimental Astronomy, 2025, 59
- [13] Ge J, Zhang H, Zang W, et al. <https://arxiv.org/abs/2206.06693>, 2022
- [14] Ge J, Zhang H, Deng H, et al. The Innovation, 2022, 3: 100271
- [15] Ge J, Chen W, Chen Y, et al. Chinese Journal of Space Science, 2024, 44: 400
- [16] Akeson R L, Chen X, Ciardi D, et al. PASP, 2013, 125: 989
- [17] Christiansen J L, McElroy D L, Harbut M, et al. PSJ, 2025, 6: 186
- [18] 胡泉泉, 葛健. 天文学进展, 2026
- [19] Auvergne M, Bodin P, Boisnard L, et al. A&A, 2009, 506: 411
- [20] Howell S B, Sobeck C, Haas M, et al. PASP, 2014, 126: 398
- [21] Pollacco D, Skillen I, Cameron A, et al. PASP, 2006, 118: 1407

- [22] Pepper J, Kuhn R B, Siverd R, et al. *PASP*, 2012, 124: 230
- [23] 陈雨, 葛健. *天文学进展*, 2026
- [24] Jenkins J M, Caldwell D A, Chandrasekaran H, et al. *ApJ*, 2010, 713: L87
- [25] Pont F, Zucker S, Queloz D. *MNRAS*, 2006, 373: 231
- [26] Koch D G, Borucki W J, Basri G, et al. *ApJ*, 2010, 713: L79
- [27] Van Cleve J E, Caldwell D A. No. KSCI-19033, NASA Ames Research Center, 2009
- [28] Jenkins J M, Caldwell D A, Chandrasekaran H, et al. *ApJ*, 2010, 713: L87
- [29] Aigrain S, Pont F, Zucker S. *MNRAS*, 2012, 419: 3147
- [30] Vanderburg A, Johnson J A. *PASP*, 2014, 126: 948
- [31] Lightkurve Collaboration, Cardoso J V d M, Hedges C, et al., 2018
- [32] Gilliland R L, Chaplin W J, Jenkins J M, et al. *AJ*, 2015, 150: 133
- [33] Mathur S, García R A, Ballot J, et al. *A&A*, 2014, 562: A124
- [34] Kovács G, Zucker S, Mazeh T. *A&A*, 2002, 391: 369
- [35] Hippke M, Heller R. *A&A*, 2019, 623: A39
- [36] Carter J A, Winn J N. *ApJ*, 2009, 704: 51
- [37] Stumpe M C, Smith J C, Van Cleve J E, et al. *PASP*, 2012, 124: 985
- [38] Smith J C, Stumpe M C, Van Cleve J E, et al. *PASP*, 2012, 124: 1000
- [39] Stumpe M C, Smith J C, Catanzarite J H, et al. *PASP*, 2014, 126: 100
- [40] Twicken J D, Jenkins J M, Seader S E, et al. *AJ*, 2016, 152: 158
- [41] Tamuz O, Mazeh T, Zucker S. *MNRAS*, 2005, 356: 1466
- [42] Mazeh T, Tamuz O, Zucker S. <https://arxiv.org/abs/astro-ph/0612418>, 2006
- [43] Aigrain S, Parviainen H, Pope B J S. *MNRAS*, 2016, 459: 2408
- [44] Mislis D, Pyrzas S, Alsubai K A, et al. *MNRAS*, 2016, 465: 3759
- [45] Aigrain S, Parviainen H, Roberts S, et al. *MNRAS*, 2017, 471: 759
- [46] Gibson N P. *MNRAS*, 2014, 445: 3401
- [47] Wang D, Hogg D W, Foreman-Mackey D, et al. *PASP*, 2016, 128: 094503
- [48] Petigura E A, Marcy G W. *PASP*, 2012, 124: 1073
- [49] Kim D W, Protopapas P, Alcock C, et al. *MNRAS*, 2009, 397: 558
- [50] Vaughan S. *A&A*, 2005, 431: 391
- [51] Krishnan S, Markowitz A G, Schwarzenberg-Czerny A, et al. *MNRAS*, 2021, 508: 3975
- [52] Angus R, Morton T, Aigrain S, et al. *MNRAS*, 2017, 474: 2094
- [53] Matesic M R B, Rowe J F, Livingston J H, et al. *AJ*, 2024, 167: 68
- [54] Gondhalekar Y, Feigelson E D, Caceres G A, et al. *ApJL*, 2023, 959: L16
- [55] Morvan M, Nikolaou N, Tsiaras A, et al. *AJ*, 2020, 159: 109
- [56] Giordano Orsini M, Ferone A, Inno L, et al. *Astronomy and Computing*, 2025, 52: 100964
- [57] Morvan M, Nikolaou N, Yip K H, et al. <https://arxiv.org/abs/2207.02777>, 2022
- [58] Taaki J S, Kembal A J, Kamalabadi F. *AJ*, 2024, 167: 60
- [59] Jenkins J M. *ApJ*, 2002, 575: 493
- [60] Carter J A, Agol E. *ApJ*, 2013, 765: 132
- [61] Scargle J D. *ApJ*, 1982, 263: 835
- [62] Stellingwerf R F. *ApJ*, 1978, 224: 953
- [63] Shahaf S, Zackay B, Mazeh T, et al. *MNRAS*, 2022, 513: 2732
- [64] Collier Cameron A, Pollacco D, Street R A, et al. *MNRAS*, 2006, 373: 799
- [65] Mandel K, Agol E. *ApJ*, 2002, 580: L171
- [66] Ofir A. *A&A*, 2014, 561: A138
- [67] North D O. RCA Tech Rep, 1947
- [68] Van Trees H L. *Detection, Estimation, and Modulation Theory*. John Wiley & Sons, 1968
- [69] Hu Q, Ge J, Jin L, et al. *GTLS: A method for speeding up periodic transit detection using GPU*. 2026
- [70] Wang K, Ge J, Willis K, et al. *MNRAS*, 2024, 534: 1913

- [71] Smith L C, Ahmed S, De Angeli F, et al. *MNRAS*, 2025, 539: 297
- [72] Thompson S E, Mullally F, Coughlin J, et al. *ApJ*, 2015, 812: 46
- [73] Mullally F, Coughlin J L, Thompson S E, et al. *PASP*, 2016, 128: 074502
- [74] Armstrong D J, Pollacco D, Santerne A. *MNRAS*, 2016, 465: 2634
- [75] Mislis D, Pyrzas S, Alsubai K A. *MNRAS*, 2018, 481: 1624
- [76] Armstrong D J, Gamper J, Damoulas T. *MNRAS*, 2020, 504: 5327
- [77] McCauliff S D, Jenkins J M, Catanzarite J, et al. *ApJ*, 2015, 806: 6
- [78] Panahi A, Zucker S. *PASP*, 2021, 133: 024502
- [79] Panahi A, Zucker S, Clementini G, et al. *A&A*, 2022, 663: A101
- [80] Malik A, Moster B P, Obermeier C. *MNRAS*, 2021
- [81] Melton E J, Feigelson E D, Montalto M, et al. *AJ*, 2024, 167: 202
- [82] Shallue C J, Vanderburg A. *AJ*, 2018, 155: 94
- [83] Pearson K A, Palafox L, Griffith C A. *MNRAS*, 2017, 474: 478
- [84] Dattilo A, Vanderburg A, Shallue C J, et al. *AJ*, 2019, 157: 169
- [85] Yu L, Vanderburg A, Huang C, et al. *AJ*, 2019, 158: 25
- [86] Tey E, Moldovan D, Kunimoto M, et al. *AJ*, 2023, 165: 95
- [87] Ansdell M, Ioannou Y, Osborn H P, et al. *ApJL*, 2018, 869: L7
- [88] Valizadegan H, Martinho M J S, Wilkens L S, et al. *ApJ*, 2022, 926: 120
- [89] Valizadegan H, Martinho M J S, Jenkins J M, et al. *AJ*, 2023, 166: 28
- [90] Zucker S, Giryes R. *AJ*, 2018, 155: 147
- [91] Iglesias Álvarez S, Díez Alonso E, Sánchez Rodríguez M L, et al. *Axioms*, 2023, 12: 348
- [92] Yeh L C, Jiang I G. *PASP*, 2020, 133: 014401
- [93] Wang K, Ge J, Willis K, et al. *MNRAS*, 2024, 528: 4053
- [94] Osborn H P, Ansdell M, Ioannou Y, et al. *A&A*, 2020, 633: A53
- [95] Chintarungruangchai P, Jiang I G. *PASP*, 2019, 131: 064502
- [96] Dvash E, Peleg Y, Zucker S, et al. *AJ*, 2022, 163: 237
- [97] Cui K, Liu J, Feng F, et al. *AJ*, 2021, 163: 23
- [98] Salinas H, Pichara K, Brahm R, et al. *MNRAS*, 2023, 522: 3201
- [99] Salinas H, Brahm R, Olmschenk G, et al. *MNRAS*, 2025, 538: 2031
- [100] Vivien H G, Deleuil M, Jannsen N, et al. *A&A*, 2025, 694: A293
- [101] Dévora-Pajares M, Pozuelos F J, Suárez J C, et al. <https://arxiv.org/abs/2511.08768>, 2025
- [102] Fiscale S, Inno L, Rotundi A, et al. *AJ*, 2025, 170: 73
- [103] Fiscale S, Ferone A, Ciaramella A, et al. *Electronics*, 2025, 14: 1738
- [104] Tardugno Poleo V, Eisner N, Hogg D W. *AJ*, 2024, 168: 100
- [105] Valizadegan H, Martinho M J S, Jenkins J M, et al. *AJ*, 2025, 170: 287
- [106] Winn J N. Seager S, ed. *Exoplanets*. Tucson, AZ: University of Arizona Press, 2010: 55
- [107] Borucki W J, Summers A L. *Icarus*, 1984, 58: 121
- [108] Burke C J, Christiansen J L, Mullally F, et al. *ApJ*, 2015, 809: 8
- [109] Rasmussen C E, Williams C K I. *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press, 2006
- [110] Guerrero N M, Seager S, Huang C X, et al. *ApJS*, 2021, 254: 39
- [111] Coughlin J L, Mullally F, Thompson S E, et al. *ApJS*, 2016, 224: 12
- [112] Mullally F, Coughlin J L, Thompson S E, et al. *Publications of the Astronomical Society of the Pacific*, 2016, 128: 074502
- [113] Bryson S T, Jenkins J M, Gilliland R L, et al. *PASP*, 2013, 125: 889
- [114] Désert J M, Charbonneau D, Torres G, et al. *ApJ*, 2015, 804: 59
- [115] Santerne A, Moutou C, Tsantaki M, et al. *A&A*, 2016, 587: A64
- [116] Ge J, Zhang H, Zhang Y, et al. Coyle L E, Matsuura S, Perrin M D, eds. *Space Telescopes and Instrumentation 2022: Optical, Infrared, and Millimeter Wave: Vol. 12180*. SPIE, 2022: 1218015
- [117] Ge J, Zhang H, Zhang Y, et al. Coyle L E, Matsuura S, Perrin M D, eds. *Space Telescopes and Instrumentation 2024: Optical, Infrared, and Millimeter Wave: Vol. 13092*. SPIE, 2024: 1309218
- [118] Damasso M, Rodrigues J, Castro-González A, et al. *A&A*, 2023, 679: A33

Review on Detecting Long-Period Weak Transiting Signals Using Artificial Intelligence

YAN Xingcheng^{1,2}, LIU Qingtian^{1,2}, GE Jian¹, RUAN Zeyu^{1,2}, LIU Zhenghong^{1,2}

(1. Shanghai Astronomical Observatory, Chinese Academy of Sciences, Shanghai 200030, China; 2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: The transit method is the primary technique for exoplanet discovery. While high-precision space-based photometric surveys have yielded massive light-curve datasets, the detection and statistical characterization of long-period, small-radius planets remain constrained compared to their short-period counterparts. Long-period targets have low a priori geometric transit probabilities, yield fewer observable transits within a finite baseline, and (for small planets) produce shallow transit depths; together these effects increase sensitivity to stellar red noise and instrumental systematics, reduce signal significance, and complicate candidate discrimination. This paper reviews the search pipelines for weak long-period transit signals, examining key technical challenges and progress in light-curve noise modeling, preprocessing, triage, and vetting. It synthesizes the strengths and limitations of cotrending and detrending methods, classical period-search algorithms, and artificial intelligence approaches under low signal-to-noise ratio (SNR) conditions. Furthermore, it discusses critical challenges in evaluating completeness and reliability. Finally, in the context of future space photometry missions, the paper outlines prospects for automated pipelines and unified benchmarking in the search for long-period Earth-like planets.

Key words: planets and satellites: detection; (stars:) planetary systems; techniques: photometric; methods: statistical; surveys; software: machine learning; deep learning